Expanding soybean cultivation boundaries by characterizing novel genes involved in early maturity

Jérôme Gélinas Bélanger

Department of Plant Science
McGill University, Montreal
December 2024

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Doctor of Philosophy

© Jérôme Gélinas Bélanger 2024

Rien n'est meilleur que l'agriculture, rien n'est plus beau, rien n'est plus digne d'un homme libre.
Elle suffit amplement aux besoins de notre vie.

Damase Potvin

Table of contents

List of Tables

List of Figures

List of Abbreviations

bp	Base pair
CDS	Coding sequence
CEN	Co-expression network
CI	Confidence interval
CIM	Composite interval mapping
cM	Centimorgan
CRE	*Cis*-regulatory elements
CRISPR-Cas9	Clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9
DBD	DNA-binding domain
DEG	Differentially expressed gene
DNA	Deoxyribonucleic acid
eQTL	Expression quantitative trait loci
FC	Fold change
FDR	False discovery rate
FRSPD	Flowering / Reproduction / Senescence / Photosynthesis / Development
FRSPD_GO	Gene ontology terms associated with FRSPD
Gbp	Gigabase pair
GBS	Genotyping-by-sequencing
GCIM	Genome-wide composite interval mapping
GM	*Glycine max*
GO	Gene ontology
GRIN	Germplasm resources information network
GRN	Gene regulatory network
GWAS	Genome-wide association studies
ICIM	Inclusive composite interval mapping
IM	Interval mapping
InDels	Insertions/Deletions
Kbp	Kilobase pair
L:D	Light:Darkness
LCH	Light-harvesting complex
LD	Long-days
LG	Linkage group
LOD	Logarithm of odds
LOF	Loss-of-function
MA	Cultivar Maple Arrow
MAD	Modified augmented design
MAF	Minor allele frequency
Mbp	Megabase pair

MD	Cultivar AAC Mandor
MG	Maturity groups
mRNA	Messenger ribonucleic acid
NEG_{TWCEN}	Negatively correlated transcriptome-wide co-expression network
NIL	Near-isogenic line
OV	Cultivar OAC Vision
PCC	Pearson correlation coefficient
PCR	Polymerase chain reaction
PFASW	Protein / Fatty Acids / Seed Weight
PI	Plant introduction
PIN	Probability in stepwise regression
POS_{TWCEN}	Positively correlated transcriptome-wide co-expression network
PSI/II	Photosystem I/II
PVE	Phenotypic variation explained
Q-Q	Quantile-quantile
qRT-PCR	Quantitative real-time PCR
QTL	Quantitative trait loci
R:FR	Red-to-far–red ratio
REML	Restricted maximum likelihood
RH	Relative humidity
RIL	Recombinant inbred line
RNA	Ribonucleic acid
SD	Short-days
SIFT4g	Sorting Intolerant From Tolerant 4G
SNP	Single nucleotide polymorphism
TF	Transcription factor
TSS	Transcription start site
TWAS	Transcriptome-wide association study
TWCEN	Transcriptome-wide co-expression network
UTR	Untranslated region
VEP	Variant Effect Predictor
WGS	Whole genome sequencing
WT	Wild type

List of Gene Abbreviations

ACBP3	*ACYL-COA-BINDING DOMAIN*
AP	*APETALA*
APL	*ALTERED PHLOEM DEVELOPMENT*
ARR	*RESPONSE REGULATOR*
CCA1	*CIRCADIAN CLOCK ASSOCIATED1*
CCT	*CONSTANS, CO-like, TOC1*
CDF	*CYCLING DOF FACTOR*
CO	*CONSTANS*
CRU	*CRUCIFERIN*
DDA	*DOWN IN DARK AND AUXIN*
EID	*EMPFINDLICHER IM DUNNKELROTEN LICHT*
ELF	*EARLY FLOWERING*
ENR	*ENOYL-ACP REDUCTASE*
FAP	*FATTY-ACID-BINDING PROTEIN*
FATB4B	*FATTY ACYL-ACP THIOESTERASES B*
FHY	*PROTEIN FAR-RED ELONGATED HYPOCOTYL*
FLC	*FLOWERING LOCUS C*
FRS	*FAR1-RELATED SEQUENCE*
FT	*FLOWERING LOCUS T*
FUL	*FRUITFULL*
FUS	*FUSCA*
GI	*GIGANTEA*
LBD21	*LOB DOMAIN-CONTAINING PROTEIN 21*
LHC	*LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT*
LHY	*LATE ELONGATED HYPOCOTYL*
MDE	*MADS-BOX DOWNREGULATED BY E1*
MFT	*MOTHER OF FT AND TFL1*
MOD1	*MOSAIC DEATH 1*
PHYA/B	*PHYTOCHROME A/B*
PRR	*PSEUDO RESPONSE REGULATOR*
TFL1	*TERMINAL FLOWER 1*
TFS1	*TARGET OF FLC AND SVP1*
TOF	*TIME OF FLOWERING*
TT8	*TRANSPARENT TESTA 8*
ZIFL1	*ZINC INDUCTED FACILITATOR-LIKE 1*

Abstract

Soybean is a short-day flowering crop originating from East Asia that is naturally adapted to sub-tropical environments. Although extensive research has been pursued on the topic, multiple aspects of the intricate molecular processes regulating the flowering and maturity processes in soybean remain elusive, mainly because of the high complexity and density of the genetic networks. The main objective of this thesis was to study the genetic structure regulating reproductive-related traits (i.e., flowering, pod-filling, and maturity) in soybean as an effort to support the accelerated breeding of this crop and understand the genetic mechanisms underlying these biological processes.

The first sub-objective of this study was to identify novel key loci involved in the regulation of four traits related to reproduction in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations. Using a combinatorial mapping strategy based on two algorithms, three quantitative trait loci signals with important additive effects on pod-filling and maturity were identified on chromosomes GM04 and GM08 in both populations. In addition, this study revealed that the *E8* locus, known to cover a broad ~37.5 Mbp region (~7-44.5 Mbp), is regulated by three distinct regions located at GM04:16,974,874-17,152,230 (*E8-r1*), GM04:35,168,111-37,664,017 (*E8-r2*), and GM04:41,808,599-42,376,237 (*E8-r3*). All of the identified regions were physically close to or encompassed major flowering genes such as *Glyma.04G124300* (*E8-r1*), *Glyma.04G156400* (*E8-r2*), *Glyma.04G167900* (*E8-r3*) and *Glyma.04G168300* (*E8-r3*).

The second sub-objective was to develop an expression quantitative trait locus mapping pipeline to identify *trans* and *cis* interactions regulating four candidate genes (*Glyma.04G168300*/*GmCDF3*, *Glyma.04G167900*/*GmLHCA4a*, *Glyma.04G166300*/*GmPRR1a*, and *Glyma.04G159300*/*GmMDE04*) and regulatory hotspots associated with Flowering / Reproduction / Senescence / Photosynthesis / Development (FRSPD) functions in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations. Using a combinatorial eQTL mapping strategy, we identified with high confidence a total of 2,218 *trans* (2,061 genes) / 7 *cis* (7 genes) in $QS15524_{F2:F3}$ and 4,073 *trans* (2,842 genes) / 3,083 *cis* (2,418 genes) in $QS15544_{RIL}$. From the *trans* signals, we have identified three hotspots (GM06:39,892,719-43,437,125, GM17:5,431,473-7,260,313, and GM18:1,434,182-1,935,386) involved in FRSPD functions in $QS15524_{F2:F3}$, and one hotspot (GM04:10,812,813-10,985,437) in $QS15544_{RIL}$. Furthermore, co-expression and eQTL analyses suggest that *ALTERED PHLOEM DEVELOPMENT* (*Glyma.15G263700*) and *DOMAIN-*

CONTAINING PROTEIN 21 (*Glyma.18G025600*) genes are the best candidates for the F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386 hotspots, respectively.

The third sub-objective was to identify the key regions involved in the regulation of six seed PFASW (Protein / Fatty Acids / Seed Weight) quality traits in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations. Using our combinatorial mapping strategy, we identified a total of four and five major QTL regions in $QS15524_{F2:F3}$ and $QS15544_{RIL}$, respectively. Moreover, three additional regions regulating the 100-seed weight trait and several other quality traits were detected in both populations. In $QS15524_{F2:F3}$, the F2_GM04.2 (GM04:36,499,381-40,206,770) oleic acid-regulating region has been found in close linkage with the *E8-r2* pod-filling and maturity region. In addition, the RIL_GM04 (GM04:16,853,028-18,312,993) and RIL_GM16 (GM16:5,841,864-5,861,155) seed weight-regulating regions have been found to be respectively in close linkage with the *E8-r1* pod-filling and GM16:5,680,173-5,730,237 maturity loci in $QS15544_{RIL}$.

Overall, the results generated in this thesis demonstrate that a few key overarching loci regulate each of these traits and thus provide useful insights to develop high-quality and high-yielding cultivars belonging to the early maturity groups MG00, MG000, and beyond.

Résumé

Originaire de l'Asie de l'Est, le soja est une plante fleurissant sous jours courts naturellement adaptée aux environnements de culture subtropicaux. Bien qu'énormément de recherche ait été effectuée sur la question, plusieurs aspects des processus moléculaires sous-jacents à la floraison de cette espèce demeurent vagues, notamment à cause de la forte complexité et l'immense densité des réseaux génétiques intrinsèques à ce phénomène. L'objectif principal de cette thèse est d'étudier la matrice génétique régulant les traits liés à la reproduction du soja afin de soutenir l'amélioration génétique accélérée de cette plante et comprendre les mécanismes sous-jacents à ces processus biologiques.

Le premier sous-objectif de cette étude fût d'identifier des gènes jouant un rôle clé dans la régulation de quatre traits liés à la reproduction des populations $QS15524_{F2:F3}$ et $QS15544_{RIL}$. En utilisant une stratégie de cartographie combinatoire basée sur deux algorithmes, trois loci de traits quantitatifs communs à ces deux populations ont été identifiés sur les chromosomes GM04 et GM08. De plus, cette étude a révélé que le locus *E8*, connu pour couvrir une large région génomique de ~37.5 Mpb (~7-44.5 Mpb), est en fait régulé par trois régions distinctes situées aux positions GM04:16,974,874-17,152,230 (*E8-r1*), GM04:35,168,111-37,664,017 (*E8-r2*) et GM04:41,808,599-42,376,237 (*E8-r3*). Toutes les régions identifiées comprennent ou sont situées à des positions physiques très rapprochées de plusieurs gènes de floraison tels que *Glyma.04G124300* (*E8-r1*), *Glyma.04G156400* (*E8-r2*), *Glyma.04G167900* (*E8-r3*) et *Glyma.04G168300* (*E8-r3*).

Le deuxième sous-objectif fût de développer un protocole de cartographie de traits quantitatifs d'expression pour identifier les interactions *trans* et *cis* régulant les fonctions associées à la Floraison / Reproduction / Sénescence / Photosynthèse / Développement (FRSPD) dans les populations $QS15524_{F2:F3}$ et $QS15544_{RIL}$. En utilisant ce protocole de cartographie, nous avons identifié un total de 2,218 *trans* (2,061 gènes) / 7 *cis* (7 gènes) dans la population $QS15524_{F2:F3}$ et 4,073 *trans* (2,842 gènes) / 3,083 *cis* (2,418 gènes) dans la population $QS15544_{RIL}$ avec un haut degré de confiance. Avec les signaux *trans*, nous avons identifié trois points chauds (GM06:39,892,719-43,437,125, GM17:5,431,473-7,260,313, et GM18:1,434,182-1,935,386) régulant les fonctions FRSPD dans la population $QS15524_{F2:F3}$, et un point chaud (GM04:10,812,813-10,985,437) dans la population $QS15544_{RIL}$. De surcroît, des analyses de co-expression de gènes et de cartographie de traits quantitatifs d'expression suggèrent que les gènes

Glyma.15G263700 et *Glyma.18G025600* sont les meilleurs candidats pour les points chauds F2_GM15:49,385,092-49,442,237 et F2_GM18:1,434,182-1,935,386, respectivement.

Le troisième sous-objectif fût d'identifier les régions clés impliquées dans la régulation de six traits de qualité PAGPG (Protéines / Acides Gras / Poids du Grain) dans les populations $QS15524_{F2:F3}$ et $QS15544_{RIL}$. En utilisant une stratégie de cartographie combinatoire, nous avons respectivement identifié un total de quatre et cinq régions majeures dans les populations $QS15524_{F2:F3}$ et $QS15544_{RIL}$. De plus, trois régions régulant le poids 100-grains et d'autres traits de qualité ont été détectées dans les deux populations. Dans la population $QS15524_{F2:F3}$, la région F2_GM04.2 (GM04:36,499,381-40,206,770) régulant le contenu en acide oléique a été identifiée comme étant en forte liaison avec la région *E8-r2* régulant le remplissage et la maturité.

De façon générale, les résultats générés lors de cette thèse démontrent que seulement quelques loci majeurs régulent chacun de ces traits et fournissent ainsi des informations clés pour développer des cultivars avec des rendements élevés et de bonne qualité appartenant aux groupes de maturité hâtifs MG00, MG000 et même plus.

Acknowledgments

My journey to building the knowledge necessary to write this thesis was challenging at all levels (physically, mentally, intellectually, and morally), but at the same time extremely rewarding in terms of personal accomplishment.

This journey would not have been possible without the help and support of my three supervisors, Drs. Louise O'Donoughue, Tanya Copley, and Valerio Hoyos-Villegas. My deepest gratitude goes to them for accepting me as their student, coaching me throughout this journey, and supporting my relentless chaotic approach to science. Altogether, we have been able to build this project from scratch and reach port as originally planned.

I want to thank Dr. Louise O'Donoughue for believing in me from the start and supervising me throughout this experience. Without your support, I do not think this experience would have been this satisfying. I want to express my deepest thanks to Dr. Tanya Copley for stepping in when the situation was dire. I have no words that can speak for all the gratitude that I have toward this decision since it saved the show. My most sincere thanks to Dr. Hoyos-Villegas for his input in this project at all levels, from the administrative duties to the precious bits of advice on science and plant breeding.

I would also like to acknowledge the input of my Ph.D. committee member, Dr. Jean-Benoit Charron, who has been a precious help throughout my studies at McGill, from my start as an undergraduate to my graduation as a PhD student. My deepest thanks to all the faculty members in the Plant Science Department, the support staff, and all the students. I consider you all as part of my first family and your support helped me throughout this adventure. I want to give a special shoutout to Luc Ouellette from the Charron lab for his support and pieces of science advice. Luc, you will be a great scientist, many thanks for your help.

I would like to thank all the members of the Soybean Breeding and Genetics Lab at CEROM, especially Éric Fortier. Thank you Éric for helping me through this adventure and supporting me every time I needed it. My thanks also go to Joannie Berthon and Daphnée Paré for your help and involvement in this project. For the people at CEROM, I consider you all as part of my second family and your thoughts and support made a big difference in the success of this PhD program.

I would also like to acknowledge the funding I received from McGill University, les Fonds de Recherche du Québec, Centre SEVE, Germination Magazine/Seed World, MITACS, and the

Natural Sciences and Engineering Research Council of Canada throughout these years. This PhD project would not have been possible without the financial contributions of all these stakeholders and I will be eternally grateful to them for choosing me as one of their recipients.

A special thanks to all the members of my family who supported me in this adventure from day one. My dad, Julien Bélanger, thank you for introducing me to agriculture and listening to my never-ending stories about the importance of plant breeding. My mom, Denise Gélinas, for her moral support throughout my studies and her unconditional support of my professional choices. My in-laws Jacques and Joanne Corbu, for believing in my potential. My dog, Chester, who has been my most precious friend during these years of hard work, stayed up at night with me to support me when I was studying and supported me morally with my letdowns in the lab. Chester, you will never be forgotten. My partner, Michelle Corbu, who has been the keystone of my success in this adventure. Thank you for your unconditional support and patience in this adventure despite the numerous challenges. Your love and patience gave me the strength to succeed and bounce back when knocked. Your view of the world helped me to see and conceptualize things differently when I was facing hardships in the lab. You are incredible.

Preface and Author Contributions

The following thesis was prepared according to the "Thesis Guidelines" of McGill University. This thesis contains four chapters (Chapters 3-6) representing three distinct research manuscripts for publication (Chapters 3-5) and a general discussion (Chapter 6). Chapters 3 and 4 were respectively published in Frontiers in Plant Science, whereas Chapter 5 was published in the Canadian Journal of Plant Science. Segments of Chapter 6 are based on a review article published in Plant Methods.

Author contributions

Jérôme Gélinas Bélanger processed the phenotyping and genotyping datasets, developed the analytic pipelines, conducted experiments, and wrote the manuscripts in all chapters.

Dr. Louise O'Donoughue provided supervision, guidance, and funding for the research, generated the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, and reviewed the manuscripts in all chapters.

Dr. Tanya Copley provided supervision and guidance, designed the research, planned the experiments, processed the phenotyping and genotyping datasets, and reviewed the manuscripts in all chapters.

Dr. Valerio Hoyos-Villegas provided supervision and guidance and reviewed the manuscripts in all chapters.

Contributions to Knowledge

Chapter 3

- In this study, novel major QTL regions (e.g., *E8-r1* to *E8-r3*) and eQTL interactions (e.g., *E8-r3* regulating *E6*) involved in the regulation of reproductive traits are identified. Using this information, several candidate SNPs and genes are proposed based on an extensive literature review and using a five-step variant analysis pipeline.

Chapter 4

- In this study, regions regulating the expression of three *E8-r3* candidate genes identified in Chapter 3 are detected using a novel combinatorial mapping pipeline. Several candidate SNPs associated with transcription factors are also proposed using a pipeline combining co-expression network and single polymorphism variant analyses.

Chapter 5

- In this study, 12 major QTL regions regulating PFASW traits are identified in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, including three found to be in close linkage with pod-filling and maturity loci identified in Chapter 3. For each of the regions, candidate single nucleotide polymorphisms are proposed using a candidate SNP prediction pipeline.

Chapter 6

- In this manuscript, the main findings of this thesis are discussed and insights about topics, such as genetic bottlenecks and future technological avenues for breeders, are provided.

Overall, the work presented in this thesis highlights important regions of the soybean genome that need to be targeted by breeders who want to expand the cultivation of this species beyond its actual northern latitudinal limits. This study also provides useful insights to understand the molecular mechanisms governing the northern distribution of the Canadian soybean germplasm.

1. Introduction

1.1 Genotypic Diversity as a Nexus for Sustainable Agriculture and Food Security

Global agroecosystems need to evolve to mitigate the consequences caused by climate changes, cope with the decrease in natural and agricultural biodiversity, and sustain the increasing global demand for agricultural goods (Dijk *et al.*, 2021; Muluneh, 2021). Novel cropping systems must be established to provide alternatives for the poorest communities and limit the pervasive effects of intensive agriculture on biodiversity hotspots such as the Amazonian tropical forest (Benton *et al.*, 2021; Eiji *et al.*, 2021; Rasche *et al.*, 2022). As currently predicted, future expansion of agriculture in subtropical and tropical countries will be limited due to the consequences of climate change on the available resources (Ericksen *et al.*, 2011; Cinner *et al.*, 2022). Farmers located closest to the Equator will be challenged with more frequent shortages in water resources and heat waves, higher pest prevalence, larger plant evapotranspiration, greater degradation of soil quality, and quicker loss in critical natural biodiversity (Challinor and Wheeler, 2008; Zhao *et al.*, 2017; Tigchelaar *et al.*, 2018; Raza *et al.*, 2019; Jägermeyr *et al.*, 2021; Masson-Delmotte *et al.*, 2022). For the rural communities of the South, these dramatic environmental changes will ultimately lead to a decrease in their total agricultural output and unstable yields and deepen their economic and cultural insecurities (Challinor and Wheeler, 2008; Zhao *et al.*, 2017; Tigchelaar *et al.*, 2018; Raza *et al.*, 2019; Jägermeyr *et al.*, 2021; Masson-Delmotte *et al.*, 2022). On the other hand, climate change will allow an expansion of agriculture, an increase in cropped acreage and yield, and crop diversification in countries located in colder climates (Wiréhn, 2018; Wolfe *et al.*, 2018; Government of Canada, 2020; Hannah *et al.*, 2020; Bahadur *et al.*, 2021; Meyfroidt, 2021; Unc *et al.*, 2021). Global warming in northern regions is generating "climate-driven agricultural frontiers" which are areas that will become suitable for agriculture or climatically appropriate for a broader range of crops (Hannah *et al.*, 2020). Often overlooked for their pivotal role in maintaining global food security, agroecosystems located in colder areas display several benefits over their tropical counterparts, including reduced pest pressure and urbanization pressure, a lower population density, and a lessened impact on naturally occurring biodiversity (Unc *et al.*, 2021). Despite being net contributors to the global economy in terms of agricultural exports, northern agroecosystems also face inherent agronomic challenges due to their harsh geographical situations.

One of the most severe limitations is the low number of cropping options available to farmers located in northern regions (e.g., oat, wheat, rye, barley, canola and pea; all crops with

short cropping cycles and long-day or day-neutral photoperiodic requirements). Although their center of origin lies in the Fertile Crescent of the Middle East (Murphy, 2007; Preece *et al.*, 2017; Haas *et al.*, 2019; Sun *et al.*, 2022), these crops were slowly adapted to high-latitudinal conditions through multiple decades of selection in northern Europe and Asia, thus enabling the emergence of new cropping systems in these harsh regions (Bakels, 2014; Liu *et al.*, 2015; Vanhanen *et al.*, 2019; Jones and Lister, 2022). As such, plant domestication through selective breeding is a pivotal aspect underpinning the evolution of crop cultivation systems, and agriculture as currently practiced in northern agro-environments relies heavily on the artificial selection of foreign plant species. Still, these long-day crops were more naturally adapted to cultivation in higher latitudes than short-day plants (e.g., rice and soybean) due to their specific genetic structure enabling them to flower under long summer days and which allows for an optimal match between their reproductive phenology and the surrounding septentrional environmental conditions (Nakamichi, 2015; Lin *et al.*, 2021a). To be cultivated under northern latitudes, farmers and breeders have deliberately selected against rice and soybean's natural photoperiodic requirements by favoring the fixation of inactivated alleles involved in the flowering and circadian clock molecular pathways (Hyten *et al.*, 2006; Iquira *et al.*, 2010; Osnato, 2023). For soybeans, this selection process promoted the development of a limited number of locally adapted genotypes which now form the basis of northern soybean cropping systems, such as the ones found in the Canadian provinces of Ontario, Quebec, and Manitoba (Hyten *et al.*, 2006; Iquira *et al.*, 2010).

Glycine max (L.) Merr., the common soybean, is a short-day flowering leguminous crop originating from East Asia and naturally adapted to sub-tropical environments. Worldwide, soybean grains are used as a human food source, in animal feed rations, and as an industrial component to produce oils and plastics (Pagano and Miransari, 2016). As such, the crop is in high demand with a worldwide production of 341.8 M metric tons in 2019 (The American Soybean Association, 2023). Because of its nitrogen-fixing abilities, the crop is introduced in crop rotations to complement non-fixing crops, such as grains crops and canola, and thus represents an ecological complement to chemical nitrogen sources. Although Canada only accounts for 2% of the worldwide soybean production and 3% of the world soybean exports (The American Soybean Association, 2023), the crop has a high farm gate value and is estimated to generate more than $2.5 billion annually for Canadian farmers (Soy Canada, 2022). To this day, soybean cultivation in the Canadian Prairies is constrained mainly to the southern tiers of Manitoba as there is only a limited

number of early-maturing cultivars currently available to farmers (Soy Canada, 2022). In Quebec and Ontario, extensive soybean cultivation is primarily practiced in southern locations due to their higher yields. As a consequence, regions from Northern and Eastern Quebec (Saguenay-Lac-Saint-Jean and Bas-Saint-Laurent) and the Clay Belt (Cochrane District in Ontario, and Abitibi County in Quebec) are still lagging to integrate this crop in their rotations. To meet the growing demand, farmers want to expand the cultivation range of soybean to the northern areas of the Canadian Prairies and Eastern Canada, meaning that cultivars with a better response to longer growing days and shorter summer seasons need to be developed. As such, developing early-maturing soybean varieties is an agricultural research priority for the Canadian soybean sector to increase its competitiveness (Saavedra, 2019). Despite strong support from the sector, several limiting factors still hinder progress in this field, such as the limited gene pool available to researchers and breeders and a complex genetic structure. Over the years, multiple studies have confirmed that the Canadian soybean genetic diversity is notoriously low in comparison to the exotic gene pool (Fu *et al.*, 2007; Iquira *et al.*, 2010) although these levels of diversity have been maintained through the incorporation of exotic germplasm into the breeding programs (Bruce *et al.*, 2019). This low genetic diversity poses multiple challenges for breeders as sources of variation are scarce and beneficial alleles, such as the ones associated with early flowering, suffer from linkage drag with important agronomic and quality traits.

One novel approach to address this issue is through the identification of crucial key genes regulating flowering and maturity along with a better understanding of the underlying genetic structures guiding the general expression patterns of these genes within this limited gene pool (Copley *et al.*, 2018). Although extensive research has been pursued on the topic, multiple aspects of the underlying intricate molecular processes regulating flowering in plants remain unknown, mainly because of the high complexity and density of the genetic networks. One of the main factors driving these intricacies is the high structural redundance of the soybean genome which has been caused by two past duplication events that occurred approximately 13 and 59 million years ago and led to an increase in the size of the soybean genome (1.1-1.15 Gbp in soybean vs 115-120 Mbp in *Arabidopsis*) and high rates of genetic redundancy, subfunctionalization, and degeneration (Shultz *et al.*, 2006; Swarbreck *et al.*, 2008; Schmutz *et al.*, 2010). In spite of these challenges, researchers have successfully identified at least 25 major flowering genes through forward and reverse molecular studies (Bouché *et al.*, 2016; Zhang *et al.*, 2017c). In complement, 844 genes have been

identified *in silico* from 306 *Arabidopsis* orthologues identified based on loss-of-function (LOF) and transgenic analyses (Bouché *et al.*, 2016; Zhang *et al.*, 2017c). At least four of them, *E1, E2, E3,* and *E4*, are now commonly used in breeding to generate early maturing varieties (Tsubokura *et al.*, 2014; Jähne *et al.*, 2020; Lin *et al.*, 2021b).

Preliminary studies conducted by Copley and O'Donoughue using genome-wide association (GWA) analysis (Copley *et al.*, 2018) and biparental QTL mapping (O'Donoughue, unpublished) suggest that multiple undeciphered genes might be participating in this regulatory process guiding early reproductive traits. In its purest form, the research project presented in this proposal is a continuation of the work performed by Copley and O'Donoughue to unravel these novel regulators. To do so, two mapping populations (one F_2 and one RIL) were generated from biparental crosses to perform QTL analysis. The first, named $QS15524_{F2:F3}$, was generated from the cross between 'Maple Arrow' (MG00; later-maturing accession) × 'OAC Vision' (PI 567787) (MG000; earlier-maturing accession). The second, named $QS15544_{RIL}$, was generated from the cross between 'AAC Mandor' (MG00; later-maturing accession) × '9004' (MG000; earlier-maturing accession). A third collection, named $MadMaturity_{86}$ and comprising 86 early-maturing accessions from diverse origins, was previously developed and used to identify novel loci involved in the regulation of reproductive and seed quality traits (Copley *et al.*, 2018).

Overall, these three populations were developed with the intent of being interconnected using 'OAC Vision' as a linker. As such, 'OAC Vision' was incorporated into the $MadMaturity_{86}$ association panel for genotyping and phenotyping. In addition, one of 'AAC Mandor' parents in the $QS15544_{RIL}$ population is 'OAC 00-07', a cultivar generated from the cross between 'SL90-655' × 'OAC Vision'. The bridging of these populations with 'OAC Vision' has several benefits over using unrelated populations, including (i) strengthening of the results through counter-validation and (ii) validation of the mapping pipelines. Based on these observations, we reason that these interconnected populations can support the identification of loci involved in the regulation of early reproductive traits, expression, and seed quality with high confidence.

1.2 Hypothesis and Objectives

The latitudinal introduction of exotic plant material requires multiple cycles of laborious selective breeding. Soybean is a short-day flowering leguminous crop naturally adapted to sub-tropical environments that is used as a model organism to understand the molecular mechanisms underlying

flowering and maturity in short-day crops (Zhang *et al.*, 2023). The adaptation of soybean to northern agro-environmental conditions has been driven by a limited number of genes that synchronize the reproductive phenology of the plant to the surrounding northern latitudinal growing conditions (i.e., short growing seasons and long-day photoperiod). Expanding the current cropping range requires a better comprehension of these mechanisms governing the early reproductive traits within Canadian soybean germplasm.

Main research question

What are the regions involved in the regulation of the reproductive and seed quality traits in early-maturing Canadian soybean accessions?

Main hypothesis

A limited number of key genes and molecular mechanisms regulate the reproductive and seed quality traits in early-maturing Canadian soybean accessions.

Main objective

Identify and functionally characterize the genes and molecular mechanisms involved in the regulation of reproductive and seed quality traits in early-maturing Canadian soybean accessions.

Sub-objectives

Chapter 3

- Identify genetic regions and candidate single nucleotide polymorphisms that are involved in the regulation of reproductive traits in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.
- Detect the effect of major QTL regions on the level of expression of major flowering loci.
- Identify loci involved in the regulation of reproductive traits that overlap those previously identified in the $MadMaturity_{86}$ association panel.

Chapter 4

- Develop an expression quantitative trait loci mapping pipeline to identify transcriptome-wide *cis* and *trans* interactions in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

- Develop a hotspot identification pipeline to detect major genomic regions regulating reproductive functions.
- Build and characterize the underlying co-expression networks for key candidate genes.
- Identify single nucleotide polymorphisms within candidate transcription factors located in the identified hotspots.

Chapter 5

- Identify genetic regions and candidate single nucleotide polymorphisms that are involved in the regulation of seed quality traits in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.
- Identify loci regulating seed quality that are in close linkage with regions regulating reproductive traits in the same populations.
- Identify loci involved in the regulation of seed quality traits that overlap those previously identified in the $MadMaturity_{86}$ association panel.

Chapter 6

- Discuss the main findings of this thesis concerning the breeding of earlier maturing soybean varieties adapted to Canadian regions.
- Discuss novel technologies (i.e., genotyping-by-sequencing and *in planta* transformation combined with genome editing strategies) that can contribute to the expansion of soybean's cultivation range further north.

2. Literature review

2.1 Expanding Soybean Cultivation Boundaries

Soybean, *Glycine max* (L.) Merr. is the most important oilseed legume crop cultivated worldwide and it is widely used for human consumption, especially in East and Southeast Asian diets, as well as for animal feed rations and industrial purposes (Pagano and Miransari, 2016). This species is known to be a short-day (SD) plant that was domesticated from wild soybean, *Glycine soja* Sieb. and Zucc., in a region between 30–45°N in China (Guo *et al.*, 2010; Wang *et al.*, 2016a; Sedivy *et al.*, 2017; Zhang *et al.*, 2017c). Although soybean has been a mainstay in crop rotation in East Asian agroecosystems, its cultivation has shifted from Asia to the Americas in the past decades (The American Soybean Association, 2023). As a consequence, the United States (28%), Brazil (37%), and Argentina (16%) now account for more than 80% of the world's production (The American Soybean Association, 2023). Canada's soybean production only accounts for 2% of the total world production as its expansion in acreage is currently restrained due to the short growing seasons and long-day (LD) photoperiod constraints of the Canadian Prairies and northern regions of eastern provinces (Soy Canada, 2022). Although *G. max* is now cultivated between the ~35°S and ~54°N due to the efforts deployed in breeding adapted elite cultivars, its ideal growing conditions are found in middle- or low-latitude regions characterized by warm and humid climatic conditions throughout the growing season (Zhang *et al.*, 2017c).

In North America, soybean cultivars are generally classified into maturity groups, from the earliest maturing, MG000, to the latest, MGX, with an approximate ten-day difference between each group (Bagg *et al.*, 2002). In Canada, the cultivar maturity requirements range from MG000 to MGIII depending on the region (Bagg *et al.*, 2002). Recently, a putative new maturity group (MG0000) has been discovered in Northeast China and the far-eastern region of Russia, demonstrating that it might be possible to expand soybean cultivation further north (Jia *et al.*, 2014; Jiang *et al.*, 2019). However, obtaining high yields in short-season regions remains a challenge due to the short time window for photosynthate accumulation and its subsequent partitioning through translocation during the summer and fall seasons (Van Roekel and Purcell, 2016). This challenge is further amplified by the fact that multiple critical features required for plant yield, such as morphological development, nutritional metabolism, organ senescence, floral induction, pod setting, and seed filling, are directly subordinated to the soybean's internal time-keeping machinery (aka photoperiodism) which has evolved during millions of years under short-day constraints

(Zhang *et al.*, 2017c). Consequently, soybean cultivation in northern latitudes is only possible because breeders have deliberately selected against soybean's natural photoperiodic requirements and have favored alleles allowing for an optimal match between the reproductive phenology of the plant and septentrional environmental conditions (i.e., a short growing season with long days) (Wolfgang and An, 2017).

2.2 Flowering Time Genetic Network in Plants

Floral induction in plants is controlled by a tight interplay between multiple transcriptional gene regulatory networks and pathways involved in the critical role of coordinating the plant's responses to endogenous cues (e.g., plant age, carbohydrate status, and gibberellin levels) and environmental cues (e.g., seasonal change, day length, and climatic conditions) (Fornara *et al.*, 2010; Zhang *et al.*, 2017c). The circadian clock enables plants, as sessile organisms, to anticipate daily and seasonal environmental cycles and to dynamically react to the surrounding stimuli by regulating the timing and pacing of the molecular response at a genome-wide scale (Wilkins *et al.*, 2016). The complex architecture of the clock has been mainly unraveled in the facultative long-day model plant *Arabidopsis* (Fornara *et al.*, 2010), and multiple aspects of its structure remain unclear in other plant species, particularly those with a complex genetic architecture such as polyploids and/or short-day flowering habits (Greenham and McClung, 2015). Although the general architecture of the circadian clock is conserved in plants, it is now widely accepted that the specific composition and roles of key genes differ between species, especially those having undergone genome duplication events during their evolution (Linde *et al.*, 2017). In *Arabidopsis* alone, there are at least six main molecular pathways modulating the floral transition process in response to distinct internal and external cues: (i) vernalization; (ii) autonomous; (iii) gibberellin; (iv) ambient temperature; (v) age; and (vi) photoperiod pathways (Jung *et al.*, 2012; Cheng *et al.*, 2017; Kim, 2020). Molecular analyses have identified at least 306 genes implicated in the control of the flowering process based on loss-of-function (LOF) and transgenic experiments in *Arabidopsis* (Bouché *et al.*, 2016). Due to the complex palaeopolyploid nature of soybean, the networks regulating flowering are even denser, with 844 predicted orthologous genes identified from these 306 genes using bioinformatical evolutionary conservation analyses (Zhang *et al.*, 2017c). Similarly, Jung *et al.* (2012) and Kim *et al.* (2012) have respectively identified 491 putative soybean flowering genes from 186 *Arabidopsis* orthologous flowering genes belonging to 99

groups, and 118 putative orthologous soybean proteins identified from 20 *Arabidopsis* proteins. The high number of genome-wide association study loci (i.e., 379 regions) and biparental QTL regions (i.e., 331 loci) for several traits related to flowering and maturity in the Composite Genetic Map Set archived in Soybase, is another striking evidence of the high level of complexity of this network (Grant *et al.*, 2009). The larger number of predicted flowering genes in soybean is directly linked to a larger genome size (i.e., 1.1-1.15 Gbp in soybean vs. 115-120 Mbp in *Arabidopsis*) caused by two past duplication events that occurred approximately 13 and 59 million years ago (Shultz *et al.*, 2006; Swarbreck *et al.*, 2008; Schmutz *et al.*, 2010). The structure of the soybean genome is highly complex as it underwent aneuploid loss ($n = 10$) followed by a polyploidization ($2n = 20$) and a diploidization ($n = 20$) (Singh and Hymowitz, 1988). As a consequence of these duplication events, a large proportion of the soybean genome is now duplicated, with almost 75% of the genes having multiple copies (Schmutz *et al.*, 2010). Approximately 25% of the soybean genome is considered to be both polyploid and highly conserved (>98% identity) and some of these regions exist in tetraploid or octoploid states (Shultz *et al.*, 2006). Despite this staggering complexity, we now know, based on recent QTL studies coupled with forward and reverse genetic approaches, that there is approximately 25 major loci governing a significant proportion of the soybean flowering and maturity phenotypic variation (Lin *et al.*, 2021b). Over the years, several major loci (e.g., *E1-E11*) associated with reproduction have been thoroughly characterized using forward and reverse genetic approaches (Lin *et al.*, 2021b). Overall, many of these loci (e.g., the ten *J* loci) have been demonstrated to be homologs with overlapping molecular and biological functions (Kong *et al.*, 2010; Lin *et al.*, 2021b). Additionally, many of these genes are known to interact together [e.g., *E1* (*Glyma.06G207800*) regulated by *E3* (*Glyma.19G224200*) and *E4* (*Glyma.20G090000*) photoreceptor genes] (Lin *et al.*, 2022), making the network difficult to decipher.

2.3 Molecular Processes Influencing Flowering and Maturity in Soybean

2.3.1 The Big Four: *E1*, *E2*, *E3*, and *E4*

Maturity loci *E1, E2*, *E3,* and *E4* are frequently reported as the most critical players for short-season maturity and are considered the cornerstones of the latitudinal geographical expansion of this species (Jiang *et al.*, 2014; Fang *et al.*, 2021b). As such, these four loci have become the pillars of short-season breeding programs although additional novel loci (e.g., *E9*, *E10*, and *Tof*

5/11/12//16/18) are slowly being implemented (Kong *et al.*, 2014; Samanfar *et al.*, 2017; Lu *et al.*, 2020; Dong *et al.*, 2021, 2022; Kou *et al.*, 2022). Loss-of-function variants in *E1-E4* contribute to photoperiod insensitivity along with the promotion of flowering under long-day conditions by repressing the expression of *FT* orthologs, such as *GmFT2a*/*E9* (*Glyma.16G150700*) and *GmFT5a* (*Glyma.16G044100*) (Buzzell and Voldeng, 1980; Saindon *et al.*, 1989; Cober *et al.*, 1996; Thakare *et al.*, 2011; Watanabe *et al.*, 2011; Xia *et al.*, 2012; Lu *et al.*, 2015; Zhang *et al.*, 2016b). The importance of these loci is such that proper haplotype combination at the *E1-E4* loci can explain more than 60 % of the variation in the observed flowering time (Liu *et al.*, 2008; Xia, 2013). In addition, it has been demonstrated that there is a high correlation between the latitudinal adaptability/photoperiod insensitivity and the number of recessive alleles for these four *E* loci (Jiang *et al.*, 2014).

E1 (*Glyma.06G207800*) is a major maturity gene encoding a legume-specific B3-like transcription factor (TF) that has two homologs, *E1-Like-a* (*E1La*; *Glyma.04G156400*) and *E1-Like-b* (*E1Lb*; *Glyma.04G143300*) (Xu *et al.*, 2013, 2015). The gene is known to be present in other legume species, such as *Medicago truncatula* and *Phaseolus vulgaris*, but its functions do not seem to be highly conserved (Zhang *et al.*, 2016b). In soybean, the role of *E1* is multifaceted with crucial functions in the modulation of plant architecture, initiation of terminal flowering, and photoperiod response (Wan *et al.*, 2022). In addition, *E1* regulates the expression of *Dt1* and *Dt2*, two important genes playing a key role in stem determination (Wan *et al.*, 2022). The dominant *E1* allele is mainly found amongst the maturity groups MGII-MGVII (Liu *et al.*, 2020), whereas the recessive *e1-nl* (null), *e1-fs* (frameshift with premature stop codon), and *e1-as* (amino acid substitution) are primarily found amongst MG0000-MGI cultivars (Liu *et al.*, 2020). The recessive forms *e1-nl* and *e1-fs* have been demonstrated to have the greatest impact on maturity and are only found within MG0000 and MG000 accessions from Northeastern China (Liu *et al.*, 2020). Additional natural *E1* variants contributing to the northern expansion of soybean include *e1-re* (retrotransposon insertion), *e1-p* (allele from the cultivar Peking), and *e1-b3a* (mutation in B3 domain) (Xia *et al.*, 2012; Tsubokura *et al.*, 2014; Liu *et al.*, 2020). The *E1La* and *E1Lb* homologs negatively regulate flowering and consequently play functions that are overlapping or redundant to those of *E1* (Liu *et al.*, 2022a). In a similar fashion to *E1*, the *e1la:K82E* variation (i.e., an *E1La* variation) has also been demonstrated to be an important contributor to the northern geographical expansion of soybean cultivation (Dietz *et al.*, 2021). In North America, the *e1la:K82E* haplotype is only found

in accessions cultivated in Canada and the northern regions of the United States (i.e., mean latitude between 48.8-49.8°N), whereas the dominant *E1La* allele is distributed both in northern and southern areas with a mean latitude between 37.8-42.1°N (Dietz *et al.*, 2021). Similar observations have been made with the recessive allelic forms of the *E1Lb* homolog as near-isogenic lines (NILs) for the *e1lb* variant have been demonstrated to flower earlier under long-day conditions in the *e3/E4* and *E3/E4* backgrounds through an increase in expression of *GmFT2a*/*E9* and *GmFT5a* (Zhu *et al.*, 2019). Moreover, it also has been demonstrated that the recessive allele *e1lb:Del* promotes early flowering, but only in the partially functional *e1-as* background (Dietz *et al.*, 2021).

E2 (*Glyma.10G221500*; *GmGIa*) is the ortholog of the *Arabidopsis* gene *GIGANTEA* which has two homologs in the soybean genome, *GmGI1a* (*Glyma.20G170000*; also known as *E2-Like a/E2La*) and *GmGI2* (*Glyma.16G163200*; also known as *E2-Like b/E2Lb*) (Watanabe *et al.*, 2011; Li *et al.*, 2013). The homolog *GmGI1a* shares a high level of sequence similarity (i.e., 97%) with *E2*, thus suggesting that these genes originate from lineage-specific duplication (Watanabe *et al.*, 2011). In *Arabidopsis*, the leading photoperiodic flowering pathway consists of the *GIGANTEA-CONSTANS-FLOWERING LOCUS T* module (Sawa and Kay, 2011). Under LD, *E2* contributes to this pathway by inhibiting the expression of *GmFT2* which results in a delayed flowering phenotype (Watanabe *et al.*, 2011). In counterpart, the recessive *e2/e2* genotype, a mutant with a premature stop codon, displays an increased level of expression of this florigen (Watanabe *et al.*, 2011). Wang *et al.* (2023) have recently demonstrated that single *e2* mutants generated with Clustered Regularly Interspaced Short Palindromic Repeats / CRISPR-associated protein 9 (CRISPR-Cas9)-mediated gene editing flower 8-10 days earlier than the wild-type accession, whereas *e2la* and *e2lb* single mutants did not display any modifications in their flowering. Three main *E2* haplotypes, designated in the literature as haplotypes H1, H2, and H3, are found naturally in cultivated and wild soybean along with 44 additional haplotypes that are present only in accessions from the wild (Wang *et al.*, 2016b). The H2 and H3 haplotypes encode full-length protein isoforms, whereas the H1 haplotype, also named *e2-ns*, encodes a protein with a premature stop codon associated with a single-base substitution located in the 10th exon (Wang *et al.*, 2016b). The H2 haplotype is found only in the Southern region of China, whereas H3 is restricted to regions near the Northeastern region (Wang *et al.*, 2016b). The *e2-ns* (H1) haplotype is derived from H2 and broadly distributed amongst cultivated accessions; however, its occurrence in the wild is low and limited to the Yellow River basin (Wang *et al.*, 2016b). In China, the *e2-ns* (H1) variant is

found within maturity groups MG0000-MGVI, whereas MGVII and MGVIII only display the dominant form of *E2* (Liu *et al.*, 2020). As a whole, this natural variant seems to be required for early maturity as none of the MG0000-MG00 Chinese accessions harbor the *E2* dominant form (Liu *et al.*, 2020).

E3 (*PHYTOCHROME A3/GmPHYA3*; *Glyma.19G224200* and *E4* (*PHYTOCHROME A2/GmPHYA2*; *Glyma.20G090000*) are photoreceptors promoting the expression of the flowering repressor *E1* and modulating the activity of *J*/*E6* (*Glyma.04G050200*), a core component of the circadian evening complex (Qin *et al.*, 2023a). The *E3* and *E4* genes interact with the EMPFINDLICHER IM DUNNKELROTEN LICHT 1 (EID1; *Glyma.03G214300*) protein upon light activation (Qin *et al.*, 2023a). This action subsequently results in a reduction in the abundance of the J/E6 protein under LD conditions due to the inhibition of EID1–J interaction (Qin *et al.*, 2023a). In addition, Qin *et al.* (2023a) demonstrated that *GmEID1* inhibits the transcription of *E1* and suggested that *E3* and *E4* might be acting as competitive inhibitors of the EID1–J interaction, thus leading to a possible increase in the expression of *E1*. As a result of their actions on the *E1* gene and J/E6 protein, *E3* and *E4* promote floral transition under LD conditions, a factor contributing to the northern geographical expansion of soybean's cultivation range (Cober *et al.*, 1996; Tsubokura *et al.*, 2013; Xu *et al.*, 2013; Qin *et al.*, 2023a). The *E3* and *E4* genes are also known to influence various soybean post-flowering functions, such as plant maturation and stem termination, by indirectly favoring the transcription of *Dt1* (Xu *et al.*, 2013). In *Arabidopsis*, the ortholog *AtPHYA* is the only photoreceptor that can perceive far-red light and several underlying physiological processes related to far-red light perception, such as de-etiolation (i.e., the process of converting etioplasts – non-green plastids - within cells into chloroplasts), are controlled by this gene (Liu *et al.*, 2008; Watanabe *et al.*, 2009; Rausenberger *et al.*, 2011). In pea, *PsphyA* and *PsphyB* double mutants demonstrate growing malformations (e.g., reduced rate of leaf production and twisted internodes) and a severely dysfunctional de-etiolation process (Weller *et al.*, 2001).

The *E3* locus modulates the flowering response in relation to the red-to-far–red (R:FR) quantum ratio of light (Buzzell, 1971; Cober *et al.*, 1996). As such, the dominant *E3* allele delays flowering based on a fluorescent-sensitive response, whereas the recessive *e3* allele provides an earlier maturity resulting from an insensitive response (Buzzell, 1971; Cober *et al.*, 1996). Three natural variants of *E3* (*e3-fs,* frameshift; *e3-ns*, non-sense; and *e3-tr,* large deletion after the third exon) have been identified within the earlier maturity groups MG0000-MGIII, whereas the later

maturity groups MGIV-MGVIII only harbor the dominant *E3* allele (Liu *et al.*, 2020). Several other recessive alleles, such as *e3-Mo* (allelic variations from the cultivar Moshidou Gong 503), *E3-Mi* (allele from the cultivar Misuzudaizu), and *E3-Ha* (allele from the cultivar Harosoy) are available for breeding and have been also identified in Chinese soybean cultivars (Watanabe *et al.*, 2009; Tsubokura *et al.*, 2014; Liu *et al.*, 2020). The *E3-Ha* or *E3-Mi* alleles were identified in 65.9% of the Chinese cultivars and the *e3-tr* allele in 31.2% of the studied germplasm (Liu *et al.*, 2020). In comparison, the recessive alleles *e3-fs* and *e3-ns* were only identified in 2.9% of the Chinese cultivars, all of which are characterized by early pod filling and maturity (Liu *et al.*, 2020).

The *E4* gene is a critical player in the adaptation of soybeans to high-latitude agro-environments as mutations associated with this locus reduce photoperiod sensitivity to long day length (Tsubokura *et al.*, 2013; Xu *et al.*, 2013). The *E4* locus was initially identified using incandescent lamps with a natural day length extended to 20 hours/day (Buzzell and Voldeng, 1980). The homozygous *e4e4* genotype is required for *e3e3* homozygous plants to flower under incandescent day-length conditions (Buzzell and Voldeng, 1980; Saindon *et al.*, 1989). Furthermore, it was demonstrated that *E1* delays flowering in combination with *e3* and *e4* under incandescent long day length (Cober *et al.*, 1996). The main mutations of *E4* are *e4-kes* (allele from cultivar Keshuang), *e4-kam* (allele from cultivar Kamaishi-17), *e4-oto* (allele from cultivar Otomewase), and *e4-tsu* (allele from cultivar Tsukue-4) (Tsubokura *et al.*, 2013) along with *e4-SORE* (insertion of Ty1/copia-like retrotransposon SORE-1 of size 6,238 bp in exon #1) (Kanazawa *et al.*, 2009). The *e4-kes*, *e4-kam*, *e4-oto,* and *e4-tsu* mutations harbor single-base deletions in exon #1 or exon #2 which generate frameshifts resulting in premature stop codons and a subsequent truncation in their synthesized proteins (Tsubokura *et al.*, 2013). The distribution of these five dysfunctional variants (*e4-kes*, *e4-kam*, *e4-oto*, *e4-tsu,* and *e4-SORE)* is limited to small geographical regions of Eastern Asia in cultivated and wild soybean accessions (Kanazawa *et al.*, 2009), thus suggesting recent and independent domestication events for these alleles. This claim is largely supported by a genotyping survey of 308 soybean cultivars originating from China which demonstrated that a significant proportion (96.8%) of the accessions harbored the wild type *E4* allele and only a small portion, respectively 1.3% and 1.9%, were *e4-SORE* and e4-kes alleles (Liu *et al.*, 2020). In addition to its role in photoperiodic flowering, *e4* is also known to improve chilling tolerance (about 15 °C) during flowering, a feature that limits the browning and cracking of the

seed coats (Matsumura *et al.*, 2008). As such, e4 is used in breeding to develop early-maturing cultivars with chilling tolerance.

Altogether, *E1, E2, E3,* and *E4* are the most crucial genetic components underlying the flowering and maturity processes in soybean. Their importance is such that several breeding programs focus solely on the different allelic combinations of these genes to understand the latitudinal distribution of soybean. Overall, the big four play a critical role in the distribution of soybean across China as mainly cultivars featuring mutations at all four loci are distributed in the northeastern areas located above the 47° N latitudinal parallel (Liu *et al.*, 2020; Zhang *et al.*, 2021b). In addition, several rare alleles of *E1* (e.g., *e1-nl* and *e1-fs*), *E3* (e.g., *e3-fs* and *e3-ns*), and *E4* (*e4-SORE* and *e4-kes*) have recently been identified in very early-maturing Chinese cultivars belonging to the maturity groups MG0000-MG00, confirming that rare mutations in these genes are drivers for northern latitudinal expansion (Liu *et al.*, 2020).

2.3.2 Supporting Acts for Flowering and Maturity

Throughout the years, several other loci (e.g., *E6, E9,* and E10) have slowly been implemented in breeding programs although their effects on flowering and maturity are generally less pronounced than *E1-E4* (Ray *et al.*, 1995; Bonato and Vello, 1999; Cober, 2011; Kong *et al.*, 2014; Samanfar *et al.*, 2017). Soybean research has identified about 25 genes and loci that have a peripheral action to *E1*, *E2*, *E3,* and *E4*, including *E5-E11*, *Time of flowering* (*Tof*) *5*/*11*/*12*/*16* and ten *J* homologs (Table 1.1) (Lin *et al.*, 2021b; Lv *et al.*, 2022). The *E5* locus was originally detected in a biparental cross derived from 'Harosoy' (*e5*; MGII) × 'PI 80837' (*E5*; MGIV) (Mcblain and Bernard, 1987), but a subsequent study from Dissayanaka *et al.* (2016) using the original cross failed to identify this QTL. This locus is located near *E2* on chromosome GM10 and it was suggested that *E5* might be an allele of *E2* (Watanabe *et al.*, 2011). To explain the discrepancies between the studies, Dissayanaka *et al.* (2016) suggested an unexpected outcrossing with an accession featuring an *E2-dl* allele might have contaminated the original cross in McBlain & Bernard (1987). However, it was recently demonstrated that an *E5* NIL exhibits a specific molecular signature at the transcriptome level, with 255 and 1,802 genes differentially expressed under non-flowering inductive long days and photoperiodic shift conditions, respectively (Wu *et al.*, 2019). Based on these observations, Wu *et al.* (2019) concluded in the existence of a unique *E5* locus and suggested that it might play a role in photoperiodic adaptation to seasonal changes.

Recently, the qMG-10.3 QTL region was reported for this locus, and two genes, *Glyma.10G251200* and *Glyma.10G284500*, with known impacts on flowering in *Arabidopsis* orthologs, were suggested as candidates (Zimmer *et al.*, 2021).

The *E6* (*Glyma.04G050200*) locus promotes flowering and is an ortholog of the *EARLY FLOWERING 3* (*AtELF3*) gene, a key component of the circadian Evening Complex in *Arabidopsis* (Fang *et al.*, 2021a). Its recessive form (*e6e6*) is known to confer the long-juvenile phenotype, a trait characterized by the inhibition of the flowering process and a prolongation of the reproductive and vegetative stages (Ray *et al.*, 1995; Bonato and Vello, 1999). The *E6* locus was identified in the MGVI cultivar Paraná due to naturally occurring mutations causing flowering inhibition (Bonato and Vello, 1999). Recent molecular analyses have revealed that *E6* and *J*, two loci both located on chromosome GM04, actually are the same locus (Nissan *et al.*, 2021). Sequencing of *Glyma.04G050200*, the *J* gene, identified that the *e6* line 'Paranagoiana' harbors a Ty1-copia retrotransposon of ~10,000 bp within exon 4, a variation now called *j-x* (Nissan *et al.*, 2021) or $e6^{PG}$ (Fang *et al.*, 2021a). This recessive mutation does not suppress the expression of *E1* by binding to its promoter and consequently allows *E1* to inhibit the expression of *GmFT2a*/*E9* and *GmFT5a* (Fang *et al.*, 2021a). Ultimately, this results in a delayed flowering process and increased yields under short-day conditions (Fang *et al.*, 2021a). Consequently, the dominant *E6* allele is required for northern adaptation as the impairment of *E6*/*J* leads to a better adaptation to tropical photoperiodic regimes (Fang *et al.*, 2021a). On the whole, the $e6^{PG}$ allele is not commonly used in modern soybean selection programs (Fang *et al.*, 2021a). An additional eight alleles, *j-1* to *j-8*, were recently discovered and demonstrated to be restricted to accessions cultivated in low-latitude regions (Lu *et al.*, 2017). Four of these alleles (*j-1*, *j-2*, *j-4*, and *j-5*) were selected during modern soybean breeding in North America, whereas four others (*j-3*, *j-6*, *j-7*, and *j-8*) were identified in landraces from Asia (Lu *et al.*, 2017; Fang *et al.*, 2021a).

The *E7* locus was originally detected because of an association between tawny pubescence, and an early-maturing phenotype in soybean lines belonging to MG000-II (Cober and Voldeng, 2001). *E7* is located on chromosome GM06, at approximately 6 cM from *E1*, between the molecular markers Satt100 and Satt460 in a region covering a total of 22.2 cM (Molnar *et al.*, 2003). The dominant form of the *E7* locus delays flowering, whereas its recessive allele flowers 6-7 days earlier with an R:FR light quality similar to natural daylight under a 20h photoperiod regime and 20 days earlier with a low R:FR light quality (Cober and Voldeng, 2001). Based on a

bioinformatic study, three genes *Glyma.06G200400* (*EXOCYST SUBUNIT EXO70 FAMILY PROTEIN B1*), *Glyma.06G200800* (*SWEET17*), and *Glyma.06G220000* (*REDUCED VERNALIZATION RESPONSE 1*), have been proposed as potential candidates for this locus (Pattang, 2022).

In a similar fashion to the *E7* locus, the dominant *E8* allele (*E8E8*) is known to delay maturity, whereas the recessive allele (*e8e8*) imparts an earlier maturity of about 5-8 days (Cober *et al.*, 2010). Cober *et al.* (2010) observed that isolines with the 'Maple Presto' and 'Harosoy' backgrounds and containing the recessive *e8e8* allele matured respectively nine and six days earlier than those with the dominant form. The *E8* locus was originally mapped between Sat_404 and Satt136 on chromosome GM04 (Cober *et al.*, 2010) and three subsequent studies have mapped broad QTL regions that could harbor *E8*: (i) GM04:13,212,370–43,843,500 bp (Kong *et al.*, 2018); (ii) GM04:7,166,748-44,508,948 bp for (Wang *et al.*, 2018); and (iii) *qRP-c-1* QTL region located between the Sat_085 and Satt294 flanking markers (Cheng *et al.*, 2011). On the whole, this large region covers approximately 37.5 Mbp on chromosome GM04 (~7 Mbp-44.5Mbp) and comprises multiple predicted or validated flowering genes such as *E1-like-a* (*Glyma.04G156400*), *E1-like-b* (*Glyma.04G143300*) and *CRYPTOCHROME 1A* (*Glyma.04G101500*). Sadowski (2020) investigated the region *in silico* and proposed seven genes as potential candidates for this locus: (i) *Glyma.04G101500* [GM04:9,337,214; also proposed by Cheng *et al.* (2011) for the *qRP-c-1* QTL region]; (ii) *Glyma.04G111200* (GM04:12,284,839); (iii) *Glyma.04G124300* (GM04:16,092,707); (iv) *Glyma.04G124600* (GM04:16,329,983); (v) *Glyma.04G126000* (GM04:16,811,865); (vi) *Glyma.04G138900* (GM04:21,797,427); and (vii) *Glyma.04G140000* (GM04:22,755,516).

The *E9* locus was originally mapped to a 245 kb region on chromosome GM16 between markers M5 and M7 using a backcross performed between 'Tokei 780' and two early-flowering recombinant inbred lines with identical *E1* to *E4* alleles (Kong *et al.*, 2014). The segregation pattern that was observed in the F_2 and F_3 progenies of this backcross demonstrated that the early flowering phenotype was controlled by *E9* which acts as a single dominant locus (Kong *et al.*, 2014). On the contrary to *E7* and *E8*, the recessive form of *E9* (*e9e9*) induces late maturity (Kong *et al.*, 2014). Subsequent fine mapping, sequencing, and expression analysis have revealed that *E9* is an ortholog of the *Arabidopsis FT2a* gene (Zhao *et al.*, 2016). Zhao *et al.* (2016) demonstrated that the *e9* allele has a *Ty1*/*copia–like* retrotransposon (SORE-1) inserted in its first intron which results in a severe decrease in *e9* transcription through allele-specific transcriptional repression.

The *E10* locus was identified at the end of chromosome GM08 using early-maturing soybean materials belonging to MG000-MG0 (Samanfar *et al.*, 2017). Its recessive form (*e10e10*) imparts an earlier maturity of about 5 to 10 days (Samanfar *et al.*, 2017). Protein-protein interaction prediction identified *Glyma.08G363100* (*GmFT4*) as the best candidate gene from a list of 75 genes and subsequent variant analysis identified three potentially deleterious SNPs located in the 5'UTR, 3'UTR, and fourth exon of this gene (Samanfar *et al.*, 2017). The expression of *GmFT4* is strongly upregulated in soybean *E1*-overexpressing mutants, thus suggesting that his gene acts downstream of *E1* as a flowering repressor (Zhai *et al.*, 2014). Moreover, it has been suggested that the balance between *GmFT4* and *GmFT2a*/*5a* (i.e., two antagonistic flowering factors) is a key component in the determination of the flowering time in soybean (Zhai *et al.*, 2014). The *E11* locus was detected in an ~ 1.03 Mbp region located on chromosome GM07 in a biparental cross between 'Archer' (MGI) and 'Minsoy' (MG0) (Wang *et al.*, 2019). Using NILs, it has been demonstrated that the plants carrying the recessive *e11e11* allele flower 9-10 later under field conditions than those with the dominant form (Wang *et al.*, 2019). Through amino acid sequence analysis, the *Glyma.07G048500* gene, an ortholog of the *Arabidopsis* circadian-clock gene *LATE ELONGATED HYPOCOTYL* (*LHY*), was proposed as the best candidate for this locus followed by *Glyma.07G049000* (*PLASTID TRANSCRIPTIONALLY ACTIVE 8*) and *Glyma.07G049200* (*METAL TOLERANCE PROTEIN A2*) (Wang *et al.*, 2019).

TIME OF FLOWERING 5 (*Tof5*) is a locus that promotes flowering and improves the adaptation of this species to high latitudes (Dong *et al.*, 2022). Using CRISPR-cas9 mutagenesis and GWA analysis, the locus has been demonstrated to be an ortholog of the *Arabidopsis FRUITFULL* (*AtFUL*) gene (Dong *et al.*, 2022). Through parallel selection, multiple *Tof5* alleles, such as *Tof5H1* and *Tof5H2*, have contributed to the expansion of the geographic distribution of soybean to high-latitude regions (Dong *et al.*, 2022). *Time of Flowering 11* (*Tof11*) and *Time of Flowering 12* (*Tof12*) are two loci coding for homologous *PSEUDO RESPONSE REGULATOR* genes (Lu *et al.*, 2020). The loci delay flowering under long photoperiods by promoting the expression of *E1* via the repression of four *LHY* homologs (Lu *et al.*, 2020). In addition, studies using *E3* and *E4* NILs have confirmed that both of these loci positively regulate *Tof11* and *Tof12* (Lu *et al.*, 2020). Loss of function of *Tof11* and *Tof12* during domestication has been associated with a gradual expansion of the northern soybean cultivation limits (Lu *et al.*, 2020). *Time of Flowering 16* (*Tof16*) delays flowering and improves soybean yield in low-latitude agricultural

areas (Dong *et al.*, 2021). Functional validation using CRISPR-cas9 mutagenesis has validated *LATE ELONGATED HYPOCOTYL* (*LHY*) as the causative gene for this locus (Dong *et al.*, 2021). Also, it has been demonstrated that *Tof16* and *E6*/*J* have cumulative effects on soybean flowering, resulting in 80% of the accessions adapted for low-latitude cultivation zones harboring recessive mutations for both of these genes (Dong *et al.*, 2021). For both genes, the delaying effects on soybean flowering are guaranteed through the binding of *Tof16* and *E6*/*J* to the promoter region of *E1* (Dong *et al.*, 2021). Altogether, these loci are supporting acts to the big four and their effects on the observed phenotype are generally smaller.

2.4 Finding the Transcription Factors and Their Targets

2.4.1 Role of Transcription Factors in the Regulation of Reproductive Traits

Transcription factors are central proteins regulating the expression of multiple downstream genes by binding with their DNA-binding domain (DBD) to the *cis*-regulatory elements (CRE) located in the promoter or enhancer regions of their targets (Bylino *et al.*, 2020). As such, TFs are the main regulators of gene transcription and thus play a critical role in gating the initiation and regulation of gene transcription in response to stimuli and sustaining the function of RNA polymerase at transcription sites in eukaryotes (Mitsis *et al.*, 2020). The functions of TFs comprise two basic features: (i) the identification and binding of short, specific DNA sequences; and (ii) the capacity to recruit or bind proteins involved in the regulation of transcriptional activities (Mitsis *et al.*, 2020). In addition to their DBD, TFs harbor other domains involved in a variety of functions such as protein-protein interaction, activation/repression of gene expression, and dimerization (Gonzalez, 2016). In eukaryotes, many TFs recruit transcriptional cofactors to modify the chromatin structure and facilitate the assembly of the pre-initiation complex formed by general TFs and RNA Polymerase II (Venters and Pugh, 2009).

Transcription factors use a broad variety of DNA-binding structural motifs to recognize the CREs of their targets (e.g., homeodomain) and can be classified into different families using this criterion. In soybean, at least 6,150 TFs (3,747 loci) belonging to 57 families have been predicted (Jin *et al.*, 2017), including a large number with validated circadian clock, flowering, and/or maturity functions [e.g., *E1* (Xia *et al.*, 2012), *E1-like-a* (Liu *et al.*, 2022a), *E1-like-b* (Zhu *et al.*, 2019), *GmLHY1a/1b/2a/2b* (Bian *et al.*, 2017), *GmFT7*/*GmFT2c* (Zhang *et al.*, 2021a), *GmNMHC5* (Wang *et al.*, 2020b), *GmTOE4b*/*Tof13* (Li *et al.*, 2023b), *GmGAMYB* (Yang *et al.*,

2021a), *Dt* genes (Ping *et al.*, 2014), and *GmIDD* (Yang *et al.*, 2021b)]. Based on their sizes, the most important TF families found in soybean are MIKC MADS (40 members), bHLH (31 members), B3 (24 members), G2-like (22 members), Dof (18 members) and GRAS (18 members) (Jin *et al.*, 2017). A recent large-scale GWAS using a large diversity panel consisting of 1,503 early-maturing soybean lines has recently confirmed the interplay between numerous TFs (*E1*, *E2*, *Dt1*, *Dt2*, and *GmAPETALA1d*) associated with QTL regulating the early-flowering time and early-maturity traits (Vollmann and Škrabišová, 2023; Zhu *et al.*, 2023). Often, these TFs are involved in specific gene regulatory networks (GRNs) involving a plethora of other TFs that regulate specific functions in the plant, such as in the case of *Dt1* (Hou *et al.*, 2022). The *Dt1* gene is expressed under LD photoperiodic conditions, but not under SD (Hou *et al.*, 2022). When soybean plants are under LD photoperiodic conditions, the expression of *Dt1* is induced by *E3* and *E4* which control photoperiod insensitivity (Xu *et al.*, 2013). When *Dt1* is expressed, it interacts with the FDc1 bZIP TF to form the FDc1-Dt1 complex (Yue *et al.*, 2021). Subsequently, the FDc1-Dt1 complex binds to the *APETALA 1* (*GmAP1*) promoter region to repress its expression (Yue *et al.*, 2021). The FT5a TF (a member of the phosphatidylethanolamine-binding protein family) upregulates *GmAP1* by inhibiting the activity of Dt1 through competitive interaction with FDc1, thus interfering with the Dt1-AP1 feedback loop (Yue *et al.*, 2021). The FT2a and FT5a TFs redundantly and differentially regulate flowering by upregulating the *GmFDL19* gene which codes for a bZIP TF that can bind to the ACGT *cis*-regulatory element of the *GmAP1* promoter (Yue *et al.*, 2021). As indicated in the model developed by Lin *et al.* (2021b), *GmFT2a* and *GmFT5a* are indirectly regulated by *E1lb*, whereas *GmFT1a* and *GmFT4* are directly regulated by *E1* (Fig. 1.1). This example clearly illustrates the complexities and densities of the GRNs involved in the regulation of reproductive traits in soybean. Unraveling these GRNs is challenging on an experimental basis due to the transient action of the TFs and the large scale of these interactions as TFs can interact with hundreds of genes. One particular approach to identifying these transient interactions on a genome-wide basis is to map the expression of quantitative trait loci and hotspots that exhibit core regulatory activities.

2.4.2 Mapping Expression Quantitative Trait Loci in Plants

Expression quantitative trait loci (eQTL) are chromosomal regions regulating the abundance of mRNA transcripts in genetic mapping populations (Druka *et al.*, 2010). By

unraveling thousands of eQTL interactions, it is possible to model gene regulatory networks and identify critical hotspots underlying the regulation of specific phenotypes (Druka *et al.*, 2010). Regulatory hotspots are loci involved in the regulation in expression of a large number of genes located in *trans* and often underlie genes coding for TFs (Neto *et al.*, 2012). On a methodological basis, the steps associated with the mapping of eQTL traits are the same as for regular traits (e.g., seed quality or reproductive traits) and can be performed either in biparental populations for linkage mapping or diversified populations for association mapping. Expression QTL interactions are often categorized into either *cis*/local or *trans*/local based on the respective locations of the regulator and regulated loci (Liu *et al.*, 2022b). At present, there is no consensus on the way to distinguish *cis* from *trans* in the literature and the filters used to do so vary from one article to another. For example, Shan *et al.* (2019) classified the interactions located within ≤ 1 Mbp on the same chromosome as *cis*, whereas all the others are categorized as *trans*. In their article on the eQTL architecture of immature soybean seed, Bolon *et al.* (2014) classified as *cis* all of the interactions located within ≤ 1.575 Mbp of the physical location of the SNP marker near the targeted gene. In Zhang *et al.* (2020b) and Li *et al.* (2018), eQTL interactions were considered *cis*/local-acting when the associated QTL region spanned the target gene on the same chromosome, whereas all the others were considered *trans*/distant when located outside of the target gene on the same chromosome or a different one. Although this classification between *cis* vs. *trans* can seem futile in some aspects, it can bear significant consequences on the results of a study as regulatory hotspots are often identified using only *trans*-acting interactions (Zhang *et al.*, 2017b, 2020a; Tan *et al.*, 2022). Importantly, it is important to keep in mind that the *cis* or *trans*-acting effects of these eQTL interactions have yet to be demonstrated and that the terms local or distant are consequently more appropriate (Rockman and Kruglyak, 2006); however, the familiar terms *cis* and *trans* are much more common in the literature in comparison to the more precise local and distant terms and are typically used for clarity's sake.

Genome-wide eQTL mapping is a relatively new technique, with the first study carried out in yeast and published in 2002 (Brem *et al.*, 2002). Over the years, the increasing affordability of RNA sequencing technology has supported the development of eQTL mapping, but this approach remains scarce in non-model plant species due to their complex and large genomes. As such, only a handful of studies have conducted genome-wide eQTL mapping in soybean (Bolon *et al.*, 2014; Li *et al.*, 2023a), although a few others have performed analyses with a limited number of genes

(e.g., Wang *et al.*, 2014b). To increase the affordability of genome-wide eQTL mapping, researchers often use smaller population sizes (<100 individuals) (Druka *et al.*, 2010; Zhang *et al.*, 2020a). In the past, several studies have demonstrated that obtaining meaningful results is feasible using this approach (Druka *et al.*, 2010; Zhang *et al.*, 2020a). According to Druka *et al.* (2010), there is no convincing statistical justification that dismisses the use of these smaller populations as gene expression traits generally have high heritabilities, an important feature that consequently ensures a high level of power to detect genetic variants. To do so, significance levels must be set such that genome-wide false-positive detection rates are low (generally below 5%) which can limit the number of eQTL signals identified during the mapping phase, but provide an accurate picture for 95 % of the interactions that are deemed as true-positives. As *cis*-acting interactions are most often easier to map on a statistical basis due to their larger impact on the phenotypic variation in comparison to *trans*-acting interactions, a smaller population should normally identify a higher ratio of *cis* vs. *trans* interactions than in a larger population (Druka *et al.*, 2010). Over the years, numerous mapping approaches, such as Genome-wide composite interval mapping (GCIM) (Zhang *et al.*, 2020b), have been proposed as a statistical solution for the identification of small-effect eQTL typically associated with *trans*/distant interactions and the reduction of the computational burden associated with transcriptome-wide mapping in large biparental populations (Westra and Franke, 2014; Wen *et al.*, 2020). To reduce the computational requirements of such studies, eQTL mapping in biparental populations generally relies on the Interval Mapping algorithm or simple tests (such as the Wilcoxon–Mann–Whitney) for linkage analysis due to their higher computational speed (Zou and Zeng, 2009), although some other approaches such as Multiple interval mapping (Zou and Zeng, 2009) and Composite interval mapping (Bolon *et al.*, 2014) have been also used.

2.4.3 Mapping of eQTL Traits in Biparental Soybean Populations

As previously detailed, expression QTL studies can be divided into two categories based on their respective population: (i) diversified association panel for GWA (also named TWAS; transcriptome-wide association studies); or (ii) biparental population for linkage mapping. The largest number of recently published high-quality genome-wide eQTL studies in soybean were using large diversified populations (Zhang *et al.*, 2017b, 2020a; Tan *et al.*, 2022; Li *et al.*, 2023a), whereas a lesser number of publications were based on biparental populations (Bolon *et al.*, 2014;

Li *et al.*, 2018). To the best of our knowledge, only one genome-wide eQTL study has been published in soybean using biparental populations (Bolon *et al.*, 2014). In Bolon *et al.* (2014), a 1,536 SNP linkage map with 24 linkage groups was generated to identify eQTL interactions underlying flavonoid biosynthesis, photosynthesis, and fatty acid biosynthetic in a RIL population generated from the 'Minsoy' × 'Noir 1' biparental cross. In this study, the global threshold was determined based on 100,000 maximum likelihood ratio test statistics derived from 1,000 permutations on 100 randomly selected gene expression traits, and traits were analyzed using the Composite interval mapping algorithm. The number of hotspots was determined in this study using a threshold calculated based on the 95th percentile of the maximum number of eQTL traits detected at any given locus and corresponded to 39 eQTL interactions per genetic locus. In this study, many of the mapped hotspots were in adjoining positions and those were considered as part of the same hotspots, thus yielding a total of 59 hotspots. Based on this approach, the researchers were able to identify many candidate TFs involved in flavonoid biosynthesis genes and seed pigmentation. On a smaller scale, the mapping of interactions associated with specific genes using qPCR was also performed in a handful of other populations: (i) mapping of interactions associated with isoflavone content in a RIL population comprising 130 individuals (Wang *et al.*, 2014); (ii) mapping of interactions associated with *RUBISCO ACTIVASE* genes in a RIL population with 184 lines (Yin *et al.*, 2010); and (iii) mapping of interactions associated with the tocopherol biosynthetic pathway in a population comprising 144 individuals (Sui *et al.*, 2020). The results obtained in these studies suggest that it might be possible to use this approach to identify regulatory hotspots underlying reproductive phenotypes on a genome-wide scale in a similar fashion than for these leaf and seed-related traits. Combined with other strategies (e.g., co-expression analysis and regular QTL analysis), eQTL mapping might provide sufficient information to locate candidate TFs regulating transient transcriptional interactions for a large number of genes involved in key traits.

Loci	Gene	Accession number	Function	Reference
E1	*E1*	*Glyma.06G207800*	Delays flowering	Xia *et al.* (2012)
	E1La	*Glyma.04G156400*	Delays flowering	Xu *et al.* (2015)
	E1Lb	*Glyma.04G143300*	Delays flowering	Watanabe *et al.* (2011); Xu *et al.* (2015)
E2	*GmGI*	*Glyma.10G221500*	Delays flowering	Watanabe *et al.* (2011)
E3	*GmPHYA3*	*Glyma.19G224200*	Delays flowering	Watanabe *et al.* (2009)
E4	*GmPHYA2*	*Glyma.20G090000*	Delays flowering	Liu *et al.* (2008)
	GmPHYA1	*Glyma.10G141400*	Unknown	Liu *et al.* (2008)
E5	Debated	Debated	Debated	Dissanayaka *et al.* (2016); Wu *et al.* (2019)
E6	Same as *GmELF3*	*Glyma.04G050200*	Promotes flowering	Nissan *et al.* (2021)

Loci	Gene	Accession number	Function	Reference
E7	Unknown	Not validated	Delays flowering	Cober and Voldeng (2001)
E8	Unknown	Not validated	Delays flowering	Cober *et al.* (2010)
E9	*GmFT2a*	*Glyma.16G150700*	Promotes flowering	Kong *et al.* (2010, 2014); Zhao *et al.* (2016)
E10	*GmFT4*	*Glyma.08G363100*	Delays flowering	Zhai *et al.* (2014); Samanfar *et al.* (2017)
J	*GmELF3*	*Glyma.04G050200*	Promotes flowering	Lu *et al.* (2017); Yue *et al.* (2017)
	GmFT5a	*Glyma.16G044100*	Promotes flowering	Kong *et al.* (2010); Fan *et al.* (2014)
	GmFT1a	*Glyma.18G298900*	Delays flowering	Guo *et al.* (2015)
	GmFT1b	*Glyma.18G299000*	Delays flowering	Guo *et al.* (2015)
	GmFT2b	*Glyma.16G151000*	Promotes flowering	Fan *et al.* (2014)

Loci	Gene	Accession number	Function	Reference
	GmFT2c	*Glyma.02G069500*	Pseudogene	Wu and Hanzawa (2018)
	GmFT3a	*Glyma.16G044200*	Promotes flowering	Fan *et al.* (2014)
	GmFT3b	*Glyma.19G108100*	Promotes flowering	Fan *et al.* (2014)
	GmFT5b	*Glyma.19G108200*	Promotes flowering	Fan *et al.* (2014)
	GmFT6	*Glyma.08G363200*	Delays flowering	Wang *et al.* (2015b)
E11	Unknown		Promotes flowering	Wang *et al.* (2019)
Tof11	*PRR3a*	*Glyma.U034500*	Delays flowering	Lu *et al.* (2017)
Tof12	*PRR3b*	*Glyma.12G073900*	Delays flowering	Lu *et al.* (2017)
	LHY1a	*Glyma.16G017400*	Promotes flowering	Lu *et al.* (2017)
	LHY1b	*Glyma.07G048500*	Promotes flowering	Lu *et al.* (2017)
	LHY2a	*Glyma.19G260900*	Promotes flowering	Lu *et al.* (2017)
	LHY2b	*Glyma.03G261800*	Promotes flowering	Lu *et al.* (2017)

*Table reproduced with permission from Lin *et al.* (2021b) but with some modifications to the *E5*, *E6*, *E7*, and *E8* loci.

Table 1.1 Major flowering genes and loci.

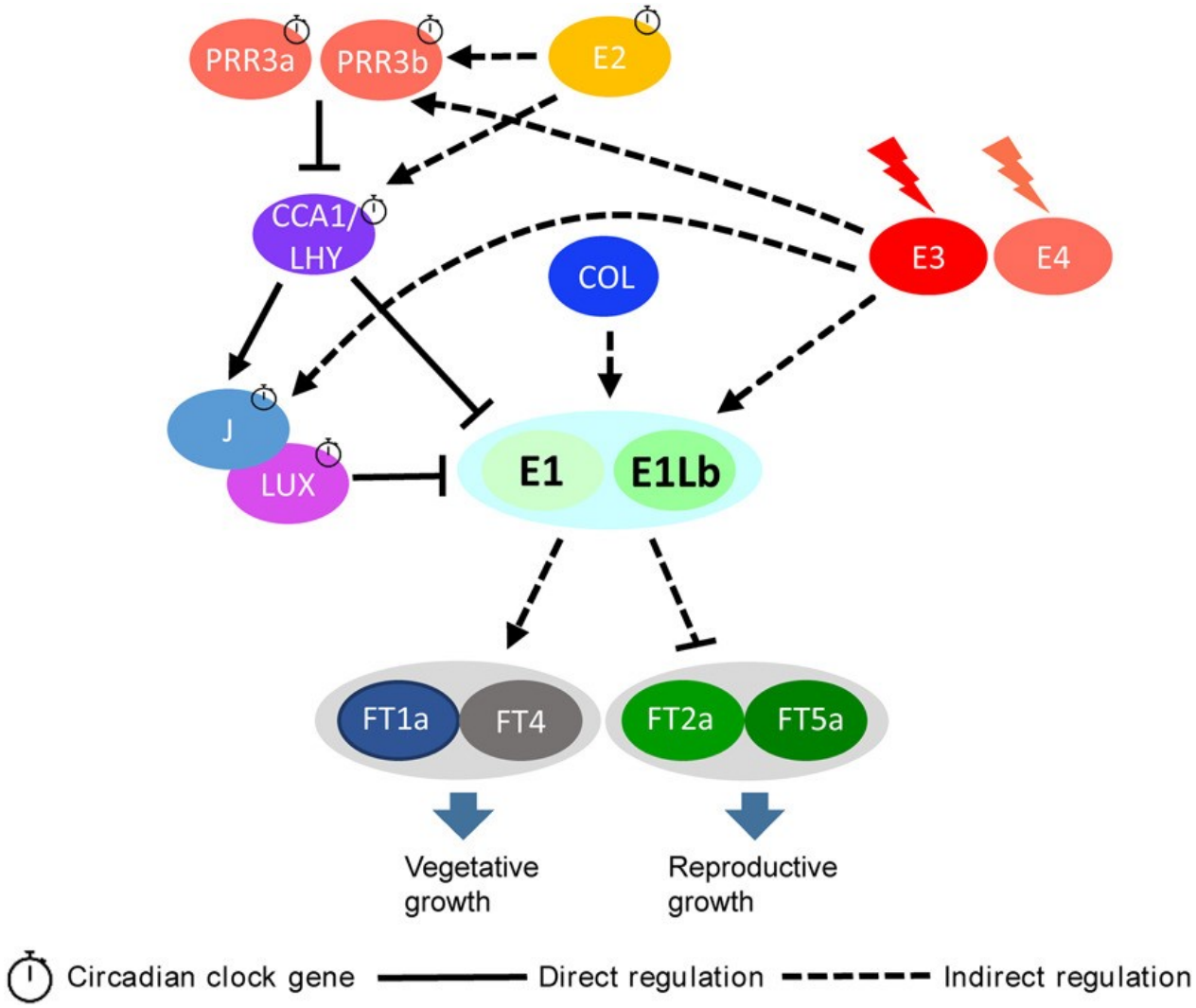

Figure 1.1 Photoperiodic flowering regulatory mechanisms in soybean. The dotted lines represent indirect regulation, whereas the solid lines represent direct regulation. The T-shape symbols indicate negative regulation, whereas the arrow symbol represents positive regulation. Figure reproduced with permission from Lin *et al.* (2021b).

3. Dissection of the *E8* Locus in Two Early Maturing Canadian Soybean Populations

Jérôme Gélinas Bélanger[1,2], Tanya Rose Copley[1], Valerio Hoyos-Villegas[2] & Louise O'Donoughue[1]

[1]CÉROM, Centre de recherche sur les grains Inc., St-Mathieu-de-Beloeil, Québec, Canada

[2]Department of Plant Science, McGill University, Montréal, Québec, Canada

* Correspondence:

Dr. Louise O'Donoughue

louise.odonoughue@cerom.qc.ca

Reproduced from Frontiers in Plant Science.

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., and O'Donoughue, L. (2024). Dissection of the E8 locus in two early maturing Canadian soybean populations. ***Front. Plant Sci.*** 15, 1329065. doi: 10.3389/fpls.2024.1329065

Minor modifications were made to conform to the McGill University thesis guidelines.

3.1 Abstract

Soybean [*Glycine max* (L.) Merr.] is a short-day crop for which breeders want to expand the cultivation range to more northern agro-environments by introgressing alleles involved in early reproductive traits. To do so, we investigated quantitative trait loci (QTL) and expression quantitative trait loci (eQTL) regions comprised within the *E8* locus, a large undeciphered region (~7.0 Mbp to 44.5 Mbp) associated with early maturity located on chromosome GM04. We used a combination of two mapping algorithms, (i) inclusive composite interval mapping (ICIM) and (ii) genome-wide composite interval mapping (GCIM), to identify major and minor regions in two soybean populations ($QS15524_{F2:F3}$ and $QS15544_{RIL}$) having fixed *E1*, *E2*, *E3*, and *E4* alleles. Using this approach, we identified three main QTL regions with high logarithm of the odds (LODs), phenotypic variation explained (PVE), and additive effects for maturity and pod-filling within the *E8* region: GM04:16,974,874-17,152,230 (*E8-r1*); GM04:35,168,111-37,664,017 (*E8-r2*); and GM04:41,808,599-42,376,237 (*E8-r3*). Using a five-step variant analysis pipeline, we identified *Protein far-red elongated hypocotyl 3* (*Glyma.04G124300*; *E8-r1*), *E1-like-a* (*Glyma.04G156400*; *E8-r2*), *Light-harvesting chlorophyll-protein complex I subunit A4* (*Glyma.04G167900*; *E8-r3*), and *Cycling dof factor 3* (*Glyma.04G168300*; *E8-r3*) as the most promising candidate genes for these regions. A combinatorial eQTL mapping approach identified significant regulatory interactions for 13 expression traits (e-traits), including *Glyma.04G050200* (*Early flowering 3/E6* locus), with the *E8-r3* region. Four other important QTL regions close to or encompassing major flowering genes were also detected on chromosomes GM07, GM08, and GM16. In GM07:5,256,305-5,404,971, a missense polymorphism was detected in the candidate gene *Glyma.07G058200* (*Protein suppressor of PHYA-105*). These findings demonstrate that the locus known as *E8* is regulated by at least three distinct genomic regions, all of which comprise major flowering genes.

3.2 Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most economically important crops worldwide and is a significant source of vegetable-based protein and oil (Pagano and Miransari, 2016). Domesticated 3,000–9,000 years ago in East Asia from wild soybean (*Glycine soja* Siebold & Zucc.) (Hyten *et al.*, 2006; Lee *et al.*, 2011), the crop has spread throughout the world and is now cultivated in Brazil (36.4%), the United States (34.3%), Argentina (12.1%), China (5.1%),

India (3%), Canada (2%), Paraguay (1%), and several other countries (6%) (The American Soybean Association, 2023). While domesticated in a region located between 30°N–45°N and encompassing the eastern Huanghe (Yellow River) basin in North China, South Korea, and Japan (Hyten *et al.*, 2006; Lee *et al.*, 2011), the plant's ability to adapt to very northern environments is still limited by its short-day photoperiod requirements. Indeed, its recent expansion into northern agricultural regions has only been possible due to major breeding efforts focused on selecting non-photosensitive lines (Zhang *et al.*, 2017c). In Canada, the cultivar maturity requirements range from MG000 to MGIII, depending on the region, with an approximate 10-day difference between each group (Bagg *et al.*, 2009). Recently, a putative new maturity group (MG0000) hailing from northeast China and far east Russia has been proposed (Jia *et al.*, 2014; Jiang *et al.*, 2019). This new maturity group demonstrates that expanding soybean's growing zone beyond its actual northern limits (~54°N) is still possible. However, to unlock soybean's northern potential, breeders still need to identify novel genes involved in the regulation of early flowering and maturity.

Over the years, several major genes and quantitative trait loci (QTL) involved in reproductive traits, such as *E1-E11, J*, *Time of flowering* (*Tof*) *5/11/12/16/18* and *Flowering locus T* (*GmFT*) homologs, have been identified and characterized using forward and reverse genetic approaches (Lin *et al.*, 2021b; Gupta *et al.*, 2022). Maturity loci *E1* (*Glyma.06G207800*), *E2* (*Glyma.10G221500*), *E3* (*Glyma.19G224200*), and *E4* (*Glyma.20G090000*) are frequently reported as the most critical players in terms of influence on the final maturity phenotype, explaining more than 60% of the variation in the observed flowering with the proper haplotype combinations (Tsubokura *et al.*, 2014). In addition, determinate habit genes *Dt1* and *Dt2* have been demonstrated to play a complementary role by regulating the growth habit, flowering time, and maturity in soybean (Liu *et al.*, 2010; Ping *et al.*, 2014; Zhu *et al.*, 2023). Loss-of-function variants in *E1–E4* contribute to photoperiod insensitivity by indirectly favoring the expression of *FT* orthologs such as *GmFT2a* (*Glyma.16G150700*) and *GmFT5a* (*Glyma.16G044100*) (Kong *et al.*, 2010; Thakare *et al.*, 2011; Watanabe *et al.*, 2011). Studies have shown a high correlation between the latitudinal adaptability/photoperiod insensitivity and the number of recessive alleles for these four *E* loci (Jiang *et al.*, 2014). Several other loci, such as *E9* (*Glyma.16G150700*), *E10* (*Glyma.08G363100*), and *Tof 5/11/12/16/18*, are slowly being implemented in early maturity breeding programs although their effects on flowering

and maturity are generally less pronounced than for *E1–E4* (Kong *et al.*, 2014; Samanfar *et al.*, 2017; Lu *et al.*, 2020; Dong *et al.*, 2021, 2022; Kou *et al.*, 2022).

The *E8* locus is an interesting locus for breeders as its recessive allele (*e8e8*) imparts a flowering date that is ~5–8 days earlier than its dominant form (Cober *et al.*, 2010). This locus has been mapped between markers Sat_404 and Satt136 on chromosome GM04 (Cober *et al.*, 2010). Two recent research articles have mapped QTL regions located on GM04 which could be *E8* (Kong *et al.*, 2018; Wang *et al.*, 2018); however, the regions identified in these papers are broad [GM04:13,212,370–43,843,500 bp for Kong *et al.* (2018) and GM04:7,166,748–44,508,948 bp for Wang *et al.* (2018)] and encompass multiple critical flowering genes such as *E1-like-a* (*Glyma.04G156400*; *E1la*) and *E1-like-b* (*Glyma.04G143300*; *E1lb*), two *E1* homologs. Consequently, these large physical locations suggest that multiple regulatory regions might be controlling flowering on chromosome GM04. Using bioinformatic analyses, seven candidate genes have been proposed for *E8* (Sadowski, 2020), all of which are located between GM04:9,337,214 and GM04:22,755,516.

Through our experiments, we observed significant differences in maturity time between Canadian lines from two early maturing populations (MG00-MG000) that were selected and fixed for identical *E1-E4* alleles, thus suggesting potential novel sources of regulation for these traits. These populations were developed to reduce the background noise generated by *E1-E4* due to their important role in maturity in terms of phenotypic variation. The narrow genetic diversity of Canadian soybean lines, especially within early maturing accessions, suggests that only a handful of regions and causal variants might be contributing to these observed phenotypes (Grainger and Rajcan, 2014). With this study, we aimed to (i) develop a combinatorial QTL analysis approach to map the regions regulating several reproductive traits under field (fluctuating photoperiod with long days during the flowering period) and greenhouse conditions (constant short days) in two plant populations; (ii) perform expression quantitative trait loci (eQTL) analyses to identify interactions with important flowering genes; and (iii) propose candidate genes involved in early maturity in relation to their gene expression level, gene ontology (GO) annotations and/or genetic polymorphism profile.

3.3 Materials and Methods

3.3.1 Plant Materials

The full mapping population of 176 $F_{2:3}$ individuals of the QS15524 population (herein named QS15524$_{F2:F3}$) was derived from a single biparental cross between "OAC Vision" ♀ (PI 567787; MG000, earlier maturing accession) × "Maple Arrow" ♂ (PI 548593; MG00, later maturing accession). The full mapping population of 162 $F_{5:8}$ individuals of the QS15544 population (recombinant inbred lines; herein named QS15544$_{RIL}$) was derived from a single biparental cross between "9004" ♀ (PI 592534, US PVP No. 9600050; MG000, earlier maturing accession) × "AAC Mandor" ♂ (MG00, later maturing accession). The "AAC Mandor" parental line is a food-grade soybean cultivar developed by Dr. Elroy Cober at the Ottawa Research and Development Centre of Agriculture and Agri-food Canada (ONT, Canada). Both populations used in this study were developed at the Centre de recherche sur les grains (CÉROM) inc. in Saint-Mathieu-de-Beloeil (QC, Canada). To generate the QS15544$_{RIL}$ population, the offspring of the "9004" × "AAC Mandor" cross were mass multiplied until reaching the F_5 generation at which point 200 plants were randomly selected and grown over one season in the greenhouse and three seasons in the field for phenotyping. To identify novel QTL, we genotyped each parent to confirm that those were fixed for *E1* (*Glyma.06G207800*) (Xia *et al.*, 2012), *E2* (*Glyma.10G221500*) (Watanabe *et al.*, 2011), *E3* (*Glyma.19G224200*) (Watanabe *et al.*, 2009; Harada *et al.*, 2011; Xu *et al.*, 2013) and *E4* (*Glyma.20G090000* (Liu *et al.*, 2008; Tsubokura *et al.*, 2013; Tardivel *et al.*, 2019) genes. As such, the genotypes for the "OAC Vision" and "Maple Arrow" parental lines were identified as *e1-nl*/*e2-ns*/*E3Ha*/*e4-SORE-1* for the QS15524$_{F2:F3}$ population. For the QS15544$_{RIL}$ population, the genotypes were *e1-as*/*e2-ns*/*e3-tr*/*e4p.T832QfsX21* for the "9004" and "AAC Mandor" parental lines.

3.3.2 Growing Conditions, Tissue Sampling and Phenotyping

For the eQTL analyses, the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations were grown following a Modified Augmented Design (Lin and Poushinsky, 1983, 1985) that was slightly adjusted for greenhouse conditions such that each table contained one parent and 19 individuals. Under these conditions, each population was phenotyped for the number of days to maturity during the winter 2017–2018 (F_2 generation of the QS15524$_{F2:F3}$ population) and winter 2019–2020 (F_5 generation of the QS15544$_{RIL}$ population), respectively. For the greenhouse experiments, the

plants were sown in one-gallon pots containing a ProMix-garden soil (1:1 v:v) (Premier Tech Horticulture, Rivière-du-Loup, QC, Canada) potting mix, with one seed per pot for the $QS15524_{F2:F3}$ population or three seeds for the $QS15544_{RIL}$ population. Seeds were sown at a depth of 4 cm and inoculated with 1×10^8 colony-forming units of liquid Cell-tech® (Novozymes BioAg, Saskatoon, SK, Canada) *Bradyrhizobium japonicum* at sowing and placed in a greenhouse with the following growing conditions: 12:12 light:dark (L:D), 27°C/24°C (L:D), and 80% relative humidity (Fehr, 1980). Plants were watered daily with a drip irrigation system with increasing volume to meet the plant needs and fertilized weekly alternating with a 15-30-15 or 20-20-20 (nitrogen-phosphorus-potassium) nutrient solution. Five pots of each parent were sown at the same time as the mapping population for a total of 190 study plants for each population. Pots were placed randomly across ten greenhouse tables with 20 pots per table. Due to extremely late maturity, or plant damage, a total of 184 and 182 individuals were retained for the eQTL and QTL analyses for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, respectively. Leaf tissue was harvested from plants grown in the greenhouse 25 days after sowing (V4 stage) for both populations (McWilliams *et al.*, 2004). Samples were taken four hours after sunrise for RNA extraction, while samples for DNA extraction were taken later in the day. All samples were immediately frozen in liquid nitrogen after harvesting and stored at −80°C until further use. These time points were taken from previously published data indicating highest expression of flowering genes four hours after sunrise (Kong *et al.*, 2010; Sun *et al.*, 2011), while the V4 stage was determined as the optimal stage according to qRT-PCR analyses of the expression of the flowering genes *Glyma.16G150700* (*GmFT2a*) and *Glyma.16G044100* (*GmFT5a*) in the parents.

For the field phenotypes, the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations were grown in Saint-Mathieu-de-Beloeil (QC, Canada) using a Modified Augmented Design (Lin and Poushinsky, 1983, 1985). The F_3 generation of the $QS15524_{F2:F3}$ population was grown in single-row plots over two seasons (summers of 2018 and 2021) and the F_6:F_8 generations of the $QS15544_{RIL}$ population were grown over the summers of 2020 (one-row plots), 2021 (two-row plots), and 2022 (two-row plots), respectively. The phenotyping of the field traits was performed on 10 plants of the F_3 generations for the $QS15524_{F2:F3}$. The field phenotypes were recorded as follows: (i) number of days to flowering, as the day of planting to the day at which 75% of the genotype was flowering; (ii) number of days to maturity, as the day of planting to the day at which 95% of the pods within the genotype were at physiological maturity; and (iii) number of days to

filling, as the number of days from flowering to maturity. To map the QTL regions for the F_3 generation of the QS15524$_{F2:F3}$ population, we averaged the observed phenotypes and used them as if those were phenotypes from the F_2 generation in the mapping pipeline. Phenotypic data distribution, Q-Q plots and Pearson correlation coefficients (PCC) were generated using R version 4.0.4 (R Core Team, 2010). Statistical analyses for the Modified Augmented Design were performed in Agrobase Generation II® (Agronomix Software Inc., 2009). Descriptive statistics (i.e., variance, standard error, kurtosis, and skewness) for the four reproductive phenotypes were calculated using QTL IciMapping 4.2 (Meng *et al.*, 2015). The broad-sense heritability values were estimated using a linear mixed model with the est_herit function implemented in R/qtl2 (Broman *et al.*, 2019). The kinship matrices used to estimate the heritability values were generated with the calc_kinship function implemented in the same package. Statistical differences between the years of data collection were calculated using a paired Student's t-test (t-test function in R) or a one-way analysis of variance (ANOVA) (aov function in R) using a threshold *p*-value of 0.01.

3.3.3 Nucleic Acid Extraction and Sequencing

Total DNA was extracted from leaf tissue using the Omega Bio-Tek Mag-bind Plant Kit (Omega Bio-tek, Norcross, GA, USA) with further purification using the Mag-Bind Total Pure NGS (Omega Bio-tek, Norcross, GA, USA). Sampling of the QS15524$_{F2:F3}$ parental lines for the whole genome sequencing (WGS) was performed by pooling the samples from the five pots used to grow each of the parental line, extracting total DNA, and having the libraries prepared at the Génome Québec Innovation Centre (Montréal, QC, Canada) using the NxSeq® AmpFREE Library Preparation kit (Lucigen, Middleton, WI, USA). To do so, the two parental libraries were barcoded, combined and sequenced to a depth of 15X on the Illumina HiSeq X platform with 150 base pair paired-end reads. The WGS data for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ parental lines were also retrieved from the GmHapMap as available and detailed in Torkamaneh *et al.* (2020). Phasing of the single-nucleotide polymorphisms (SNPs) data for the QS15524$_{F2:F3}$ population was performed using the resequenced WGS data, whereas the identification of the candidate SNP was performed using the GmHapMap WGS datasets.

To generate the genotyping-by-sequencing (GBS) datasets of the QS15524$_{F2:F3}$ (F_2 generation) and QS15544$_{RIL}$ (F_5 generation) mapping populations, the libraries were prepared at the Institute of Integrative Biology and Systems (Laval University, Quebec City,

QC, Canada) using the *PstI/MspI* enzymes as detailed in Abed *et al.* (2019). Samples were randomly divided into two sets of 91 individuals, which were barcoded and pooled to form two libraries per population. Sequencing of the QS15524$_{F2:F3}$ GBS libraries was done by combining a total of 91 barcoded samples per library. Sequencing of each library was done on four Ion PI V3 Chips per library with sequencing performed on the Ion Proton Sequencer and HiQ chemistry at the Institute of Integrative Biology and Systems, for a total of eight sequenced chips. For the QS15544$_{RIL}$ population, samples were randomly divided into two sets of 91 samples and sequenced using the same technologies, with two chips per library.

Total RNA was extracted from samples using a standard Trizol™ (Invitrogen, Waltham, MA, USA) RNA extraction procedure as detailed in the company's protocol, with two additional ethanol rinses to improve purity. Isolation of messenger RNA (mRNA) was performed using the NEBNext mRNA stranded library preparation kit (New England Biolabs, Ipswich, MA, USA) at the Génome Québec Innovation Centre. A total of 96 samples were barcoded and pooled per final library with one population per library. Each of the libraries was sequenced on two Illumina NovaSeq6000 lanes using S2 or S4 flow cells with 100 base pair paired-end sequencing at the Génome Québec Innovation Centre, for a total of four sequencing lanes. Genome coverage was evaluated to be ≈43.9 M paired-end reads per sample for the QS15524$_{F2:F3}$ and ≈50M reads per sample for the QS15544$_{RIL}$ population.

3.3.4 Bioinformatics

All sequencing alignment was done using version 2 of the *Glycine max* reference genome (Gmax_275_v2.0). Whole genome sequencing data were processed using the fast-WGS pipeline (Torkamaneh *et al.*, 2018) for the QS15524$_{F2:F3}$ parental lines. Briefly, raw data were aligned to the genome using Burrows-Wheeler Alignment (Li and Durbin, 2009) with the command: bwa mem refGenome Input. Variants were called using Platypus version 0.8.1 (Rimmer *et al.*, 2014) with the following commands: –minReads=2, –minMapQual=20, and –minBaseQual=20. GBS data were processed using the fast-GBS pipeline (Torkamaneh *et al.*, 2017). Briefly, samples were demultiplexed using Sabre version 1.00 (Joshi, 2011), and their adapters removed using Cutadapt (Martin, 2011). The samples were subsequently aligned to the reference genome using Burrows-Wheeler Alignment with the command: bwa mem refGenome Input. Quality checks on the raw data were performed using FastQC software version 0.11.9 (Andrews, 2010). Variants were then

called using Platypus version 0.8.1 with the following commands: –minReads=2, –minMapQual=20, and -minBaseQual=20. Genotypes were filtered using vcftools version 0.1.16 (Danecek *et al.*, 2011) to (i) maintain only biallelic sites, (ii) remove InDels, (iii) keep polymorphisms located only on chromosomes and not scaffolds, and (iv) filter allele frequency and count with the –maxmissing 0.2, –maf 0.3, and –mac 4 commands. For the $QS15544_{RIL}$ population only, each SNP and offspring was then filtered based on their heterozygosity using an interquartile range approach $\{[Q_1-k(Q_3-Q_1), Q_3+k(Q_3-Q_1)]$, $k = 3$, as per Tukey (1977). As such, loci with >14.85% heterozygous calls and offspring with >18.57% heterozygous calls were considered outliers and removed. Missing genotypes for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations were then self-imputed using Beagle version 4.1.0 with 12 iterations (Browning and Browning, 2016). Genotypes were phased with Convert2map https://bitbucket.org/jerlar73/convert-genotypes-to-mapping-files/src/master/ (accessed 12 December 2023) using the parental data from the GmHapMap for the $QS15544_{RIL}$ population and the fast-WGS resequenced data for the $QS15524_{F2:F3}$ parental lines. Subsequently, correction of the genotype calls for the $QS15524_{F2:F3}$ population was performed using Genotype Corrector (Miao *et al.*, 2018) with the software default options (sliding window size of 11 and error rates for homo1 and homo2 of 0.03 and 0.01, respectively) and all the implemented quality checks. For the $QS15544_{RIL}$ population, the removal of the double crossovers was performed using Convert2map. Finally, all genotypes with >10% heterozygous calls were removed from the $QS15544_{RIL}$ dataset before binning with QTL IciMapping version 4.2. For the $QS15524_{F2:F3}$ population, binning was performed with the binning option implemented in Genotype Corrector.

RNA dataset processing was performed using an in-house script comprising multiple publicly available software tools. Briefly, adapters were removed using Trimmomatic version 0.33 33 (Bolger *et al.*, 2014) with the following options: ILLUMINACLIP:$prog/Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:15\, LEADING:3\ and TRAILING:3\, SLIDINGWINDOW:3:20\ and MINLEN:32\. Filtered reads were then aligned to the soybean reference genome using TopHat2 version 2.1.1 (Kim *et al.*, 2013).

3.3.5 Map Construction and QTL Analysis

The genetic linkage maps of the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations were generated using QTL IciMapping version 4.2 with the Kosambi mapping function to convert the

recombination frequency into centimorgans (cM). The QS15524$_{F2:F3}$ map was generated with "Maple Arrow" as parent A (positive additive effect) and "OAC Vision" as parent B (negative additive effect), whereas the QS15544$_{RIL}$ map was generated with "AAC Mandor" as parent A (positive additive effect) and "9004" as parent B (negative additive effect). In this specific case, a positive additivity relates to the increase in the number of days to flowering, pod-filling, and maturity. The linkage groups were split when gaps exceeded 30 cM and the markers were anchored to their physical positions. The linkage maps with the displayed QTL regions were drawn using the Linkage Map View version 2.1.2 package in R (Ouellette *et al.*, 2018). The condensed versions of the full linkage maps were plotted by https://www.bioinformatics.com.cn/en (accessed 12 December 2022), a free online platform for data analysis and visualization.

For the QTL analyses, we opted for a combinatorial approach using two standard mapping algorithms: ICIM approach implemented in QTL IciMapping version 4.2 (Meng *et al.*, 2015), and Genome-wide compositive interval mapping (GCIM) method in the QTL.gCIMapping.GUI.v2.0.GUI package (Zhang *et al.*, 2020c). Briefly, ICIM was performed using the following mapping parameters: (i) deletion of the missing phenotypes; (ii) a scanning interval step of 1 cM and a PIN of 0.001; and (iii) a logarithm of the odds (LODs) threshold determined with 1,000 permutations and α of 0.05. GCIM was performed using the fixed model and a walking speed of 1 cM for both populations. In addition, mapping in the QS15524$_{F2:F3}$ population was performed by choosing the restricted maximum likelihood (REML) function implemented in the same software. To remove the minor QTL regions, we subsequently increased the identified ICIM LOD thresholds from 3.99–4.24 (QS15524$_{F2:F3}$) and 3.43–3.57 (QS15544$_{RIL}$) to 5.25, and the GCIM LOD threshold from 2.5 (i.e., the default parameter) to 7.1 (Zhang *et al.*, 2020c). Subsequently, we decided to only retain GCIM with a phenotypic variation explained (PVE) ≥3.5% and ICIM regions with a PVE ≥5.5%. Finally, we only retained regions that were either: (i) identified within both populations; (ii) identified by ICIM and GCIM within the same population; or (iii) identified with only one algorithm within only one population but with LOD ≥ 12 and PVE ≥ 20%. For the GCIM regions for which both flanking markers were the same, we considered a ± 100,000 bp region upstream and downstream of the flanking markers when investigating candidate variants. The recombination fraction figures were calculated using the PlotRF function implemented in R/QTL version 1.50 (Broman *et al.*, 2003) and visualized using ASMap version 1.0-4 (Taylor and Butler, 2017) in R. The QTL regions identified with this

combinatorial pipeline were named using the following nomenclature: (i) Method (i.e., ICIM or GCIM); (ii) Population (i.e., 24/QS15524$_{F2:F3}$ or 44/QS15544$_{RIL}$); and (iii) QTL trait and associated number (e.g., mat1 for maturity region 1, fill2 for filling region 2, and flow 3 for flowering region 3). To reduce the number of studied regions, we merged the loci that were found in both populations using the following nomenclature: (i) Merg; (ii) chromosome number; and (iii) field (f) or greenhouse conditions (gh). To increase the precision of our QTL mapping procedure, we generated the results both for (i) each year of data and (ii) phenotypic averages for all the studied years. Based on our observations, the results between both types of analysis (i.e., each year and phenotypic averages) were largely comparable for most regions and a preference was given to the phenotypic averages for the main analysis.

3.3.6 Expression QTL Mapping

The mapping of eQTL regions was performed as in Gélinas Bélanger *et al.* (unpublished). Briefly, eQTL analysis was performed on the DESeq2 normalized transcript abundances of 38,692 genes of the 176 F_2 lines of the QS15524$_{F2:F3}$ population and 40,218 genes of the 162 F_5 lines of the QS15544$_{RIL}$ population. Mapping of the eQTL traits was performed using a combinatorial approach that includes the use of three different algorithms: (i) ICIM; (ii) GCIM; and (iii) Interval mapping (IM) from QTL IciMapping version 4.2. The LOD thresholds for ICIM and IM were calculated in QTL IciMapping with 1,000 permutations of 100 sampled expression traits (e-traits) with α of 0.05 and a walking step of 1 cM for genome-wide scanning. Subsequently, global permutation thresholds were calculated as the 95th percentile of the representative null distribution and equaled to (i) 4.01 for ICIM in QS15544$_{RIL}$; (ii) 3.99 for IM in QS15544$_{RIL}$; (iii) 4.13 for ICIM in QS15524$_{F2:F3}$; and (iv) 4.12 for IM in QS15524$_{F2:F3}$. For GCIM, the REML-fixed and fixed model components were respectively chosen for the QS15524$_{F2:F3}$ population and QS15544$_{RIL}$ populations, both with a walking speed of 1 cM. In the QTL.gCIMapping.GUI v2.0 package, the likelihood function is only available to F_2 populations and was chosen based on prior testing. The GCIM LOD threshold was increased from 2.5 to 7.5 for QS15524$_{F2:F3}$ and 4.0 for QS15544$_{RIL}$ to improve the reliability of the results.

Expression QTL generated by the three algorithms were retained only if they fell within ± 1 Mbp in at least two of the three methods. To do so, the interactions were divided between *trans*-acting and *cis*-acting, and the size of each of the identified eQTL regions (i.e., all of the loci

identified with the three aforementioned algorithms) was manually adjusted by adding 500,000 bp both upstream and downstream. The overlapping sets of regions were then identified using the genomic peak Venn function implemented in https://www.bioinformatics.com.cn/en (accessed 12 December 2022), a free online platform for data analysis and visualization. The overlaps were identified using a pairwise comparison (e.g., ICIM vs. IM) using the ICIM interactions as the reference regions in the ICIM versus IM and ICIM versus GCIM analyses. In addition, the IM regions were used as references in the IM vs GCIM analysis. *Trans* interactions overlapping *cis* regions were *de facto* considered as *cis* and excluded from *trans*-interactions hotspot mapping.

3.3.7 Identification of Candidate SNPs and Genes

Candidate SNPs and genes were identified using a five-step custom pipeline. First, the prediction of the deleterious effects of the SNPs was performed using Ensembl Variant Effect Predictor (VEP) with Glycine_max_v2.1 (McLaren *et al.*, 2016). Second, putative effects of identified non-synonymous missense polymorphisms were then predicted using Sorting Intolerant From Tolerant 4G (SIFT4g) using "William 82" as the wild-type allele (Ng and Henikoff, 2003; Kumar *et al.*, 2009). To do so, we generated a soybean database using the annotations of *G. max* Wm82.a2.v1 from Ensembl Plants (Yates *et al.*, 2022) and by following the SIFT4G_Create_Genomic_DB guidelines available at https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB (accessed 12 December 2022). The SNPs with SIFT scores <0.05 were classified as putatively deleterious and the ones ≥0.05 were considered as tolerated. Third, we matched the parental genotypes from the GmHapMap (Torkamaneh *et al.*, 2021) dataset with the parental allele causing the additive effect. Fourth, we retained only polymorphisms that were predicted as having moderate or high consequences on the protein structure. Variants located in the 3′ and 5′ (UTR) regions were also retained if those were identified within the sequence of a gene with a validated reproductive function in soybean. Fifth, we generated one custom GO database by retrieving 162 terms flagged as linked to (i) flowering and maturity and (ii) photosynthesis and photoperiodic response from Soybase (Grant *et al.*, 2009) as detailed in Gélinas Bélanger *et al.* (unpublished). Also, we retrieved 836 soybean genes identified as putatively involved in flowering based on comparative analysis using *Arabidopsis* orthologs (Zhang *et al.*, 2017c). Genes identified as having ≥3 GO annotations, flagged as being an *Arabidopsis* flowering

ortholog, validated for a reproductive function, and/or harboring one or multiple deleterious polymorphisms were prioritized in the downstream analysis.

3.4 Results

3.4.1 Generation of the Populations and Phenotypic Analysis

To perform our experiment, we generated the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations and phenotyped both in the greenhouse (one trait; maturity) during the winter and in the field (three traits; flowering, pod-filling, and maturity) during the summer. Both populations exhibited an agronomically important difference in terms of the number of days to flowering in the field, pod-filling in the field, maturity in the field, and maturity in the greenhouse for each year of data (Appendix 1.1, 1.2) and also their phenotypic average for each trait (Fig. 2.1; Appendix 1.1, 1.2). When comparing both populations, the QS15524$_{F2:F3}$ population always displayed an earlier phenotype for all reproductive traits. Transgressive segregation was mainly observed in the QS15544$_{RIL}$ population based on the respective distribution pattern of the offspring for each trait. The distribution of all of the phenotypes followed a normal distribution, except for the number of days to flowering of the QS15524$_{F2:F3}$ population (Fig. 2.2; Appendix 1.1, 1.2). The broad-sense heritability values for each of the trait and years of data collection were high (i.e., $H^2 \geq 0.5$), except for the number of days to flowering in both populations, thus indicating that genotypes contribute to most of the variation observed in the studied traits (Appendix 1.2). Likewise, the pairwise PCC for each of the trait and years of data collection were also high (PCC ≥ 0.5), except for the flowering trait (Appendix 1.3). A significant year effect was detected for all phenotypes based on the t-test and ANOVA analyses (Appendix 1.4); however, the high-heritability values and PCC between the years suggest that this observation was most likely due to a magnitude effect on the trait. Consequently, the traits were further analyzed using the phenotypic averages for all the studied years.

3.4.2 Construction of the Linkage Maps

Linkage maps based on the segregation of GBS-derived SNP markers for 176 F_2 lines of the QS15524$_{F2:F3}$ population (Fig. 2.3A) and 162 F_5 lines of the QS15544$_{RIL}$ population were generated (Fig. 2.4A). A total of 541,106,451 and 286,844,986 unique single-end reads were generated in the sequencing step for QS15524$_{F2:F3}$ and QS15544$_{RIL}$, respectively. For the final

linkage maps, 1,613 ($QS15524_{F2:F3}$; Appendix 1.5, 1.7) and 2,746 ($QS15544_{RIL}$; Appendix 1.6, 1.7) polymorphic markers were retained after applying our SNP filtering pipeline. Splitting of the markers distanced by a gap >30 cM resulted in a map with 26 linkage groups with a total length of 2,971 cM, an average genetic distance between the markers of 1.84 cM, and an average length per linkage group of 114.27 cM for $QS15524_{F2:F3}$ (Table 2.1). The same procedure generated 34 linkage groups measuring an average length of 148.77 cM with an average distance between markers of 1.84 cM, and a total length of 5,058 cM in $QS15544_{RIL}$ (Table 2.2). The high quality of both maps was confirmed by plotting the genetic distance versus the physical position (Fig. 2.3B, 2.4B) and the pairwise recombination fraction and LOD score (Fig. 2.3C, 2.4C).

3.4.3 Quantitative Trait Loci Mapping

Mapping of the QTL regions was performed using a combinatorial approach with two algorithms (ICIM from the QTL IciMapping software and GCIM from the QTL.gCIMapping R package) for all four traits. The QTL regions were identified both for each year of data (Appendix 1.8) and the phenotypic averages for all the studied years (presented below). Overall, we identified a total of three regions (*MergGM04f*, *MergGM04gh*, and *MergGM08f*) that were present in both populations (Table 2.3) and also four unique regions that were identified only in $QS15544_{RIL}$ (Table 2.4) using the phenotypic averages. In addition to these major regions, several minor QTL loci were also mapped in both populations (Appendix 1.9).

For the $QS15524_{F2:F3}$ population, ICIM and GCIM identified a total of 10 QTL on chromosomes GM04 and GM08 (Table 2.3). Overall, the most significant QTL in terms of LOD, PVE, and additive effects were identified on GM04 (Fig. 2.5A; Table 2.3). Four QTL were detected in a ≈450 Kbp region located between the GM04:36,499,381 and GM04:36,941,521 flanking markers with ICIM (ICIM_24_fill1 and ICIM_24_mat1) and GCIM (GCIM_24_fill1 and GCIM_24_mat2). These QTL were displaying high LOD (33.80–51.60), PVE (28.00%–48.20%), and additive effects (3.19–3.85 days to maturity; 2.85–3.58 days to pod-filling). For the greenhouse maturity trait, we also identified two QTL for $QS15524_{F2:F3}$ (ICIM_24_matgh1 and GCIM_24_matgh1) that were located between the GM04:41,808,599 and GM04:42,156,365 flanking markers (Fig. 2.5A; Table 2.3). These regions were in close physical proximity (± 5 Mbp) to the region encompassing the four field QTL, but those were clearly distinct. For the maturity in the greenhouse QTL, the LOD (11.90 and 12.40), PVE (18.70% and 29.40%), and additive effects

(3.2–3.73 difference in the number of days to maturity) were also high, albeit slightly inferior to those observed for the field phenotypes. Four QTL from the field data (ICIM_24_fill2, GCIM_24_fill5, ICIM_24_mat4, and GCIM_24_mat6) were also detected on chromosome GM08 between the GM08:47,258,336 and GM08:47,289,756 flanking markers (Fig. 2.5C; Table 2.3). For the four regions located on GM08, the LOD scores were between 6.30 and 13.80 and the PVE between 4.40 and 8.70%.

Using the same approach as for the $QS15524_{F2:F3}$ population, we identified a total of 12 QTL (four with GCIM and eight with ICIM) for the $QS15544_{RIL}$ population (Tables 2.3, 2.4). Three QTL for the number of days to maturity in the field (ICIM_44_mat1, GCIM_44_mat1, and ICIM_44_mat2) were detected on chromosome GM04 in a region comprised between the GM04:35,168,111 and GM04:37,664,017 flanking markers (Fig. 2.5B; Table 2.3). The LOD scores for these three traits ranged from 7.10 to 19.60, while the PVE varied between 8.70 and 22.10%. One QTL for the greenhouse maturity trait, ICIM_44_matgh1, with a high additive effect (2.07 days) and PVE (15.70%), was identified between the GM04:42,368,274 and GM04:42,376,237 flanking markers (Fig. 2.5B; Table 2.3). Another significant QTL for pod-filling in the field (ICIM_44_fill2), located between the GM04:16,974,874 and GM04:17,152,230 flanking markers, was also identified in the $QS15544_{RIL}$ population, but only with ICIM and not GCIM (Fig. 2.6A; Table 2.4). To confirm that this hit was not an artifact of the algorithm, we performed QTL analyses for each season's data for the pod-filling and maturity traits and also computed their pairwise average for each season's pair (e.g., 2020 and 2021). A total of nine QTL (ICIM, seven hits; GCIM, two hits) with LOD scores ranging between 6.43 and 20.54 were identified within a ≈2.5 Mbp region starting at GM04:15,748,916 and ending at GM04:18,312,993, thus reinforcing our confidence that this observation was not an artifact (Appendix 1.9).

The field data also yielded QTL in other regions of the genome. One QTL (ICIM_44_mat6) with a lower LOD score (5.40) and PVE (4.80%) was detected on chromosome GM08 (Fig. 2.5D; Table 2.3) in a physically close region (≈500 Kbp) to the one identified in $QS15524_{F2:F3}$. Two QTL, ICIM_44_mat5 and GCIM_44_mat5, with a high-statistical significance (LOD scores of 8.52 and 11.35, respectively) were identified on chromosome GM07 (Fig. 2.6B; Table 2.4). Two QTL related to the number of days to maturity, ICIM_44_mat3 and GCIM_44_mat2, were identified on chromosome GM16 between the GM16:5,680,173 and GM16:5,730,237 flanking markers (Fig. 2.6C; Table 2.4). In addition, two other QTL were identified on the same linkage

group in a region located between the GM16:22,756,017 and GM16:23,154,638 flanking markers (Fig. 2.6C; Table 2.4). All of the QTL identified on GM16 had important LOD, PVE, and additive effects.

3.4.4 Identification of Candidate SNPs and Genes

As described in the Material and Methods section, we developed a five-step analytical pipeline to discover the best candidate SNPs and genes. This pipeline was subsequently applied to the seven QTL regions identified with ICIM and GCIM (three merged regions and four unique for $QS15544_{RIL}$). For the merged regions, we identified a total of 14 missense polymorphisms (9 SIFT-Tolerated and 5 SIFT-Deleterious), five 3′UTR, and one 5′UTR variant (Table 2.5). For the regions unique to $QS15544_{RIL}$, 10 missense polymorphisms (7 SIFT-Tolerated and 3 SIFT-Deleterious) were identified along with two 3′UTR variants, one splice donor and one stop-gain variant (Table 2.6). Amongst these polymorphisms, several were located in genes known to be involved in maturity and reproduction. Polymorphisms located in the 3′UTR regions were identified in *E1la* and *Glyma.04G167900* (*Light-harvesting chlorophyll-protein complex I subunit A4; GmLHCA4*) for the merged regions, and in *Glyma.16G044100* (*GmFT5a*) and *Glyma.07G049400* (*Pseudo-response regulator 5d; GmPRR5d*) for the unique regions identified in $QS15544_{RIL}$. The 5′UTR variant was identified in *Glyma.04G166300* (*Pseudo-response regulator 1a; GmPPR1a*). For the *MergGM04gh* region, a SIFT-Tolerated missense polymorphism was detected in *Glyma.04G168300* (*Cycling dof factor 3*; *GmCDF3*), a transcription factor with a known impact on flowering in *Arabidopsis*. For the unique regions identified in $QS15544_{RIL}$, multiple missense variants were identified in important flowering genes. In the GM04:16,974,874-17,152,230 region, we identified a SIFT-Tolerated missense polymorphism in *Glyma.04G124300* (*Protein far-red elongated hypocotyl 3*; *GmFHY3*) and a SIFT-Deleterious missense polymorphism in *Glyma.04G124600* (*Far1-related sequence 5*; *GmFRS5*). A stop-gain polymorphism was also identified in *Glyma.04G124800* (*Zinc inducted facilitator-like 1*; *GmZIFL1*) in the same region. A SIFT-Tolerated polymorphism was also identified in *Glyma.07G058200* (*Protein suppressor of PHYA-105; GmSPA1*) for the GM07:5,256,305–5,404,971 region. A splice donor variant predicted to have a high impact on the protein structure was identified in *Glyma.16G110700* (*Cytochrome P450*; *GmCYP450*) in the GM16:22,756,017–23,154,638 region.

3.4.5 Mapping of eQTL Interactions

Using the greenhouse data from the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations, we performed a transcriptome-wide eQTL study (Gélinas Bélanger *et al.*, unpublished) using a combinatorial mapping approach with three algorithms (IM, ICIM, and GCIM) designed specifically to identify *cis* and *trans* quantitative e-traits. From these results, we identified several e-traits regulated by the *MergGM04gh* region identified in this present study in the QS15544$_{RIL}$ population. For the QS15544$_{RIL}$ population, we identified a total of 13 e-traits regulated by the *MergGM04gh* region (Fig. 2.7; Appendix 1.10). The e-traits were identified on six chromosomes (GM01, GM04, GM11, GM12, GM19, and GM20) with chromosome GM04 having the highest number of e-traits, seven in total. The *Glyma.04G165400* gene was found to be regulated by *cis* and *trans* interactions from regions located in close physical proximity. Two of the regulated genes were *Glyma.04G050200* (*Early flowering 3/E6* locus; *GmELF3*) and *Glyma.12G048500* (*Target of FLC And SVP1; GmTFS1*), the former being as a light *Zeitnehmer (*"*time-taker*"*)* and thermosensor circadian clock component in *Arabidopsis* and the latter being an AP2/B3-like transcriptional factor promoting floral transition in *Arabidopsis*. Both of them were regulated by *trans* interactions. No eQTL interactions were identified for the *MergGM04gh* region in the QS15524$_{F2:F3}$ population.

3.5 Discussion

3.5.1 Chromosome GM04 is a Hub for Early Reproductive Traits

Chromosome GM04 is known to host several major loci (e.g., *E6* and *E8*) and genes (e.g., *E1La* and *E1Lb*) that are involved in the regulation of early reproductive traits (Zhang *et al.*, 2017c; Gupta *et al.*, 2022). In addition, this chromosome is known to host a large number of *Arabidopsis* orthologs (52 genes out of 836) involved in flowering (Zhang *et al.*, 2017c). Dissecting QTL regions from this chromosome is challenging due to the close proximity and interplay of several of these orthologous flowering genes, as can be observed in Kong *et al.* (2018) and Wang *et al.* (2018) in which the QTL regions encompassed GM04:13,212,370–43,843,500 and GM04:7,166,748-44,508,948, respectively. In this study, we generated high-density GBS-derived linkage maps for chromosome GM04 in two plant populations and performed QTL mapping using a combinatorial approach composed of two mapping algorithms (ICIM and GCIM) with the intent of dissecting the large QTL region normally identified on this chromosome.

In the present study, three distinct loci were identified within the *E8* locus: GM04:16,974,874-17,152,230 (*ICIM_44_fill2*), *MergGM04f*, and *MergGM04gh*. In both populations, the greenhouse (*MergGM04gh*; GM04:41,808,599-42,376,237) and field (*MergGM04f* region; GM04:35,168,111-37,664,017) QTL were identified nearby on the same chromosome. We consider that the *MergGM04gh* and *MergGM04f* regions are distinct due to the large distance separating the regions and the different photoperiod of each growth system (e.g., fluctuating long days in the field vs. constant short days in the greenhouse). Our results demonstrate that *E8* is regulated by three distinct genomic regions on chromosome GM04, which all encompass or are closely located to flowering genes. To dissect *E8* into smaller regions, we decided to split the locus into three distinct regions using the following nomenclature; (i) *E8-r1*, which corresponds to the GM04:16,974,874-17,152,230 identified in $QS15544_{RIL}$ (Table 2.4); (ii) *E8-r2*, which corresponds to the *MergGM04f* (position GM04:35,168,111-37,664,017) identified in both populations (Table 2.3); and (iii) *E8-r3*, which corresponds to the *MergGM04gh* (position GM04:41,808,599-42,376,237) region identified in both populations under greenhouse conditions (Table 2.3). All three regions, listed as ECqMG-4.1 for *E8-r1*, qMG-4.3 for *E8-r2*, and ECqMG-4.4 for *E8-r3*, were previously identified in a genome-wide association study (GWAS) performed with a 16,879 accessions panel (Zimmer *et al.*, 2021); however, all of them were only associated with late maturity (MG0 and above) and none with super early maturity (i.e., MG000-MG00) such as the lines used in this study. To the best of our knowledge, this is the first time these alleles are reported for cultivars belonging to the MGs 000 and 00. Additionally, this is the first time these alleles have been demonstrated to have cumulative additive effects to generate an early maturity phenotype. Overall, the high-heritability values for each of the pod-filling and number of days to maturity traits suggest that these QTL could be used in the breeding of early maturing cultivars.

3.5.2 *E8-r1* Locus

The *E8-r1* (GM04:16,974,874-17,152,230) region comprises nine genes and has a high impact on pod-filling (LOD 13.2 and PVE 27.4%), leading to earlier phenotype by 1.81 (ICIM) days in $QS15544_{RIL}$ (Table 2.4). As previously mentioned, the statistical associations with this region were more challenging to map, with QTL identified starting at GM04:15,748,916 and ending at GM04:18,312,993 with each season's data and pairwise average for each season's pair (Appendix 1.9). None of the nine genes found within the region were previously found to be

associated with reproductive phenotypes in soybean or *Arabidopsis* in the literature; however, we identified two variants, GM04:16,097,210 (*Glyma.04G124300*) and GM04:16,331,703 (*Glyma.04G124600*), located in neighbouring genes that were previously identified as potential candidates for *E8* in Sadowski (2020) (Table 2.6). The GM04:16,097,210 SNP is a G→T SIFT-Tolerated missense polymorphism located at the amino acid position 375 of *Glyma.04G124300*. This polymorphism was found to be present only in "AAC Mandor" and possibly causes a longer pod-filling. The *Glyma.04G124300* gene belongs to the FAR1/FHY3 family which are essential proteins involved in the phytochrome A controlled far-red responses (Lin and Wang, 2004) and positive regulators of chlorophyll biosynthesis (Tang *et al.*, 2012) in *Arabidopsis*. Furthermore, this family is also involved in the activation of the gene expression of *Circadian clock associated1* (*AtCCA1*) in *Arabidopsis* which serves as a key component of the core oscillator of the circadian clock (Liu *et al.*, 2020b).

In Sadowski (2020), *Glyma.04G124300* was considered as a promising candidate, but inferior to *Glyma.04G124600*, another member of the FAR1/FHY3 family. In our variant analysis, *Glyma.04G124600* exhibits a SIFT-Deleterious missense polymorphism C→T on the third exon at amino acid position 350 in "AAC Mandor"; however, "AAC Mandor" is heterozygous for this polymorphism, and more investigation would be required to know if this SNP could be causal. In addition, a T→G stop-gain variant was identified in *Glyma.04G124800*, an ortholog of the *Arabidopsis* gene *AtZIFL1*. In *Arabidopsis*, this gene is known to be involved in root development, gravitropism, stomatal movements, and basipetal auxin transport (Remy *et al.*, 2013). Its unconfirmed role in maturity makes *GmZIFL1* less likely to be the regulator at the source of the GM04:16,974,874-17,152,230 region although its polymorphism is predicted to be highly deleterious. On the whole, our results suggest that *Glyma.04G124300* and *Glyma.04G124600* are currently the best candidate genes for the *E8-r1* locus.

3.5.3 *E8-r2* Locus

The *E8-r2* locus (*MergGM04f* region) comprises seven QTL (four in $QS15524_{F2:F3}$ and three in $QS15544_{RIL}$) with important effects on the observed phenotypes, especially those identified for the $QS15524_{F2:F3}$ population (Table 2.3). In the $QS15524_{F2:F3}$ population, the additive effects identified for this region represented an average earlier pod-filling phenotype of 2.85 (GCIM)/3.58 (ICIM) days and an average earlier maturing phenotype of 3.19 (GCIM)/3.85 (ICIM) days for the

"OAC Vision" allele. In the QS15544$_{RIL}$ population, this additive effect caused an average earlier maturity of 1.27 (ICIM; GM04:35,168,111-35,533,929 sub-region) and 1.81 (ICIM; GM04:37,662,935-37,664,017 sub-region) days in the offspring having the "9004" allele. It is currently impossible to attest if the QTL observed in the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations stem from the same or different regulators; however, based on an analysis of the SNPs identified in the GmHapMap dataset and located within coding regions of genes located within *E8-r2*, a high homology exists within SNPs of the later maturing parental lines ("Maple Arrow" and "AAC Mandor") versus the earlier maturing parental lines ("OAC Vision" and "9004") (data not shown). Consequently, this evidence suggests that the causal variants might be the same. To the best of our knowledge, no candidate genes have yet been proposed for this locus, despite being located within the *E8* large-range region and close to a GWAS hit (GM04:38,274,140) identified by Zhang *et al.* (2015).

The narrow *E8-r2* sub-region of the QS15524$_{F2:F3}$ population (GM04:36,499,381-36,941,521) comprises only six genes, including *E1la*, a major transcription factor involved in flowering and maturity that has been validated for the *Tof4* QTL (Liu *et al.*, 2022a; Dong *et al.*, 2023) (Table 2.5). Silencing of *E1la* using virus-induced gene silencing upregulates the expression of *GmFT2a* and *GmFT5a*, leading to earlier flowering (Xu *et al.*, 2015). In our study, a G→A 3′UTR polymorphism was identified at position GM04:36,758,687 in both "OAC Vision" and "9004", which are the providers of the allele causing an earlier maturity. The 3′UTR region is involved in a plethora of functions, such as RNA stability, translation, and localization, and harbors binding sites for microRNAs and RNA-binding proteins (Steri *et al.*, 2018). In consequence, polymorphisms in a binding site can lead to modifications in the level of gene expression. The presence of *E1la* in the narrow *E8-r2* QS15524$_{F2:F3}$ sub-region of the QS15524$_{F2:F3}$ population and the fact that none of the five other proposed SNPs are located in flowering orthologs suggest that *E1la* is the best candidate for the *E8-r2* region.

3.5.4 *E8-r3* Locus

The *MergGM04gh* region is the only region associated with the number of days to maturity in the greenhouse phenotype and was identified in both populations with ≈200 Kbp separating the QS15524$_{F2:F3}$ QTL from those observed in QS15544$_{RIL}$, suggesting that the causal variant could be the same (Table 2.3). The *MergGM04gh* is related to an earlier maturity phenotype by 3.21

(GCIM)/3.73 (ICIM) days in the $QS15524_{F2:F3}$ population and 2.07 (ICIM) days in the $QS15544_{RIL}$ population under constant short days. Based on our QTL analysis, this earlier flowering phenotype is provided by 'OAC Vision' and "9004" in $QS15524_{F2:F3}$ and $QS15544_{RIL}$, respectively. Overlapping or closely located biparental and GWAS QTL have been previously identified by Wang *et al.* (2015), Sun *et al.* (2013), Zhang *et al.* (2015), Mao *et al.* (2017) and Liu *et al.* (2021), with several candidate genes being proposed. In our study, the *MergGM04gh* region comprises 28 genes, including two candidate genes with polymorphisms of interest: *Glyma.04G168300* (*GmCDF3*) and *Glyma.04G167900* (*GmLHCA4*) (Table 2.5). Another gene of interest, *Glyma.04G166300* (*GmPRR1a*), is located at ≈50 Kbp upstream of the region. *Glyma.04G168300* (*GmCDF3*) is a Dof-type zinc finger domain-containing transcription factor that was suggested as a candidate maturity gene by Mao *et al.* (2017). Corrales *et al.* (2017) recently demonstrated that *AtCDF3* overexpression promotes late flowering partly by controlling the expression of the CBF/DREB2A-CRT/DRE and ZAT10/12 modules in the Columbia (Col-0) ecotype. To the best of our knowledge, its impact on soybean flowering has not been validated yet.

In our study, a C→A missense SIFT-Tolerated missense polymorphism has been identified at amino acid position 306 in *Glyma.04G168300*/*GmCDF3*. Based on our analysis of the variants, "OAC Vision" and "9004" exhibit the same genotypes for this polymorphism, supporting it as a potential candidate gene for this region. Additionally, we detected four SNPs (positions GM04:42,126,107, GM04:42,126,847, GM04:42,126,965, and GM04:42,127,008) located in the 3'UTR region of *Glyma.04G167900* (*GmLHCA4*). Overall, these four variants all display the same genotype pattern, with "OAC Vision" and "9004" being the providers of the early flowering alleles. Liu *et al.* (2021) investigated the role of *Glyma.04G167900* (*GmLHCA4*) and observed a 1.8-day difference in the number of days to flowering between two *GmLHCA4* haplotypes. The PSEUDO RESPONSE REGULATOR (PRR) family regulates many biological processes in *Arabidopsis*, including photoperiodic flowering, growth, stress response, and regulation of the circadian clock (Hayama *et al.*, 2017; Li *et al.*, 2019; Kim *et al.*, 2022), with several homologs found within the soybean genome. The domestication of the *Glyma.12G073900* (*GmPRR3b*) gene in soybean has been associated with an early flowering phenotype due to the presence of a causal SNP at position GM12:5,520,945 (Li *et al.*, 2020a). In our study, we identified a G→T polymorphism in the 5'UTR region at position GM04:41,757,388 of the *Glyma.04G166300* (*GmPRR1a*) gene that is present in "OAC Vision" and "9004"

(heterozygous). On the whole, our results suggest that *Glyma.04G168300* (*GmCDF3*), *Glyma.04G167900* (*GmLHCA4*), and *Glyma.04G166300* (*GmPRR1a*) are the best candidates for *E8-r3*.

3.5.5 Unique QTL in the QS15544$_{RIL}$ Population

Using our combinatorial approach, we detected four additional QTL regions (i.e., GM07:5,256,305-5,404,971; GM16:5,680,173-5,730,237; and GM16:22,756,017-23,154,638) that were identified only in the QS15544$_{RIL}$ population, possibly due to a higher number of recombination events and a greater statistical power due to the decreased number of heterozygotes in comparison to QS15524$_{F2:F3}$. Following the identification of these unique regions, those were narrowed to a total of 11 candidates with our five-step variant calling pipeline. For the GM07:5,256,305-5,404,971 region, we identified that the inbred lines carrying the "9004" allele mature between 1.15 (GCIM) and 1.30 (ICIM) days earlier than those harboring the "AAC Mandor" allele. This region was previously identified by Wang *et al.* (2004) with the Satt567 (position GM07:4,559,602) and Satt463 (position GM07:8,283,465) markers, with four QTL reported in Soybase (i.e., Pod maturity 14-4, First flower 6-1, Pod maturity 10-2 and Reproductive stage length 4-3). Cheng *et al.* (2011) also identified a QTL between Satt540 (position GM04:5,010,696) and Satt435 (Soybase biparental QTL Reproductive stage length 5-4). For the GM07:5,256,305-5,404,971 region, we identified a SIFT-Tolerated missense polymorphism at position GM07:5,200,811 of the *Glyma.07G058200* (*GmSPA1*) gene. Han *et al.* (2021) identified a GWAS QTL at position GM7:5,059,730 for the number of days to flowering in soybean and proposed *GmSPA1* as the best candidate for this hit. In *Arabidopsis*, *AtSPA1* is a WD (tryptophan–aspartic acid)–repeat protein involved in the regulation of the circadian clock and photomorphogenesis in a light-responsive repressor manner (Hoecker *et al.*, 1999; Yang *et al.*, 2005).

The GM16:5,680,173-5,730,237 region has an impact on the number of days to maturity of the QS15544$_{RIL}$ population, with the offspring harboring the "AAC Mandor" allele reaching maturity 1.51 (GCIM)/1.55 (ICIM) days before the ones harboring the "9004" allele. This region lies close (~1.5 Mbp) to *Glyma.16G044100* (*GmFT5a*) and *Glyma.16G044200* (*GmFT3a*), two major homologs involved in flowering and maturity (Liu *et al.*, 2018b; Lee *et al.*, 2021). The region is close to the GWAS QTL First Flower 4-g63 (position GM16:5,799,540) (Mao *et al.*, 2017). No

gene has been proposed by Mao *et al.* (2017) for this region. Using our pipeline, we did not find any promising variants within the region; however, four putatively deleterious SNPs were identified upstream of the region in *Glyma.16G044100* (*GmFT5a*), *Glyma.16G050300* (*Fusca3*; *GmFUS3*), and *Glyma.16G057200* (*Baf60*; *GmBAF60*).

3.6 Conclusion

In conclusion, the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ plant populations were generated using fixed alleles for *E1–E4*, which enabled us to identify overlapping regions and unique QTL regions involved in reproductive traits. Our results demonstrate that the major *E8* locus is composed of three separate regions (*E8-r1, E8-r2*, and *E8-r3*) with major additive effects. In addition, we demonstrate that eQTL interactions with the major flowering gene *GmELF3*/*E6* and 12 other e-traits stem from regions located within *E8-r3* or nearby. Several other unique QTL regions regulating reproductive traits were also identified in QS15544$_{RIL}$ on chromosomes GM07, GM08, and GM16. With our five-step variant calling pipeline, we were able to identify candidate SNPs and genes located within or near all of the identified QTL regions. Altogether, our results demonstrate that novel major genes controlling early maturity can still be identified and incorporated into early maturing material. Nonetheless, in-depth functional characterization of these candidate genes remains necessary to confirm their role in early pod-filling and maturity.

3.7 Supplemental data

Appendix 1.1 Phenotypic data associated with each of the lines of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.2 Descriptive statistics associated with the four phenotypes for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.3 Pearson correlations associated with each of the phenotypes of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.4 Statistical analyses associated with each of the field phenotypes of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.5 Genotypes associated with each marker of the QS15524$_{F2:F3}$ population.

Appendix 1.6 Genotypes associated with each marker of the QS15544$_{RIL}$ population.

Appendix 1.7 Linkage maps for both populations.

Appendix 1.8 QTL analyses for each of the studied year for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Appendix 1.9 Minor QTL regions identified in each of the populations.

Appendix 1.10 Expression QTL regions for the *E8-r3* locus.

3.8 Information

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: PRJNA1035514.

Author contribution

JG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Software. TC: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing, Data curation, Investigation, Software. VH-V: Supervision, Writing – review & editing. LO'D: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing, Investigation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Génome Québec and Génome Canada with funds awarded to the SoyaGen Project (project #5801) and by the Canadian Field Crop Research Alliance and Agriculture and Agri-Food Canada under the Agri-Innovation Program (#ASC-09). JG was supported by the Natural Sciences and Engineering Research Council of Canada, les Fonds de recherche du Québec volet Nature et Technologie, Centre SÈVE, and Seed World Group.

Acknowledgments

We would like to thank Dr. Martine Jean and Vincent-Thomas Boucher St-Amour for their advice on QTL mapping and linkage map construction. We would like to acknowledge the work of Éric Fortier for the extraction of RNA and phenotypic data collection. Thanks to Joannie Berthon, Maxime Carrier, Dominique Poulin, and Daphnée Paré for the phenotypic data collection in the greenhouse and the field.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

3.9 References

Abed, A., Légaré, G., Pomerleau, S., St-Cyr, J., Boyle, B., Belzile, F. J. (2019). Genotyping-by-sequencing on the ion torrent platform in barley. In Harwood, W. (eds) *Barley. Methods in Molecular Biology* (New York, NY: Humana Press), vol 1900. doi: 10.1007/978-1-4939-8944-7_15

Agronomix Software Inc. (2009). *Agrobase Generation II*. Available at: https://www.agronomix.com/.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed May 25, 2021).

Bagg, J., Banks, S., Baute, T., Bohner, H., Brown, C., Cowbrough, M., *et al.* (2009). *Agronomy Guide for Field Crops*. Ed. Brown, C. (Toronto, Ontario, Canada: Ministry of Agriculture, Food and Rural Affairs).

Bolger, A. M., Lohse, M., Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., *et al.* (2019). R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics* 211, 495–502. doi: 10.1534/genetics.118.301595

Broman, K. W., Wu, H., Sen, Ś., Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Browning, B. L., Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Cheng, L., Wang, Y., Zhang, C., Wu, C., Xu, J., Zhu, H., *et al.* (2011). Genetic analysis and QTL detection of reproductive period and post-flowering photoperiod responses in soybean. *Theor. Appl. Genet.* 123, 421–429. doi: 10.1007/s00122-011-1594-8

Cober, E. R., Molnar, S. J., Charette, M., Voldeng, H. D. (2010). A new locus for early maturity in soybean. *Crop Sci.* 50, 524–527. doi: 10.2135/cropsci2009.04.0174

Corrales, A. R., Carrillo, L., Lasierra, P., Nebauer, S. G., Dominguez-Figueroa, J., Renau-Morata, B., *et al.* (2017). Multifaceted role of cycling DOF factor 3 (CDF3) in the regulation of flowering time and abiotic stress responses in *Arabidopsis*. *Plant Cell Environ.* 40, 748–764. doi: 10.1111/pce.12894

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dong, L., Cheng, Q., Fang, C., Kong, L., Yang, H., Hou, Z., *et al.* (2022). Parallel selection of distinct Tof5 alleles drove the adaptation of cultivated and wild soybean to high latitudes. *Mol. Plant* 15, 308–321. doi: 10.1016/j.molp.2021.10.004

Dong, L., Fang, C., Cheng, Q., Su, T., Kou, K., Kong, L., *et al.* (2021). Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nat. Commun.* 12, 1–11. doi: 10.1038/s41467-021-25800-3

Dong, L., Li, S., Wang, L., Su, T., Zhang, C., Bi, Y., *et al.* (2023). The genetic basis of high-latitude adaptation in wild soybean. *Curr. Biol.* 33, 252–262.e4. doi: 10.1016/j.cub.2022.11.061

Fehr, W. R. (1980). *Soybean. Hybrid. Crop plants*. (Wisconsin, USA: American Society of Agronomy and Crop Science Society of America, Publishers Madison) 589–599.

Grainger, C. M., Rajcan, I. (2014). Characterization of the genetic changes in a multi-generational pedigree of an elite Canadian soybean cultivar. *Theor. Appl. Genet.* 127, 211–229. doi: 10.1007/s00122-013-2211-9

Grant, D., Nelson, R. T., Cannon, S. B., Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, 843–846. doi: 10.1093/nar/gkp798

Gupta, S., Kumawat, G., Agrawal, N., Tripathi, R., Rajesh, V., Nataraj, V., *et al.* (2022). Photoperiod trait: Insight in molecular mechanism for growth and maturity adaptation of soybean (Glycine max) to different latitudes. *Plant Breed.* 141, 483–500. doi: 10.1111/pbr.13041

Han, X., Xu, Z. R., Zhou, L., Han, C. Y., Zhang, Y. M. (2021). Identification of QTNs and their candidate genes for flowering time and plant height in soybean using multi-locus genome-wide association studies. *Mol. Breed.* 41, 39. doi: 10.1007/s11032-021-01230-3

Harada, K., Watanabe, S., Zhengjun, X., Tsubokura, Y., Yamanaka, N., Anai, T. (2011). Positional cloning of the responsible genes for maturity loci E1, E2 and E3 in soybean. *Soybean - Genet. Nov. Tech. Yield Enhanc.*, 51–76. doi: 10.5772/21085

Hayama, R., Sarid-Krebs, L., Richter, R., Fernández, V., Jang, S., Coupland, G. (2017). PSEUDO RESPONSE REGULATORs stabilize CONSTANS protein to promote flowering in response to day length. *EMBO J.* 36, 904–918. doi: 10.15252/embj.201693907

Hoecker, U., Tepperman, J. M., Quail, P. H. (1999). SPA1, a WD-repeat protein specific to phytochrome A signal transduction. *Sci. (80-.).* 284, 496–499. doi: 10.1126/science.284.5413.496

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y. Y., Nelson, R. L., Costa, J. M., *et al.* (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci.* 103, 16666–16671. doi: 10.1073/pnas.0604379103

Jia, H., Jiang, B., Wu, C., Lu, W., Hou, W., Sun, S., *et al.* (2014). Maturity group classification and maturity locus genotyping of early-maturing soybean varieties from high-latitude cold regions. *PloS One* 9 (4). doi: 10.1371/journal.pone.0094139

Jiang, B., Nan, H., Gao, Y., Tang, L., Yue, Y., Lu, S., *et al.* (2014). Allelic combinations of soybean maturity loci E1, E2, E3 and E4 result in diversity of maturity and adaptation to different latitudes. *PloS One* 9 (8). doi: 10.1371/journal.pone.0106042

Jiang, B., Zhang, S., Song, W., Khan, M. A. A., Sun, S., Zhang, C., *et al.* (2019). Natural variations of FT family genes in soybean varieties covering a wide range of maturity groups. *BMC Genomics* 20, 1–16. doi: 10.1186/s12864-019-5577-5

Joshi, N. (2011) *Sabre - A barcode demultiplexing and trimming tool for FastQ files*. Available at: https://github.com/najoshi/sabre (Accessed May 25, 2021).

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14, R36. doi: 10.1186/gb-2013-14-4-r36

Kim, N. S., Yu, J., Bae, S., Kim, H. S., Park, S., Lee, K., *et al.* (2022). Identification and characterization of PSEUDO-RESPONSE REGULATOR (PRR) 1a and 1b genes by CRISPR/cas9-targeted mutagenesis in Chinese cabbage (Brassica rapa L.). *Int. J. Mol. Sci.* 23, 1–15. doi: 10.3390/ijms23136963

Kong, F., Liu, B., Xia, Z., Sato, S., Kim, B. M., Watanabe, S., *et al.* (2010). Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in Soybean. *Plant Physiol.* 154, 1220–1231. doi: 10.1104/pp.110.160796

Kong, L., Lu, S., Wang, Y., Fang, C., Wang, F., Nan, H., *et al.* (2018). Quantitative trait locus mapping of flowering time and maturity in soybean using next-generation sequencing-based analysis. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00995

Kong, F., Nan, H., Cao, D., Li, Y., Wu, F., Wang, J., *et al.* (2014). A new dominant gene E9 conditions early flowering and maturity in soybean. *Crop Sci.* 54, 2529–2535. doi: 10.2135/cropsci2014.03.0228

Kou, K., Yang, H., Li, H., Fang, C., Chen, L., Yue, L., *et al.* (2022). A functionally divergent SOC1 homolog improves soybean yield and latitudinal adaptation. *Curr. Biol.* 32, 1728–1742. doi: 10.1016/j.cub.2022.02.046

Kumar, P., Henikoff, S., Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1082. doi: 10.1038/nprot.2009.86

Lee, S. H., Choi, C. W., Park, K. M., Jung, W. H., Chun, H. J., Baek, D., *et al.* (2021). Diversification in functions and expressions of soybean FLOWERING LOCUS T genes fine-tunes seasonal flowering. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.613675

Lee, G. A., Crawford, G. W., Liu, L., Sasaki, Y., Chen, X. (2011). Archaeological soybean (Glycine max) in East Asia: Does size matter? *PloS One* 6 (11). doi: 10.1371/journal.pone.0026720

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, C., Li, Y. H., Li, Y., Lu, H., Hong, H., Tian, Y., *et al.* (2020). A domestication-associated gene gmPRR3b regulates the circadian clock and flowering time in soybean. *Mol. Plant* 13, 745–759. doi: 10.1016/j.molp.2020.01.014

Li, B., Wang, Y., Zhang, Y., Tian, W., Chong, K., Jang, J. C., *et al.* (2019). PRR5, 7 and 9 positively modulate TOR signaling-mediated root cell proliferation by repressing TANDEM ZINC FINGER 1 in *Arabidopsis*. *Nucleic Acids Res.* 47, 5001–5015. doi: 10.1093/nar/gkz191

Lin, X., Liu, B., Weller, J. L., Abe, J., Kong, F. (2020). Molecular mechanisms for the photoperiodic regulation of flowering in soybean. *J. Integr. Plant Biol.* 63 (6), 981–994. doi: 10.1111/jipb.13021

Lin, C.-S., Poushinsky, G. (1983). A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. *Biometrics* 39 (3), 553–561. doi: 10.2307/2531083

Lin, C.-S., Poushinsky, G. (1985). A modified augmented design (type 2) for rectangular plots. *Can. J. Plant Sci.* 65, 743–749. doi: 10.4141/cjps85-094

Lin, R., Wang, H. (2004). *Arabidopsis* FHY3/FAR1 gene family and distinct roles of its members in light control of *Arabidopsis* development. *Plant Physiol.* 136, 4010–4022. doi: 10.1104/pp.104.052191

Liu, C., Chen, X., Wang, W., Hu, X., Han, W., He, Q., *et al.* (2021). Identifying wild versus cultivated gene-alleles conferring seed coat color and days to flowering in Soybean. *Int. J. Mol. Sci.* 22, 1–22. doi: 10.3390/ijms22041559

Liu, L., Gao, L., Zhang, L., Cai, Y., Song, W., Chen, L., *et al.* (2022). Co-silencing E1 and its homologs in an extremely late-maturing soybean cultivar confers super-early maturity and adaptation to high-latitude short-season regions. *J. Integr. Agric.* 21, 326–335. doi: 10.1016/S2095-3119(20)63391-3

Liu, W., Jiang, B., Ma, L., Zhang, S., Zhai, H., Xu, X., *et al.* (2018). Functional diversification of Flowering Locus T homologs in soybean: GmFT1a and GmFT2a/5a have opposite roles in controlling flowering and maturation. *New Phytol.* 217, 1335–1345. doi: 10.1111/nph.14884

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180, 995–1007. doi: 10.1534/genetics.108.092742

Liu, Y., Ma, M., Li, G., Yuan, L., Xie, Y., Wei, H., *et al.* (2020). Transcription factors FHY3 and FAR1 regulate light-induced CIRCADIAN CLOCK ASSOCIATED1 gene expression in *Arabidopsis*. *Plant Cell* 32, 1464–1478. doi: 10.1105/tpc.19.00981

Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., *et al.* (2010). The soybean stem growth habit gene Dt1 is an ortholog of *Arabidopsis* TERMINAL FLOWER1. *Plant Physiol.* 153, 198–210. doi: 10.1104/pp.109.150607

Lu, S., Dong, L., Fang, C., Liu, S., Kong, L., Cheng, Q., *et al.* (2020). Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. *Nat. Genet.* 52, 428–436. doi: 10.1038/s41588-020-0604-7

Mao, T., Li, J., Wen, Z., Wu, T., Wu, C., Sun, S., *et al.* (2017). Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. *BMC Genomics* 18, 1–17. doi: 10.1186/s12864-017-3778-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17 (1), 10–12. doi: 10.14806/ej.17.1.200

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., *et al.* (2016). The ensembl variant effect predictor. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0974-4

McWilliams, D., Berglund, D. R., Endres, G. J. (2004). *Soybean - Growth and Management Quick Guide* (Fargo, North Dakota, USA: North Dakota State University, 1–8. Available at: http://www.marchutletseeds.ca/uploads/soybeans_soybeanstages.pdf.

Meng, L., Li, H., Zhang, L., Wang, J. (2015). QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Miao, C., Fang, J., Li, D., Liang, P., Zhang, X., Yang, J., *et al.* (2018). Genotype-Corrector: Improved genotype calls for genetic mapping in F2 and RIL populations. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-28294-0

Ng, P. C., Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Ouellette, L. A., Reid, R. W., Blanchard, S. G., Brouwer, C. R. (2018). LinkageMapView-rendering high-resolution linkage and QTL maps. *Bioinformatics* 34, 306–307. doi: 10.1093/bioinformatics/btx576

Pagano, M. C., Miransari, M. (2016). "The importance of soybean production worldwide," in *Abiotic and Biotic Stresses in Soybean Production* (Waltham, MA, USA: Elsevier Inc), 1–26. doi: 10.1016/B978-0-12-801536-0.00001-3

Ping, J., Liu, Y., Sun, L., Zhao, M., Li, Y., She, M., *et al.* (2014). Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. *Plant Cell* 26, 2831–2842. doi: 10.1105/tpc.114.126938

R Core Team (2010). *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Found. Stat. Comput). Available at: http://www.gnu.org/copyleft/gpl.html.

Remy, E., Cabrito, T. R., Baster, P., Batista, R. A., Teixeira, M. C., Friml, J., *et al.* (2013). A Major Facilitator Superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in *Arabidopsis*. *Plant Cell* 25, 901–926. doi: 10.1105/tpc.113.110353

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., *et al.* (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Sadowski, M. (2020). *A functional genomics approach in identifying the underlying gene for the E8 maturity locus in soybean (Glycine max)* (Department of Biology, Carleton University). Master thesis. Available at: https://curve.carleton.ca/5c96c758-a146-4cd6-aadd-d7d83398fd3e.

Samanfar, B., Molnar, S. J., Charette, M., Schoenrock, A., Dehne, F., Golshani, A., *et al.* (2017). Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. *Theor. Appl. Genet.* 130, 377–390. doi: 10.1007/s00122-016-2819-7

Steri, M., Idda, M. L., Whalen, M. B., Orrù, V. (2018). Genetic variants in mRNA untranslated regions. *Wiley Interdiscip. Rev. RNA* 9, e1474. doi: 10.1002/wrna.1474

Sun, H., Jia, Z., Cao, D., Jiang, B., Wu, C., Hou, W., *et al.* (2011). GmFT2a, a soybean homolog of flowering locus T, is involved in flowering transition and maintenance. *PloS One* 6, 18–20. doi: 10.1371/journal.pone.0029238

Sun, S., Kim, M. Y., Van, K., Lee, Y. W., Li, B., Lee, S. H. (2013). *QTL* for resistance to Phomopsis seed decay are associated with days to maturity in soybean (Glycine max). *Theor. Appl. Genet.* 126, 2029–2038. doi: 10.1007/s00122-013-2115-8

Tang, W., Wang, W., Chen, D., Ji, Q., Jing, Y., Wang, H., *et al.* (2012). Transposase-derived proteins FHY3/FAR1 interact with PHYTOCHROME-INTERACTING FACTOR1 to regulate chlorophyll biosynthesis by modulating HEMB1 during deetiolation in *Arabidopsis*. *Plant Cell* 24, 1984–2000. doi: 10.1105/tpc.112.097022

Tardivel, A., Torkamaneh, D., Lemay, M.-A., Belzile, F., O'Donoughue, L. S. (2019). A systematic gene-centric approach to define haplotypes and identify alleles on the basis of dense single nucleotide polymorphism datasets. *Plant Genome* 12, 180061. doi: 10.3835/plantgenome2018.08.0061

Taylor, J., Butler, D. (2017). R package ASMap: efficient genetic linkage map construction and diagnosis. *J. Stat. Software* 79, 1–29. doi: 10.18637/jss.v079.i06

Thakare, D., Kumudini, S., Dinkins, R. D. (2011). The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in *Arabidopsis*. *Planta* 234, 933–943. doi: 10.1007/s00425-011-1450-8

The American Soybean Association (2023) *SoyStats - International: World Soybean Production, (2021/2022 year). Am. Soybean Assoc*. Available at: http://soystats.com/ (Accessed December 11, 2023).

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., Belzile, F. (2017). Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinf.* 18, 1–7. doi: 10.1186/s12859-016-1431-9

Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., *et al.* (2018). Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnol. J.* 16, 749–759. doi: 10.1111/pbi.12825

Torkamaneh, D., Laroche, J., Valliyodan, B., O'Donoughue, L., Cober, E., Rajcan, I., *et al.* (2020). Soybean (Glycine max) Haplotype Map (GmHapMap): a universal resource for soybean translational and functional genomics. *Plant Biotechnol. J.* 19 (2), 1–11. doi: 10.1111/pbi.13466

Tsubokura, Y., Matsumura, H., Xu, M., Liu, B., Nakashima, H., Anai, T., *et al.* (2013). Genetic variation in soybean at the maturity locus e4 is involved in adaptation to long days at high latitudes. *Agronomy* 3, 117–134. doi: 10.3390/agronomy3010117

Tsubokura, Y., Watanabe, S., Xia, Z., Kanamori, H., Yamagata, H., Kaga, A., *et al.* (2014). Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean. *Ann. Bot.* 113, 429–441. doi: 10.1093/aob/mct269

Tukey, J. W. (1977). *Exploratory data analysis* (Reading, Mass: Addison-Wesley Pub. Co.).

Wang, Y., Cheng, L., Leng, J., Wu, C., Shao, G., Hou, W., *et al.* (2015). Genetic analysis and quantitative trait locus identification of the reproductive to vegetative growth period ratio in soybean (Glycine max (L.) Merr.). *Euphytica* 201, 275–284. doi: 10.1007/s10681-014-1209-y

Wang, D., Graef, G. L., Procopiuk, A. M., Diers, B. W. (2004). Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theor. Appl. Genet.* 108, 458–467. doi: 10.1007/s00122-003-1449-z

Wang, J., Kong, L., Yu, K., Zhang, F., Shi, X., Wang, Y., *et al.* (2018). Development and validation of InDel markers for identification of QTL underlying flowering time in soybean. *Crop J.* 6, 126–135. doi: 10.1016/j.cj.2017.08.001

Watanabe, S., Hideshima, R., Zhengjun, X., Tsubokura, Y., Sato, S., Nakamoto, Y., *et al.* (2009). Map-based cloning of the gene associated with the soybean maturity locus E3. *Genetics* 182, 1251–1262. doi: 10.1534/genetics.108.098772

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., *et al.* (2011). A map-based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering. *Genetics* 188, 395–407. doi: 10.1534/genetics.110.125062

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., *et al.* (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2155–E2164. doi: 10.1073/pnas.1117982109

Xu, M., Xu, Z., Liu, B., Kong, F., Tsubokura, Y., Watanabe, S., *et al.* (2013). Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. *BMC Plant Biol.* 13, 91. doi: 10.1186/1471-2229-13-91

Xu, M., Yamagishi, N., Zhao, C., Takeshima, R., Kasai, M., Watanabe, S., *et al.* (2015). The soybean-specific maturity gene E1 family of floral repressors controls night-break responses through down-regulation of FLOWERING LOCUS T orthologs. *Plant Physiol.* 168, 1735–1746. doi: 10.1104/pp.15.00763

Yang, J., Lin, R., Hoecker, U., Liu, B., Xu, L., Wang, H. (2005). Repression of light signaling by *Arabidopsis* SPA1 involves post-translational regulation of HFR1 protein accumulation. *Plant J.* 43, 131–141. doi: 10.1111/j.1365-313X.2005.02433.x

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., *et al.* (2022). Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* 50, D996–D1003. doi: 10.1093/nar/gkab1007

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., *et al.* (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. *BMC Genomics* 16, 1–11. doi: 10.1186/s12864-015-1441-4

Zhang, S.-R. R., Wang, H., Wang, Z., Ren, Y., Niu, L., Liu, J., *et al.* (2017). Photoperiodism dynamics during the domestication and improvement of soybean. *Sci. China Life Sci.* 60, 1416–1427. doi: 10.1007/s11427-016-9154-x

Zhang, Y. W., Wen, Y. J., Dunwell, J. M., Zhang, Y. M. (2020). QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked *QTL* for quantitative traits in bi-parental segregation populations. *Comput. Struct. Biotechnol. J.* 18, 59–65. doi: 10.1016/j.csbj.2019.11.005

Zhu, X., Leiser, W. L., Hahn, V., Würschum, T. (2023). The genetic architecture of soybean photothermal adaptation to high latitudes. *J. Exp. Bot.* 74, 2987–3002. doi: 10.1093/jxb/erad064

Zimmer, G., Miller, M. J., Steketee, C. J., Jackson, S. A., de Tunes, L. V. M., Li, Z. (2021). Genetic control and allele variation among soybean maturity groups 000 through IX. *Plant Genome* 14, 1–25. doi: 10.1002/tpg2.20146

Linkage group	Number of markers	LG length (cM)	Average interval (cM)	Linkage group	Number of markers	LG length (cM)	Average interval (cM)
1	47	78.21	1.66	11a	52	54.60	1.05
2	134	200.20	1.49	11b	28	58.63	2.09
3	95	155.42	1.64	12	82	127.97	1.56
4	121	158.38	1.31	13a	11	33.30	3.03
5	86	153.66	1.79	13b	45	107.07	2.38
6	101	197.36	1.95	14a	46	79.91	1.74
7	86	174.26	2.03	14b	16	33.19	2.07
8a	4	26.24	6.56	15	85	204.10	2.40
8b	43	69.38	1.61	16	74	113.13	1.53
8c	12	24.66	2.06	17	100	191.03	1.91
9a	17	40.55	2.39	18	106	181.48	1.71
9b	87	183.60	2.11	19	55	128.08	2.33
10	51	93.34	1.83	20	29	103.29	3.56

Table 2.1 Linkage map characteristics of the QS15524$_{F2:F3}$ population.

Linkage group	Number of markers	LG length (cM)	Average interval (cM)	Linkage group	Number of markers	LG length (cM)	Average interval (cM)
1	123	224.10	1.82	10b	2	1.32	0.66
2	187	351.10	1.88	11a	50	99.87	2.00
3	228	419.03	1.84	11b	12	27.83	2.32
4	302	596.60	1.98	11c	15	44.36	2.96
5	137	258.57	1.89	12a	77	156.95	2.04
6a	51	84.74	1.66	12b	20	23.10	1.15
6b	181	268.96	1.49	13a	98	210.99	2.15
7a	72	111.90	1.55	13b	31	73.61	2.37
7b	4	14.31	3.58	14a	33	41.34	1.25
8a	130	280.55	2.16	14b	89	145.58	1.64
8b	9	12.31	1.37	15	84	180.38	2.15
9a	55	75.07	1.36	16	167	261.25	1.56
9b	3	3.39	1.13	17	78	208.70	2.68
9c	4	52.69	13.17	18a	4	2.32	0.58
9d	85	122.31	1.44	18b	86	172.01	2.00
9e	29	51.68	1.78	19	72	150.58	2.09
10a	132	168.30	1.28	20	96	162.52	1.69

Table 2.2 Linkage map characteristics of the QS15544$_{RIL}$ population.

Region	Trait	Population	QTL name	Linkage group	Position of the QTL peak (cM)	Confidence interval (cM)		QTL position		LOD	PVE (%)	Additive effect	Dominance effect
						Low	High	Left	Right				
MergGM04f (GM04:35,168,111-37,664,017)	Maturity	$QS15544_{RIL}$	*ICIM_44_mat1*	4	400.00	399.50	401.50	GM04:35,168,111	GM04:35,533,929	19.60	22.10	1.81	N/A
	Maturity		*GCIM_44_mat1*	4	400.49	N/A	N/A	GM04:35,533,929	GM04:35,533,929	7.10	8.70	1.79	N/A
	Maturity		*ICIM_44_mat2*	4	418.00	417.50	418.50	GM04:37,662,935	GM04:37,664,017	11.30	10.90	1.27	N/A
	Pod-filling	$QS15524_{F2:F3}$	*ICIM_24_fill1*	4	80.00	79.50	80.50	GM04:36,499,381	GM04:36,941,521	33.80	47.00	3.58	0.05
	Maturity		*ICIM_24_mat1*	4	80.00	79.50	80.50	GM04:36,499,381	GM04:36,941,521	41.40	48.20	3.85	0.26
	Pod-filling		*GCIM_24_fill1*	4	79.88	N/A	N/A	GM04:36,499,381	GM04:36,499,381	46.30	28.00	2.85	0
	Maturity		*GCIM_24_mat2*	4	79.88	N/A	N/A	GM04:36,499,381	GM04:36,499,381	51.60	29.60	3.19	0
MergGM04gh (GM04:41,808,599-42,376,237)	Maturity (greenhouse)	$QS15544_{RIL}$	*ICIM_44_matgh1*	4	497.00	495.50	497.50	GM04:42,368,274	GM04:42,376,237	5.80	15.70	2.07	N/A
	Maturity (greenhouse)	$QS15524_{F2:F3}$	*ICIM_24_matgh1*	4	84.00	83.50	84.50	GM04:41,808,599	GM04:42,156,365	12.40	29.40	3.73	-1.19
	Maturity (greenhouse)		*GCIM_24_matgh1*	4	83.57	N/A	N/A	GM04:41,808,599	GM04:41,808,599	11.90	18.70	3.21	0
MergGM08f (GM08:47,258,336-47,770,836)	Pod-filling	$QS15524_{F2:F3}$	*ICIM_24_fill2*	8c	14.00	10.50	17.50	GM08:47,258,336	GM08:47,289,756	6.30	5.90	-1.31	-0.10
	Maturity		*ICIM_24_mat4*	8c	14.00	11.50	16.50	GM08:47,258,336	GM08:47,289,756	11.60	8.70	-1.69	0.26
	Pod-filling		*GCIM_24_fill5*	8c	14.00	N/A	N/A	GM08:47,258,336	GM08:47,289,756	13.80	5.20	-1.22	0
	Maturity		*GCIM_24_mat6*	8c	13.65	N/A	N/A	GM08:47,258,336	GM08:47,258,336	13.60	4.40	-1.23	0
	Maturity	$QS15544_{RIL}$	*ICIM_44_mat6*	8b	12.00	11.50	12.00	GM08:47,706,704	GM08:47,770,836	5.40	4.80	0.85	N/A

N/A, not available.

Table 2.3 Overlapping quantitative trait loci regions between the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Region	Trait	QTL name	Linkage group	Position of the QTL peak	Confidence interval (cM)		QTL position		LOD	PVE (%)	Additive effect
				(cM)	Low	High	Left	Right			
GM04:16,974,874-17,152,230	Pod-filling	*ICIM_44_fill2*	4	226.00	225.50	226.50	GM04:16,974,874	GM04:17,152,230	13.21	27.41	1.81
GM07:5,256,305-5,4049,71	Maturity	*ICIM_44_mat5*	7a	17.00	15.50	18.50	GM07:5,256,305	GM07:5,279,354	11.35	11.28	1.30
	Maturity	*GCIM_44_mat5*	7a	22.52	N/A	N/A	GM07:5,404,971	GM07:5,404,971	8.52	3.60	1.15
GM16:5,680,173-5,730,237	Maturity	*GCIM_44_mat2*	16	53.40	N/A	N/A	GM16:5,680,173	GM16:5,730,237	13.07	6.16	-1.51
	Maturity	*ICIM_44_mat3*	16	53.00	52.50	54.50	GM16:5,680,173	GM16:5,730,237	14.67	14.89	-1.55
GM16:22,756,017-23,154,638	Pod-filling	*ICIM_44_fill1*	16	159.00	157.50	161.50	GM16:22,756,017	GM16:23,154,638	5.27	9.65	-1.12
	Pod-filling	*GCIM_44_fill1*	16	159.68	N/A	N/A	GM16:23,154,638	GM16:23,154,638	7.47	13.35	-1.35

N/A, not available.

Table 2.4 Unique quantitative trait loci regions identified in the QS15544$_{RIL}$ population.

Region	Locus, gene	Nucleotide variant		Amino acid variant		Type	Consequence	Gene name	Provider early/ shorter phenotype
		Position	W82/MA/ OV/ MD/90[1]	Position	W82/MA/ OV/ MD/90[1]				
MergGM04f (GM04:35,168,111-37,664,017)	*Glyma.04G156200*	GM04:36,554,187	A/A/G/A/G	1,794	L/L/P/L/P	Missense	Tolerated	*Tetratricopeptide repeat domain protein/Reduced chloroplast coverage 2*	OV & 9004
	Glyma.04G156400	GM04:36,758,687	G/G/A/G/A	N/A	N/A	3'UTR	N/A	*E1-like-a*	
	Glyma.04G156700	GM04:36,985,491	T/T/A/T/*	270	L/L/Q/L/*	Missense	Tolerated	*Strictosidine synthase-related*	
	Glyma.04G157000	GM04:37,049,412	C/C/A/C/A	108	W/W/L/W/L	Missense	Deleterious	*Enolase 1, Chloroplastic*	
	Glyma.04G157100	GM04:37,068,001	T/T/G/T/*	352	N/N/J/N/*	Missense	Tolerated	*Pollen-expressed transcription factor 2*	
	Glyma.04G157100	GM04:37,069,037	C/C/T/C/T	26	C/C/Y/C/Y	Missense	Deleterious	*Pollen-expressed transcription factor 2*	
MergGM04gh (GM04:41,808,599-42,376,237)	*Glyma.04G166300*	GM04:41,757,388	G/G/T/G/*	N/A	N/A	5'UTR	N/A	*Pseudo-response regulator 1a*	OV & 9004
	Glyma.04G167900	GM04:42,126,107	T/T/A/T/A	N/A	N/A	3'UTR	N/A	*Light-harvesting chlorophyll-protein complex I subunit A4*	
	Glyma.04G167900	GM04:42,126,847	A/A/G/A/G	N/A	N/A	3'UTR	N/A	*Light-harvesting chlorophyll-protein complex I subunit A4*	
	Glyma.04G167900	GM04:42,126,965	G/G/A/G/A	N/A	N/A	3'UTR	N/A	*Light-harvesting chlorophyll-protein complex I subunit A4*	
	Glyma.04G167900	GM04:42,127,008	G/G/T/G/T	N/A	N/A	3'UTR	N/A	*Light-harvesting chlorophyll-protein complex I subunit A4*	
	Glyma.04G168300	GM04:42,192,025	C/C/A/C/A	306	Q/Q/H/Q/H	Missense	Tolerated	*Cycling dof factor 3*	
	Glyma.04G169200	GM04:42,358,749	G/G/C/G/C	238	R/R/G/R/G	Missense	Deleterious	*EMB514*	
	Glyma.04G169200	GM04:42,359,864	A/A/G/A/G	115	L/L/S/L/S	Missense	Tolerated	*EMB514*	
MergGM08f (GM08:47,258,336-47,770,836)	*Glyma.08G362400*	GM08:47,378,990	C/C/T/T/C	867	G/G/E/E/G	Missense	Deleterious	*Clathrin interactor 1*	MA & 9004
	Glyma.08G366200	GM08:47,712,475	C/C/T/T/C	237	D/D/N/N/D	Missense	Deleterious	*U11/U12 small nuclear ribonucleoprotein 35 Kda protein*	
	Glyma.08G366400	GM08:47,724,957	G/G/A/A/G	115	G/G/S/S/G	Missense	Tolerated	*F1O19.11 protein*	
	Glyma.08G366400	GM08:47,725,261	A/A/G/G/A	216	H/H/R/R/H	Missense	Tolerated	*F1O19.11 protein*	
	Glyma.08G366600	GM08:47,736,890	A/A/G/G/A	655	D/D/G/G/D	Missense	Tolerated	*Phosphatidylinositol N-acetylglucosaminyltransferase subunit P down syndrome critical region protein 5-related*	
	Glyma.08G367000	GM08:47,756,614	C/C/A/A/C	53	A/A/S/S/A	Missense	Tolerated	*UDP-glycosyltransferase 72B2-related*	

[1] W82, William 82; MA, Maple Arrow; OV, OAC Vision; MD, AAC Mandor; 90, 9004. An asterisk (*) indicates a heterozygote for the SNP of interest.
N/A, not available.

Table 2.5 Candidate variants for the overlapping quantitative trait loci regions.

Region	Locus, gene	Nucleotide variant		Amino acid variant		Type	Consequence	Gene name	Provider early/ shorter phenotype
		Position	W82/MA/OV/ MD/90[1]	Position	W82/MA/OV/ MD/90[1]				
GM04:16,974,874-17,152,230	*Glyma.04G124300*	GM04:16,097,210	G/G/G/T/G	375	C/C/C/F/C	Missense	Tolerated	*Protein far-red elongated hypocotyl 3*	9004
	Glyma.04G124600	GM04:16,331,703	C/C/*/*/T	350	C/C/*/*/Y	Missense	Deleterious	*Far1-related sequence 5*	
	Glyma.04G124800	GM04:16,374,159	T/T/T/G/T	188	N/A	Stop-gain	N/A	*Zinc inducted facilitator-like 1*	
GM07:5,256,305-5,404,971	*Glyma.07G049400*	GM07:4,202,517	T/T/T/T/C	N/A	N/A	3'UTR	N/A	*Pseudo-response regulator 5d*	9004
	Glyma.07G058200	GM07:5,200,811	C/C/C/C/*	731	R/R/R/R/*	Missense	Tolerated	*Protein suppressor of PHYA-105*	
GM16:5,680,173-5,730,237	*Glyma.16G044100*	GM16:4,136,378	C/A/A/A/C	N/A	N/A	3'UTR	N/A	*Flowering locus T*	MD
	Glyma.16G050300	GM16:4,820,456	G/A/*/A/G	122	T/I/*/I/T	Missense	Tolerated	*Fusca3*	
	Glyma.16G057200	GM16:5,596,915	T/A/A/A/T	301	N/I/I/I/N	Missense	Tolerated	*Baf60/Chc1*	
	Glyma.16G057200	GM16:5,597,545	G/A/A/A/G	117	R/C/C/C/R	Missense	Tolerated	*Baf60/Chc1*	
GM16:22,756,017-23,154,638	*Glyma.16G109600*	GM16:23,645,426	G/G/G/G/C	24	E/E/E/E/D	Missense	Tolerated	*RNA-binding glycine-rich protein D4*	MD
	Glyma.16G109600	GM16:23,645,448	G/G/G/G/A	32	A/A/A/A/T	Missense	Deleterious	*RNA-binding glycine-rich protein D4*	
	Glyma.16G110400	GM16:24,358,614	A/A/A/A/*	178	K/K/K/K/*	Missense	Tolerated	*Apyrase 7*	
	Glyma.16G110400	GM16:24,358,915	T/T/T/T/G	278	H/H/H/H/Q	Missense	Deleterious	*Apyrase 7*	
	Glyma.16G110700	GM16:24,403,586	T/T/T/T/C	N/A	N/A	Splice donor variant	N/A	*Cytochrome P450*	

[1] W82, William 82; MA, Maple Arrow; OV, OAC Vision; MD, AAC Mandor; 90, 9004. An asterisk (*) indicates a heterozygote for the SNP of interest.
N/A, not available.

Table 2.6 Candidate variants for the unique quantitative trait loci regions identified in the QS15544$_{RIL}$ population.

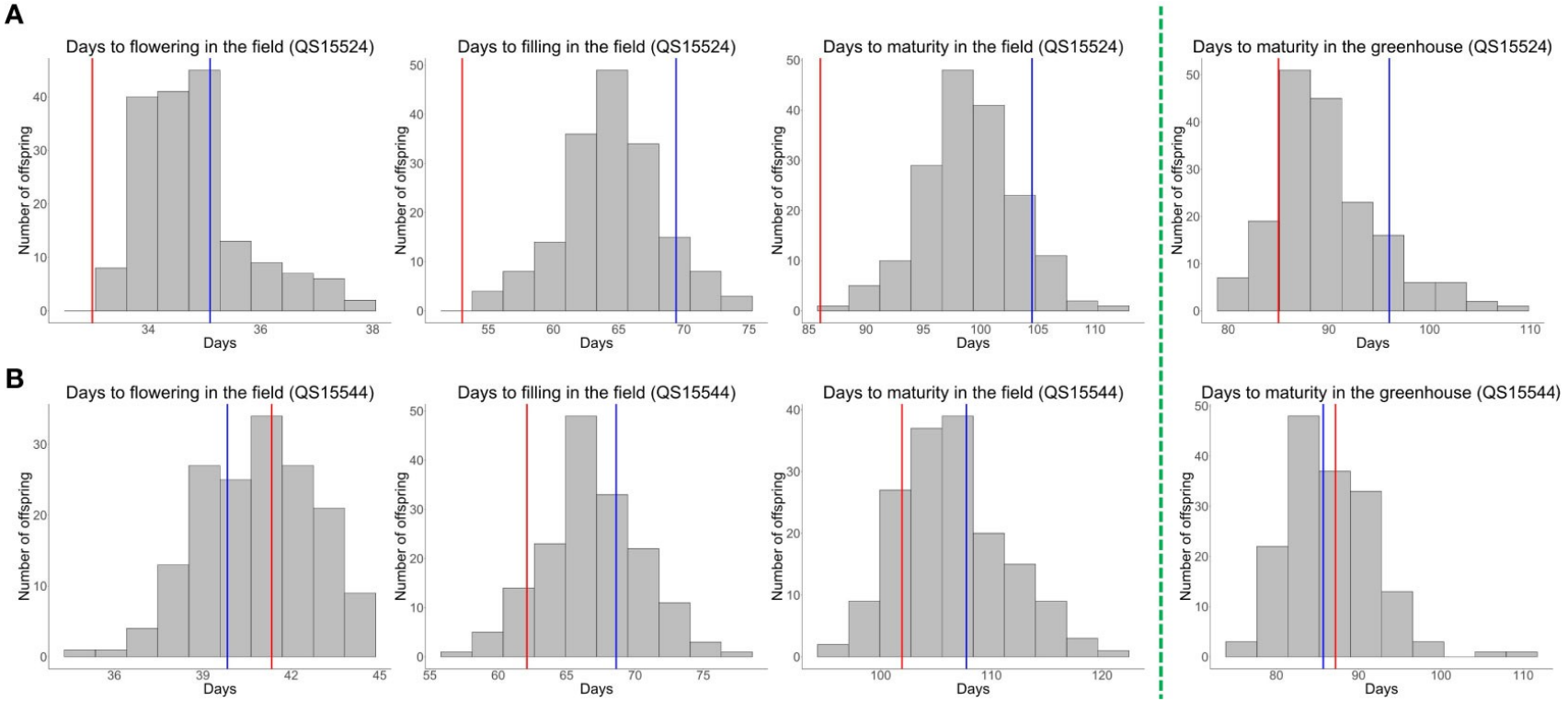

Figure 2.1 Phenotypic trait data distribution for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations. (A) Distribution of the phenotypes for the QS15524$_{F2:F3}$ population in the greenhouse (winter 2017–2018) and in the field (phenotypic average for the summers of 2018 and 2021). Parental lines are indicated with vertical-colored lines. Red lines, "OAC Vision"; blue lines, "Maple Arrow". **(B)** Distribution of the phenotypes for the QS15544$_{RIL}$ population in the greenhouse (winter 2019-2020) and in the field (phenotypic average for the summers of 2020, 2021, and 2022). Parental lines are indicated with vertical-colored lines. Red lines, "9004"; blue lines, "AAC Mandor". The green dotted line delineates the field (left-hand side) and the greenhouse (right-hand side) phenotypes.

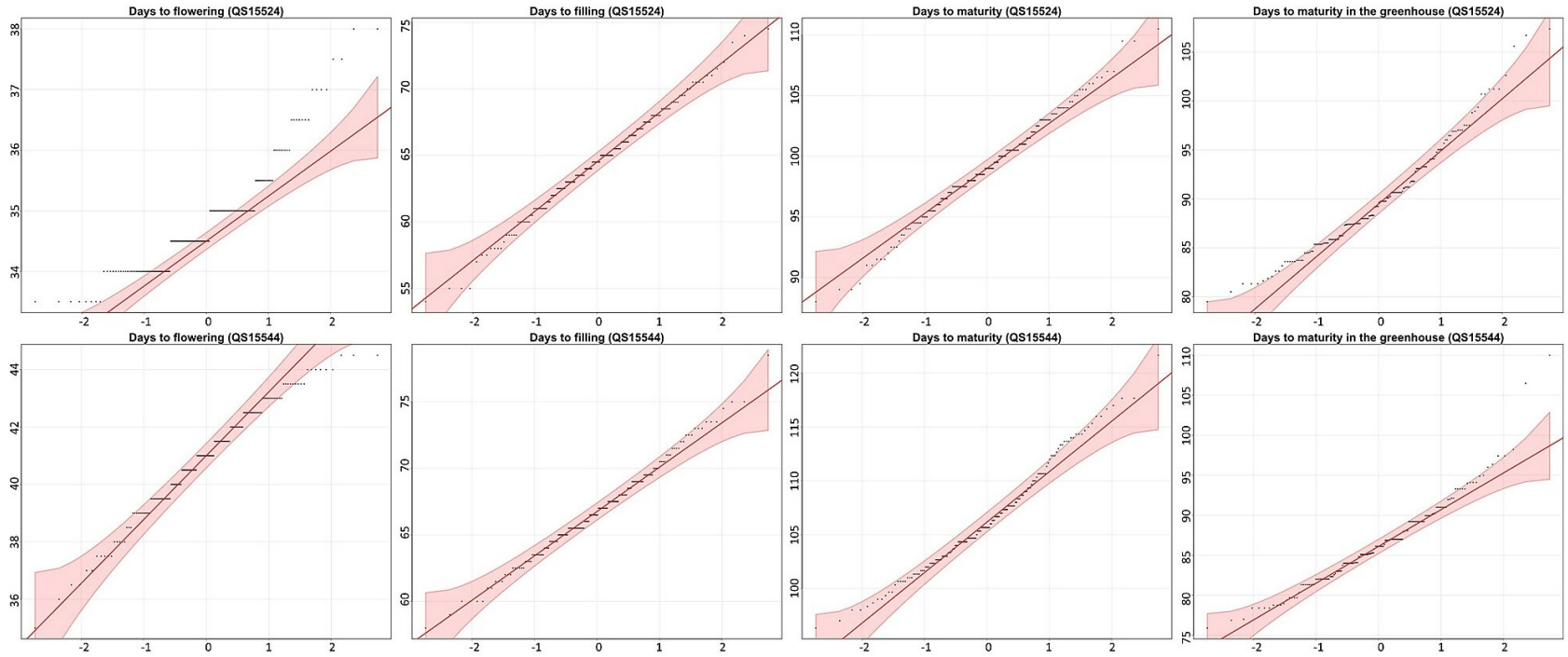

Figure 2.2 Quantile-quantile (Q-Q) plots of phenotypic traits.

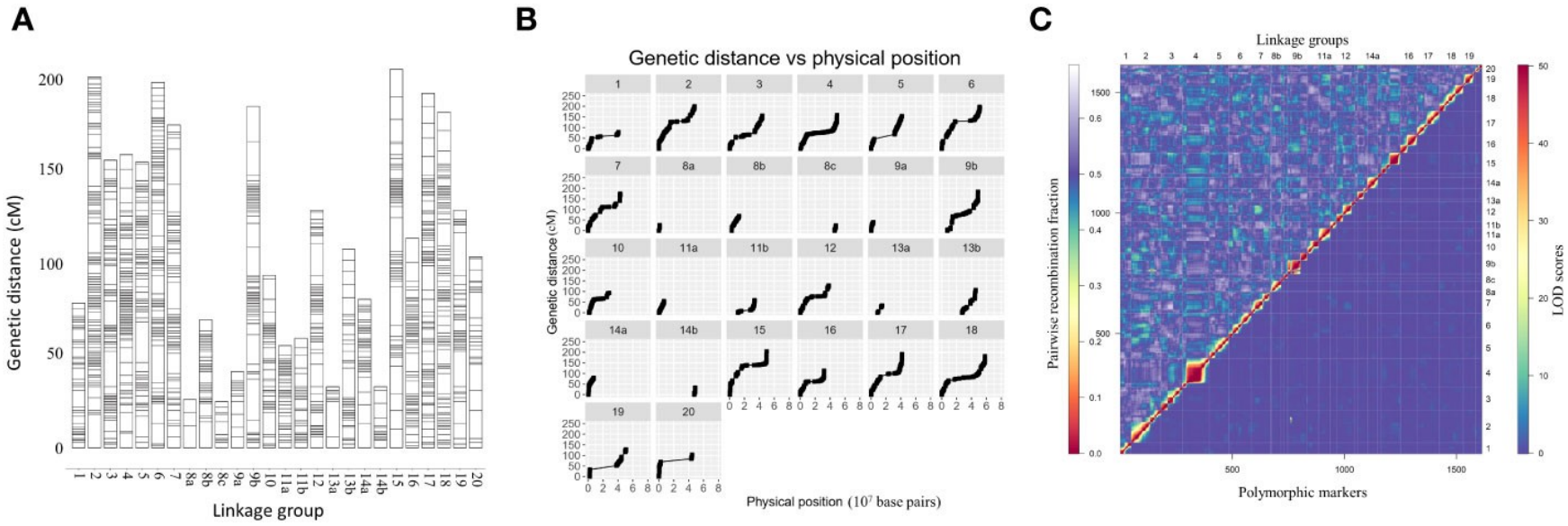

Figure 2.3 Construction of the linkage map for the QS15524$_{F2:F3}$ population. (A) Full linkage map displaying the 26 linkage groups and 1,613 polymorphic markers. **(B)** Plot of the genetic distance vs. the physical position of the markers. **(C)** Pairwise recombination fraction (upper left) and LOD scores for tests of linkage (bottom right) for all 1,613 markers. The upper half represents the recombination fraction between the markers, from the lowest (red color) to the highest (white color). The bottom half displays the LOD score associated with the linkage between each marker pair, from the lowest (blue color) to the highest (red color). Smaller linkage groups have been removed to facilitate visualization.

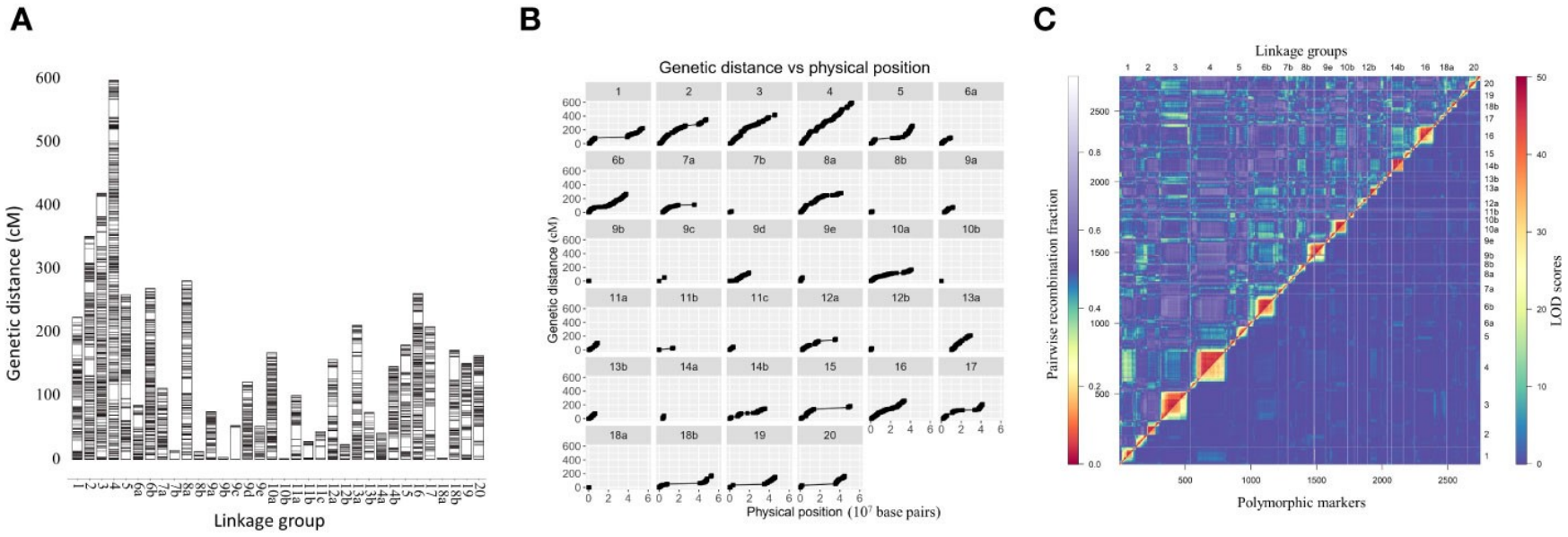

Figure 2.4 Construction of the linkage map for the QS15544$_{RIL}$ population. (A) Full linkage map displaying the 34 linkage groups and 2,746 polymorphic markers. **(B)** Plot of the genetic distance vs. the physical position of the markers. **(C)** Pairwise recombination fraction (upper left) and LOD scores for tests of linkage (bottom right) for all 2,746 markers. The upper half represents the recombination fraction between the markers, from the lowest (red color) to the highest (white color). The bottom half displays the LOD score associated with the linkage between each marker pair, from the lowest (blue color) to the highest (red color). Smaller linkage groups have been removed to facilitate visualization.

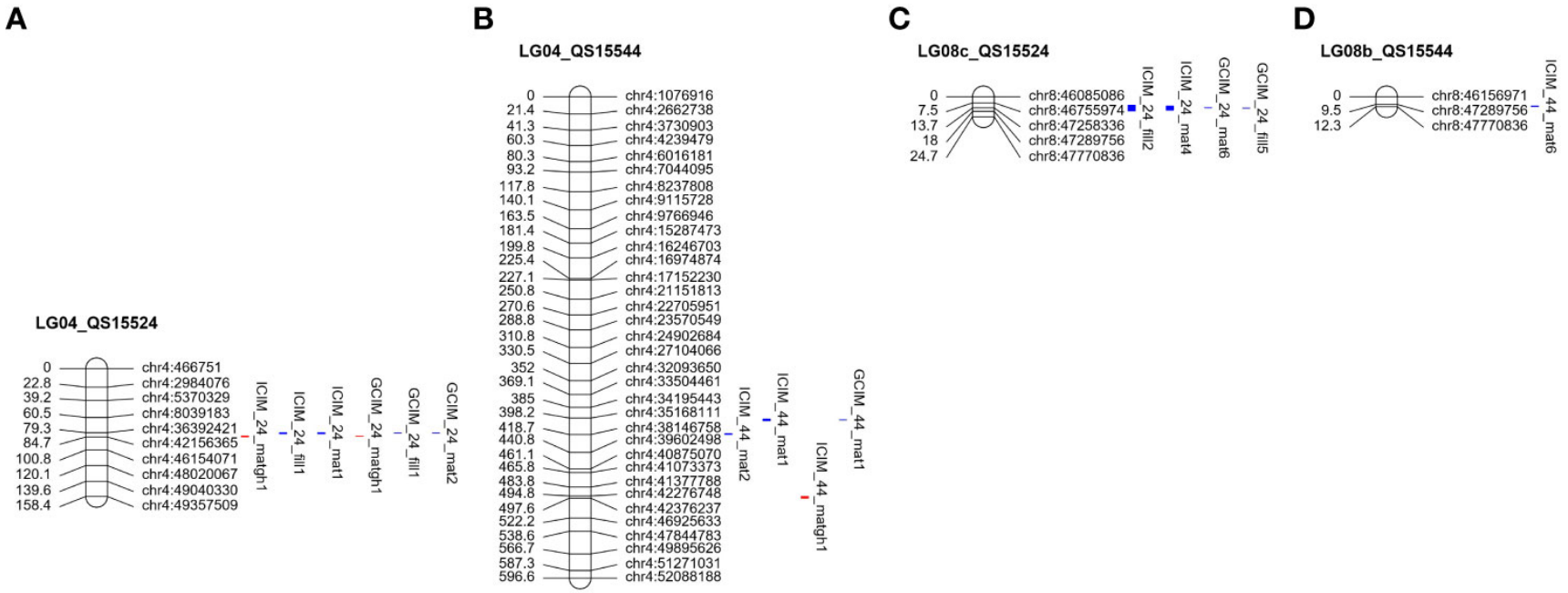

Figure 2.5 Overlapping quantitative trait loci signals between the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations. Red-marked traits indicate the number of days to maturity in the greenhouse, whereas blue-marked traits are field phenotypes. The QTL regions identified for the QS15524$_{F2:F3}$ **(A)** and QS15544$_{RIL}$ **(B)** populations on chromosome GM04. Two overlapping regions were identified on this chromosome, *MergGM04f* (GM04:35,168,111-37,664,017) and *MergGM04gh* (GM04:41,808,599-42,376,237). A third overlapping region, *MergGM08f* (GM08:47,258,336-47,770,836) was found on chromosome GM08. The identified QTL in this genetic region included populations QS15524$_{F2:F3}$ **(C)** and QS15544$_{RIL}$ **(D)**. The number of markers has been decreased for both chromosomes to facilitate visualization.

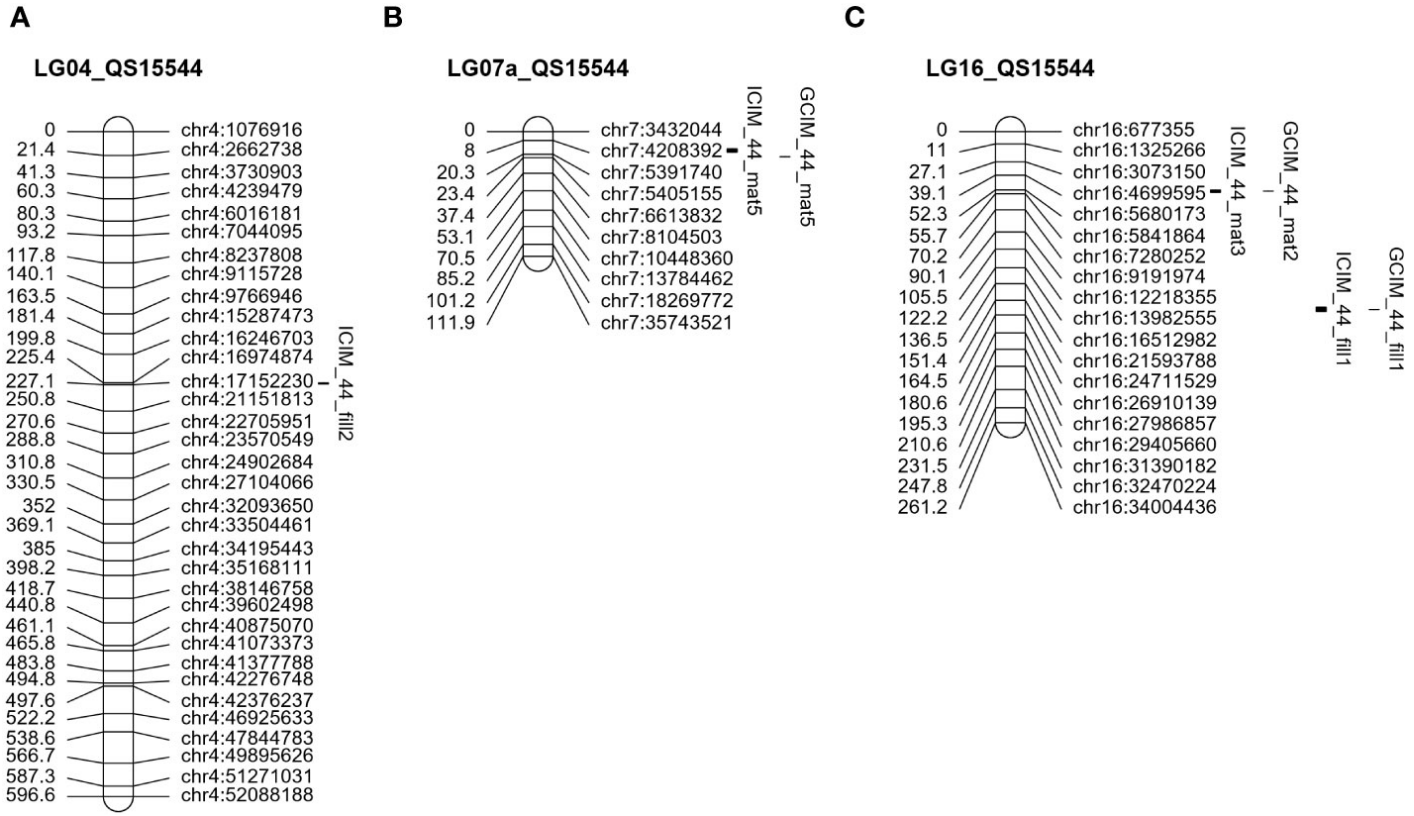

Figure 2.6 Unique QTL regions identified in the QS15544$_{RIL}$ population. Significant QTL identified on LG04 **(A)**, LG07a **(B)**, and LG16 **(C)**. The number of markers has been decreased on all chromosomes to facilitate visualization.

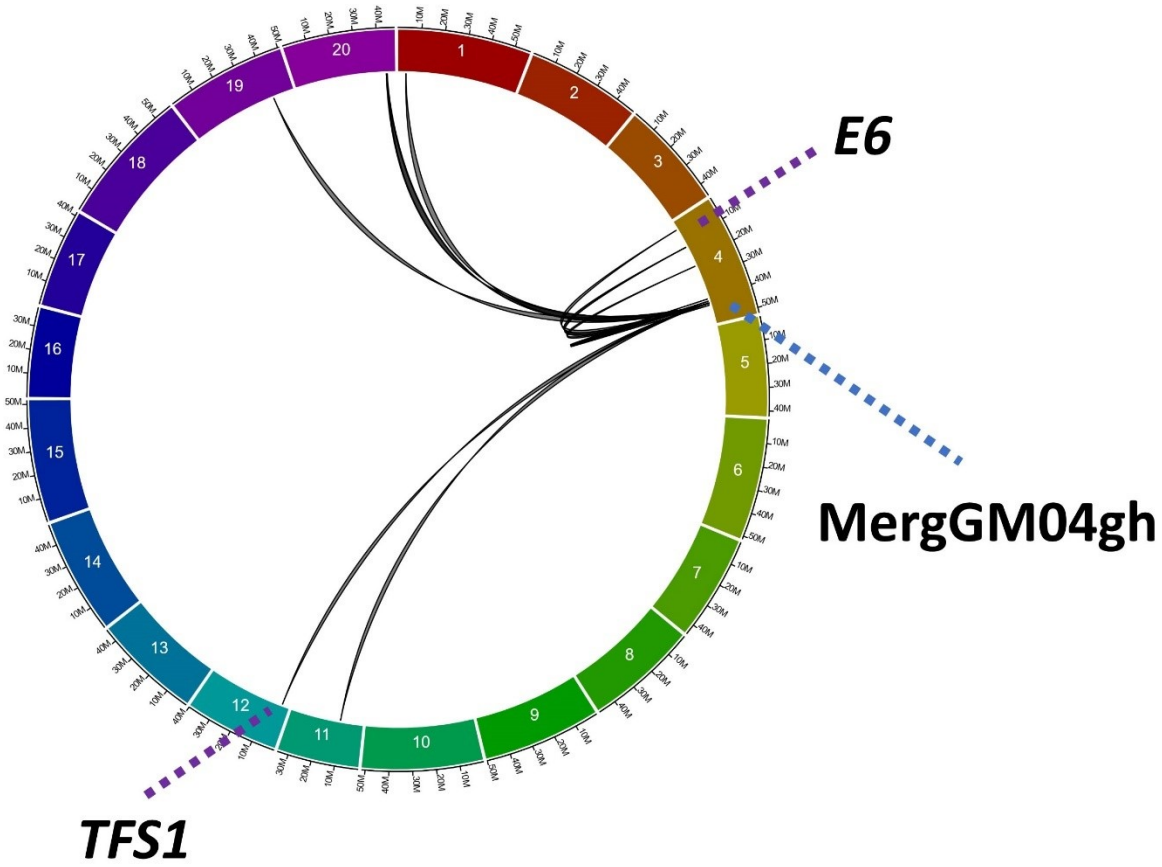

Figure 2.7 *Trans* and *cis* expression quantitative trait loci signals for the *MergGM04gh* region. Interactions between this region and 13 different e-traits have been identified using a combination of three algorithms (IM, ICIM, and GCIM). Black lines underline the eQTL interactions between the *MergGM04gh* region and its target genes. Purple dotted lines indicate the positions of two genes involved in flowering: *Glyma.04G050200* (*GmELF3/E6* locus) and *Glyma.12G048500* (*GmTFS1*). Blue dotted line indicates the location of the *MergGM04gh* region.

3.10 Connecting text

Chapter 3 demonstrated that multiple critical loci regulate shorter pod-filling and earlier maturity in two soybean populations adapted for MGs 00 and 000 cultivation areas. In addition, this chapter showed that a specific region, *E8-r3*, regulates the expression level of *E6*, a critical locus characterized with the long-juvenile trait, in *trans*. In the literature, many of the proposed candidates (e.g., *GmCDF3* for *E8-r3* and *E1la* for *E8-r2*) have been suggested to exhibit transcription factor activity. As such, Chapter 4 further investigates the gene regulatory networks governing early reproductive traits in both populations using a novel expression quantitative trait loci mapping pipeline and an innovative *trans* hotspot detection strategy. Overall, the following study identifies the hotspots regulating transiently the level of expression of several candidate genes for the *E8-r3* locus at the V4 stage in soybean, just before the initiation of floral meristematic transition.

4. Integrated eQTL Mapping Approach Reveals Genomic Regions Regulating Candidate Genes of the *E8-r3* Locus in Soybean

Jérôme Gélinas Bélanger[1,2], Tanya Rose Copley[1], Valerio Hoyos-Villegas[2] & Louise O'Donoughue[1]

[1]CÉROM, Centre de recherche sur les grains Inc., St-Mathieu-de-Beloeil, Québec, Canada

[2]Department of Plant Science, McGill University, Montréal, Québec, Canada

* Correspondence:

Dr. Louise O'Donoughue

louise.odonoughue@cerom.qc.ca

Reproduced from Frontiers in Plant Science.

Gélinas Bélanger J, Copley TR, Hoyos-Villegas V and O'Donoughue L (2024) Integrated eQTL mapping approach reveals genomic regions regulating candidate genes of the *E8-r3* locus in soybean. ***Front. Plant Sci.*** 15:1463300. doi: 10.3389/fpls.2024.1463300

Minor modifications were made to conform to the McGill University thesis guidelines.

4.1 Abstract

Deciphering the gene regulatory networks of critical quantitative trait loci associated with early maturity provides information for breeders to unlock soybean's (*Glycine max* (L.) Merr.) northern potential and expand its cultivation range. The *E8-r3* locus is a genomic region regulating the number of days to maturity under constant short-day photoperiodic conditions in two early-maturing soybean populations ($QS15524_{F2:F3}$ and $QS15544_{RIL}$) belonging to maturity groups MG00 and MG000. In this study, we developed a combinatorial expression quantitative trait loci mapping approach using three algorithms (ICIM, IM, and GCIM) to identify the regions that regulate three candidate genes of the *E8-r3* locus (*Glyma.04G167900*/*GmLHCA4a*, *Glyma.04G166300*/*GmPRR1a*, and *Glyma.04G159300*/*GmMDE04*). Using this approach, a total of 2,218 *trans* (2,061 genes)/7 *cis* (7 genes) and 4,073 *trans* (2,842 genes)/3,083 *cis* (2,418 genes) interactions were mapped in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, respectively. From these interactions, we successfully identified two hotspots (F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386) and three minor regions (RIL_GM04:17,227,512-20,251,662, RIL_GM04:31,408,946-31,525,671 and RIL_GM13:37,289,785-38,620,690) regulating the candidate genes of *E8-r3* and several of their homologs. Based on co-expression network and single nucleotide variant analyses, we identified *ALTERED PHLOEM DEVELOPMENT* (*Glyma.15G263700*) and *DOMAIN-CONTAINING PROTEIN 21* (*Glyma.18G025600*) as the best candidates for the F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386 hotspots. These findings demonstrate that a few key regions are involved in the regulation of the E8-r3 candidates *GmLHCA4a*, *GmPRR1a*, and *GmMDE04*.

4.2 Introduction

Soybean (*Glycine max* (L.) Merr.) is the most important leguminous oilseed crop and significantly contributes to maintaining food security on a global scale. This crop is mainly cultivated in countries located in warm temperate, subtropical, and/or tropical areas, and the contribution of northern countries such as Canada (2%) to the global soybean output remains modest (The American Soybean Association, 2023). Current projections suggest limited growth in soybean production for countries located in tropical and subtropical countries (Ali *et al.*, 2022); however, due to the projected rise in world population and anticipated growth in the international need for soybean-related food and industrial goods, global production will need to increase to

supply the growing demand (Unc *et al.*, 2021). One approach to partly solve this problem is to improve soybean adaptability to northern regions beyond its actual limits (~54°N) and fine-tune its reproductive phenology by identifying the critical transcription factors regulating the extra-early flowering and maturity phenotypes.

Transcription factors (TFs) are critical proteins that regulate the transcription of one or multiple downstream targets by binding to *cis*-regulatory elements (CRE) with their DNA-binding domains (DBD) (Bylino *et al.*, 2020). In soybean, 6,150 TFs (3,747 loci) belonging to 57 families have been predicted (Jin *et al.*, 2017), with several having reported flowering regulatory functions such as *E1* (Xia *et al.*, 2012), *E1-like-a* (Liu *et al.*, 2022), *E1-like-b* (Zhu *et al.*, 2019), and *LHY1a/1b/2a/2b* (*Glyma.16G017400*, *Glyma.07G048500*, *Glyma.19G260900*, and *Glyma.03G261800*, respectively) (Bian *et al.*, 2017). In *Arabidopsis*, core regulators of the circadian clock are all TFs and include *CIRCADIAN CLOCK-ASSOCIATED1* (*AtCCA1*), *LATE ELONGATED HYPOCOTYL* (*AtLHY*), and the evening-expressed gene, *TIMING OF CAB EXPRESSION1* (*AtTOC1*) (Wang and Ma, 2013). As the main biological timekeeper, the circadian clock gates the global molecular response to the environmental cues, the zeitgebers, in a timely fashion. From an agronomical standpoint, these multiple interlocked transcription-translation feedback loops comprised within the circadian clock regulate essential metabolic functions (e.g. photosynthesis and reproductive phenology) with potential effects on critical traits such as maturity, yield, and disease resistance (Hotta, 2021). In particular, the cryptochrome and phytochrome photoreceptors regulate many key aspects of the circadian clock and act as a molecular bridge between photosynthesis, development, and reproduction (Venkat and Muneer, 2022). As a consequence, photosynthesis and reproduction are intertwined at the molecular level due to specific genes (e.g. *PHYTOCHROME A2/A3*) acting to control photoperiodic flowering (Lin *et al.*, 2022).

Generating a compendium of interactions for one specific TF is challenging due to the transient nature of the regulatory mechanisms and the intricate density of the underlying regulatory networks. One approach to solve this issue is to perform expression quantitative trait loci (eQTL) mapping on a genome-wide scale to identify proximal/*cis* (within a 1-Mbp window of the transcription start site) and distal/*trans* single nucleotide polymorphisms influencing the level of messenger RNA (mRNA) expression (Gilad *et al.*, 2008; Westra and Franke, 2014). Expression quantitative trait loci hotspots are genetic variations, most often located in genes coding for TFs,

that regulate the expression level of numerous genes, often in the hundreds to thousands (Choi *et al.*, 2020). Obtaining sufficient statistical power is often challenging in eQTL mapping due to the prohibitive financial cost associated with sufficient reading depth and the computational burden of transcriptome-wide measurements on hundreds of lines. To overcome this challenge, numerous mapping algorithms, such as Genome-wide composite interval mapping (GCIM) (Zhang *et al.*, 2020) and Inclusive Composite Interval mapping (Li *et al.*, 2007), have been developed to improve the identification of small-effect eQTLs, which are most often located in *trans* (Westra and Franke, 2014). We believe that used in conjunction, these methods have an increased ability to identify regions of interest for given phenotypes and can also be used to map eQTL interactions and associated regulatory hotspots with increased precision.

In a previous study, we identified a QTL region named *E8-r3* located between the GM04:41,808,599 and GM04:42,376,237 flanking markers that regulates the number of days to maturity under a constant short-day photoperiod in two early-maturing soybean populations ($QS15524_{F2:F3}$ and $QS15544_{RIL}$) (Gélinas Bélanger *et al.*, 2024a). The same region was not identified when these populations were grown under fluctuating long-day conditions under Canadian field conditions, suggesting that this region is specifically involved in photoperiodic responses under short days. In this previous study, we also identified that this region regulates the expression of several genes, including *E6* (*Glyma.04G050200*), an ortholog of *Arabidopsis thaliana EARLY FLOWERING 3* that has been demonstrated to have an effect on the flowering of soybean using a combinatorial eQTL mapping approach (Fang *et al.*, 2021; Gélinas Bélanger *et al.*, 2024a). The associated QTL region identified for the short-day phenotypic response encompasses 29 genes and is implicated in the 'Photosynthesis - antenna proteins' KEGG pathway. In total, we have proposed three candidate genes (*Glyma.04G168300*, *Glyma.04G167900*, and *Glyma.04G166300*) for this region based on a candidate single nucleotide polymorphisms (SNPs) analysis. Two of these genes, *Glyma.04G166300* (*PSEUDO–RESPONSE REGULATOR 1a*; *GmPRR1a*) (Liu *et al.*, 2009; Dietz *et al.*, 2022) and *Glyma.04G168300* (*CYCLING DOF FACTOR 3*; *GmCDF3*) (Corrales *et al.*, 2017), encode TFs involved in the circadian clock, developmental processes and regulation of maturity, thus suggesting that TFs might be involved in the regulation of the *E8-r3* locus. The other gene, *Glyma.04G167900* (*LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT A4a; LHCA4a*), is involved in photosynthetic activities and possibly regulated by TFs located in *cis* or in *trans*. Recently, a gene coding for a

MADS-box transcription factor, *Glyma.04G159300* (*MADS-BOX DOWNREGULATED BY E1 04*; *GmMDE04*), was found to be statistically associated with the GM04:39,294,836 marker for the flowering time (i.e. R1 stage), maturity (i.e. R8 stage), and reproductive length (i.e. the difference between R8 and R1) traits (Escamilla *et al.*, 2024). Although this gene is located outside of *E8-r3* flanking markers, our lab is currently reconsidering its potential role as a regulator for this locus based on the results found by Escamilla *et al.* (2024). The objective of the present study is to identify novel eQTLs using an approach combining multiple mapping techniques in two early-maturing soybean populations. Overall, this study aims at (i) validating an eQTL mapping pipeline based on a combinatorial mapping strategy; (ii) identifying eQTL signals and hotspots regulating the genes involved in flowering, maturity, and photosynthesis; (iii) locating the eQTL signals interacting with the *E8-r3* region; and (iv) identifying candidate TFs and characterizing their co-expression networks.

4.3 Materials and Methods

4.3.1 Plant Materials, Growing Conditions, and Phenotyping

The populations and phenotyping procedures were generated and performed as detailed in Gélinas Bélanger *et al.* (2024a). Briefly, the QS15524$_{F2:F3}$ population was generated from a biparental cross between 'Maple Arrow' (MG00; later-maturing accession) × 'OAC Vision' (PI 567787) (MG000; earlier-maturing accession), now herein respectively referred as MA and OV. The QS15544$_{RIL}$ population was generated from the biparental cross between 'AAC Mandor' (MG00; later-maturing accession) × '9004' (MG000; earlier-maturing accession), the former now being herein referred to as MD. The parental lines in each population were fixed for their *E1* (*Glyma.06G207800*) (Xia *et al.*, 2012), *E2* (*Glyma.10G221500*) (Watanabe *et al.*, 2011), *E3* (*Glyma.19G224200*) (Watanabe *et al.*, 2009), and *E4* (*Glyma.20G090000*) (Liu *et al.*, 2008) alleles. As such, the genotypes for the QS15524$_{F2:F3}$ parental lines were *e1-nl*/*e2-ns*/*E3Ha*/*e4-SORE-1* and *e1-as*/*e2-ns*/*e3-tr*/*e4p.T832QfsX21* for the QS15544$_{RIL}$ parental lines. The *e4p.T832QfsX21* allele is a rare premature stop codon mutation previously identified in Tardivel *et al.* (2019).

The QS15524$_{F2:F3}$ population was grown and phenotyped in a greenhouse during the winter of 2017-2018 at the Centre de recherche sur les grains inc. (CÉROM) in St-Mathieu-de-Beloeil (QC, Canada), GPS coordinates 45°34'57.9"N 73°14'11.4"W. In the case of QS15544$_{RIL}$, the

population (F_5:F_6 generation) was grown and phenotyped in a greenhouse during the winter of 2019-2020. Plants for the offspring and parental lines were sown on December 14th 2017 and October 25th 2019 for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, respectively. During the experiments, natural photoperiod was below 12h but maintained artificially at 12h using sodium halogen lights at all time before flowering since flowering for all plants happened before the March Equinox. Both populations were grown following a custom Modified Augmented Design (Lin & Poushinsky, 1983, 1985) with 19 individuals per table and one parent per table. For each population, the plants were sown in one-gallon pots containing a ProMix-garden soil (1:1 v:v) (Premier Tech Horticulture, Rivière-du-Loup, QC, Canada) potting mix. For the $QS15524_{F2:F3}$ offspring, one seed was planted per pot, whereas three seeds were sown per pot for the $QS15544_{RIL}$ offspring. As reported by Gélinas Bélanger *et al.* (2024a), the OV and MA parents of the $QS15524_{F2:F3}$ population respectively matured in 85 and 96 days. For the $QS15544_{RIL}$ population, it was observed that the MD and '9004' lines matured in 87.5 and 87.2 days, respectively.

4.3.2 Sampling, Nucleic Acid Extraction and Sequencing

The sampling and sequencing procedures were performed as detailed in Gélinas Bélanger *et al.* (2024a). Briefly, leaf tissue for RNA extraction was harvested by making six 4 mm plugs in the uppermost expanding middle leaflet of the trifoliate leaf 4 hours after sunrise at the V4 leaf stage (25 days post-seeding), frozen immediately in liquid nitrogen and stored at -80°C until further use (Fig. 3.1A). The time points were chosen based on previously published data indicating highest expression of flowering genes four hours after sunrise (Kong *et al.*, 2010; Sun *et al.*, 2011). Furthermore, the V4 stage was determined based on preliminary qRT-PCR analyses of the expression of the flowering genes *Glyma.16G150700* (*FLOWERING LOCUS T 2A*; *GmFT2a*) and *Glyma.16G044100* (*FLOWERING LOCUS T 5A*; *GmFT5a*) in the parental lines (data not shown). To do so, we compared the expression for the V1 to V5 stages and chose the stage which exhibited the highest expression for both of these genes as the *FT* florigens promote the transition to reproductive development and flowering. The extraction and purification of total DNA from leaf tissue was performed using the Omega Bio-Tek Mag-bind Plant Kit and Mag-Bind Total Pure NGS kit (Omega Biotek, Georgia state, USA). Construction of the whole genome sequencing (WGS) libraries for the $QS15524_{F2:F3}$ parental lines was performed by pooling the leaf tissue from the five

pots of each parent. Extraction of total DNA and library preparation was performed at the Génome Québec Innovation Centre (Montréal, Canada) using the NxSeq® AmpFREE Library Preparation kit (Lucigen, Wisconsin, U.S.A.). The two parental libraries were combined and sequenced at a 15X depth on the Illumina HiSeq X platform with 150 bp paired-end reads.

The genotyping-by-sequencing (GBS) libraries of the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ mapping populations were prepared using the *PstI/MspI* enzymes as described in Abed *et al.* (2019) at the Institute of Integrative Biology and Systems (Laval University, Québec, Canada). Sequencing of the $QS15524_{F2:F3}$ GBS libraries was performed by randomly combining a total of 91 barcoded samples per library and by sequencing with four Ion PI V3 chips per library (Fig. 3.1B). For the $QS15544_{RIL}$ population, the samples were randomly divided into two sets of 91 samples and sequenced with two Ion PI V3 chips per library. Sequencing for all libraries was performed on the Ion Proton Sequencer and HiQ chemistry at the Institute of Integrative Biology and Systems (Laval University, Québec, Canada).

Total RNA was extracted from samples using a standard Trizol RNA extraction procedure with two extra ethanol rinses to improve purity. Messenger RNA (mRNA) was isolated using the NEBNext mRNA stranded library preparation kit (New England Biolabs, Ontario, Canada) at the Génome Québec Innovation Centre (Montréal, Canada). Two libraries containing 96 pooled samples were prepared per population. Each library was then sequenced on two Illumina NovaSeq6000 S2 ($QS15524_{F2:F3}$) or S4 ($QS15544_{RIL}$) lanes at the Génome Québec Innovation Centre (Montréal, Canada), with four sequencing lanes per population and a total of 8000 M and 9600 M paired-end reads per population, respectively (Fig. 3.1C).

4.3.3 Bioinformatics

The bioinformatic analyses were performed as detailed in Gélinas Bélanger *et al.* (2024a). Briefly, alignment of all the sequences was performed using version 2 of the *Glycine max* reference genome (Gmax_275_v2.0) (https://phytozome-next.jgi.doe.gov/) (Accessed 8 December 2017; https://data.jgi.doe.gov/). Processing of the WGS sequencing datasets of the $QS15524_{F2:F3}$ parental lines was performed using the fast-WGS pipeline with the default settings (Torkamaneh *et al.*, 2017). The processing of GBS datasets was performed using the fast-GBS pipeline (Torkamaneh *et al.*, 2017) (Fig. 3.1B). Variant calling was performed with Platypus version 0.8.1 (Rimmer *et al.*, 2014) with the following commands: –minReads=2, –

minMapQual=20 and –minBaseQual=20. Subsequently, a filtering step using vcftools version 0.1.16 (Danecek *et al.*, 2011) was performed with the following parameters: (i) remove non-biallelic sites; (ii) remove InDels; (iii) remove scaffolds; and (iv) filter alleles using the –maxmissing 0.2, –maf 0.3 and –mac 4 commands. Self-imputation was then performed on the missing data for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations using Beagle version 4.1.0 (Browning and Browning, 2007) with twelve iterations. Phasing was then performed with Convert2Map (https://bitbucket.org/jerlar73/convert-genotypes-to-mapping-files/src/master/) using the fast-WGS resequencing data for the $QS15524_{F2:F3}$ parental lines and the GmHapMap dataset for the QS15544 parental lines. A last round of filtering was performed in the $QS15544_{RIL}$ dataset by removing all SNPs with > 10% heterozygous calls before binning with QTL IciMapping (Meng *et al.*, 2015). For $QS15524_{F2:F3}$, the binning step was performed with Genotype Corrector.

Processing of the RNA datasets was performed using multiple publicly available software tools with an in-house script (Fig. 3.1D). Briefly, adapters were removed using Trimmomatic version 0.33 (Bolger *et al.*, 2014) with the following options: ILLUMINACLIPTruSeq3-SE.fa:2:30:15, LEADING:3 and TRAILING:3, SLIDINGWINDOW:3:20, and MINLEN:32. Filtered reads were then aligned to the soybean reference transcriptome using TopHat2 version 2.1.1 (Kim *et al.*, 2013). Aligned reads were then counted using HTSeq-count version 0.6.1 (Anders *et al.*, 2015) and were filtered to be considered expressed only if they met the following criteria: (i) min raw counts of at least two to be considered active in a given line; and (ii) transcription recorded in a minimum of 25% of the population. This filtering step resulted in gene sets comprising 38,692 and 40,218 genes for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, respectively.

4.3.4 Linkage Map Construction

The linkage maps were built as described in Gélinas Bélanger *et al.* (2024a) (Fig. 3.1B). Briefly, the maps were generated using QTL IciMapping version 4.2 (Meng *et al.*, 2015) with the Kosambi mapping function. For both maps, the markers were anchored to their physical positions when ordering and the resulting linkage groups (LGs) were split when gaps exceeded 30 cM. The robustness of both linkage maps was previously demonstrated in two previous studies that aimed at mapping reproductive (Gélinas Bélanger *et al.*, 2024a) and seed quality (Gélinas Bélanger *et al.*,

2024b) traits by plotting the (i) genetic distance versus the physical position and (ii) the pairwise recombination fraction and LOD score (Fig. 3.1B). In addition, the high-quality of the linkage maps was assessed by confirming the synteny between the physical and genetic positions of the markers (data not shown).

4.3.5 Measurement of Differential Gene Expression

Measurement of differential gene expression was performed in the $QS15524_{F2:F3}$ (OV vs MA; Appendix 2.1) and $QS15544_{RIL}$ (MD vs '9004'; Appendix 2.2) parental lines (Fig. 3.1E). Each analysis was performed using the early-maturing parent ($QS15524_{F2:F3,}$ OV; $QS15544_{RIL}$, MD) as the reference line. Due to low-quality data in the RNA-seq datasets in the $QS15524_{F2:F3}$ parental lines, two replicates were removed each from the OV and MA samples, thus resulting in a total of three replicates per parent. Similarly, one replicate was removed from the MD and '9004' samples in the $QS15544_{RIL}$ parental lines, resulting in a total of four replicates per parent. The expressed gene sets comprised 38,692 genes for the $QS15524_{F2:F3}$ parents and 40,218 genes for the $QS15544_{RIL}$ parents which were filtered using the aforementioned parameters. Differentially expressed gene analysis was performed in iDEP.96 (Ge *et al.*, 2018) using the DESeq2 function with a false discovery rate (FDR) adjusted p-value threshold fixed at 0.05 and a minimum fold change of 2.0. The normalization of the transcripts for the GO analysis and the eQTL mapping (see below) was performed using the DESeq2 R package (Love *et al.*, 2014). Volcano plots and heatmaps were respectively generated using the online version of VolcaNoseR (Goedhart and Luijsterburg, 2020) and iDEP.96.

4.3.6 Gene Ontology Enrichment

Gene ontology (GO) enrichment was performed on the parental downregulated and upregulated gene sets using the Soybase_GOtool (https://www.soybase.org/goslimgraphic_v2/dashboard.php) as detailed in Morales *et al.* (2013) (Fig. 3.1E). The Fisher test p-values obtained with the Soybase_GOtool were adjusted using the Bonferroni correction with a threshold for these corrected p-values fixed at 0.01. In Soybase, the obtained p-value is automatically multiplied by the number of scanned genes (e.g. p-value 0.003 X 4000 genes = Bonferroni corrected p-value of 12), leading to p-values that can be above 1. From this list of results, the GO terms associated with molecular functions and cellular components were

manually removed using https://biodbnet-abcc.ncifcrf.gov/db/db2db.php, and only the GO terms associated with biological processes were retained. Following this step, we manually curated and retained GO terms associated with the following biological functions from Soybase (Grant *et al.*, 2009): (i) flowering; (ii) reproduction; (iii) senescence; (iv) photosynthesis; and (v) development. This list of GO terms included a total of 162 annotations (Appendix 2.3) and is herein referred to as FRSPD_GO (Flowering / Reproduction / Senescence / Photosynthesis / Development). In this paper, this list was used to annotate the enriched FRSPD terms in the parental differentially expressed gene (DEG) datasets, mapped eQTL interactions, and genes found in co-expression networks (CEN).

4.3.7 Expression Quantitative Trait Loci Analysis

Transcriptome-wide eQTL analysis was performed on normalized transcript abundances for the 176 lines of the $QS15524_{F2:F3}$ population (38,692 genes) and the 162 lines of the $QS15544_{RIL}$ (40,218 genes) population (Fig. 3.1F). The mapping of eQTL was performed using a combinatorial approach which includes the use of three different algorithms: (i) Inclusive composite interval mapping (ICIM) approach implemented in QTL IciMapping version 4.2 (Meng *et al.*, 2015); (ii) Interval mapping (IM) from QTL IciMapping version 4.2 (Meng *et al.*, 2015); and (iii) Genome-wide compositive interval mapping (GCIM) method in the QTL.gCIMapping.GUI.v2.0.GUI package (Zhang *et al.*, 2020). The LOD thresholds for ICIM and IM were calculated in QTL IciMapping with 1000 permutations using an α of 0.05 and a walking step of 1 cM for genome-wide scanning. To limit the computational burden (i.e. at least 1,000 permutations for 38,692 and 40,218 genes), we performed permutations on 100 randomly sampled gene transcripts (i.e. 1,000 permutations X transcripts for 100 randomly selected genes = 100,000 permutations) as performed in West *et al.* (2007); Wang *et al.* (2010, 2014), and Huang *et al.* (2020). Subsequently, the global permutation threshold was calculated as the 95th percentile of the representative null distribution and equaled to (i) 4.01 for ICIM in $QS15544_{RIL}$; (ii) 3.99 for IM in $QS15544_{RIL}$; (iii) 4.13 for ICIM in $QS15524_{F2:F3}$; and (iv) 4.12 for IM in $QS15524_{F2:F3}$. For GCIM, the fixed model component was chosen for the $QS15544_{RIL}$ population and the fixed-restricted maximum likelihood (REML) component was chosen for the $QS15524_{F2:F3}$ population, both with a walking speed of 1 cM. In the QTL.gCIMapping.GUI.v2.0.GUI package, the likelihood function is only available for F_2 populations and was chosen based on prior testing. For GCIM, the default

LOD threshold suggested in the literature is 2.5 for QTL studies; however, the thresholds were increased to 7.5 for the $QS15524_{F2:F3}$ and 4.0 for the $QS15544_{RIL}$ populations to reduce the noise and remove minor eQTL interactions. The contrasting LOD thresholds for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations were chosen based on preliminary tests performed using the different functions implemented in the QTL.gCIMapping.GUI.v2.0.GUI package. Following the mapping of interactions with the three algorithms, all of the significant interactions were classified either as *cis*-acting or *trans*-acting. Interactions were classified as *cis*-acting if within 1,000,000 bp region from the transcription start site (TSS) of the studied gene, whereas interactions were considered *trans*-acting if identified outside this 1,000,000 bp region or on another chromosome.

To increase our confidence in the eQTL regions identified by the three methods, only signals identified by at least two methods and within 1 Mbp of each other were retained. To do so, the interactions were split between *cis*-acting and *trans*-acting, and the size of each of the mapped eQTL regions (i.e. all of the interactions identified with the three aforementioned algorithms) was manually adjusted by adding 500,000 bp both upstream and downstream of the loci. The overlapping regions were subsequently identified using the genomic peak Venn function implemented in https://www.bioinformatics.com.cn/en, a free online platform for data analysis and visualization. To compute the interactions using this software, each interaction was codified as the following: cis/trans_genename_interactingchromosome startregioninteraction endregioninteraction. For example, *trans_Glyma.01G123600_GM05 40000 200000* would represent the region 40,000 – 200,000 on chromosome GM05 interacting in *trans* with *Glyma.01G123600*. The overlaps were identified using a pairwise comparison using the ICIM interactions as the reference signals in the ICIM vs. IM and ICIM vs. GCIM analyses. In addition, the IM signals were used as references in the IM vs. GCIM analysis. *Trans* interactions overlapping *cis* regions were *de facto* considered as *cis*

4.3.8 Regulatory Hotspot Mapping

To uncover regions associated with the regulation of the expression of multiple genes, we decided to identify the hotspots involved in the modulation of a high number of *trans* interactions (Fig. 3.1F). To do so, marker pairs delineating *trans* hotspots were qualified based on their respective (i) number of *trans* interactions and (ii) *trans* interaction density, and only those meeting both of these criteria were considered as markers flanking a hotspot region. The number of

interactions was identified by summing the number of *trans* interactions associated with a specific pair of markers and only the pairs of markers that were above the 95th (minor hotspot) or the 99th (major hotspot) percentiles threshold of all marker pairs from that population were retained. The *trans* interaction density was quantified by identifying the average number of *trans* interactions per kbp associated with the distance between the flanking of markers. A specific marker pair was deemed significant if its density was above the 80th percentile of all of the calculated *trans*-interaction densities. To facilitate the reading and understanding of the paper, each of the loci presented in this article is distinguished using either F2 ($QS15524_{F2:F3}$) or RIL ($QS15544_{RIL}$) in front of the region's name (e.g., F2_GM18:1,911,667-1,935,386). All of the Circos plots found in this paper were drawn using Circa V1 https://omgenomics.gumroad.com/l/circa.

4.3.9 Co-expression Network Analysis and Identification of Homologous Genes Using Protein Homology

Co-expression networks were built for the target genes and candidate TFs to understand their global expression pattern within the transcriptome (Fig. 3.1G). To understand the general expression pattern of the interactions associated with a specific hotspot, we generated the pairwise Pearson correlation coefficients (PCC) for the queried genes and clustered them using the pheatmap package (https://github.com/raivokolde/pheatmap) implemented in R. Transcriptome-wide CENs (TWCENs) were also generated using the $QS15524_{F2:F3}$ (38,692 genes) and $QS15544_{RIL}$ (40,218 genes) expression datasets. To do so, PCCs were generated using ≥ 0.85 (positive TWCEN, herein named POS_{TWCEN}) or ≤ -0.85 (negative TWCEN, herein named NEG_{TWCEN}) as thresholds for the expression datasets for the 176 and 162 lines for $QS15524_{F2:F3}$ and $QS15544_{RIL}$, respectively. The significant genes based on these thresholds were then annotated using the Soybase_GOtool to identify FRSPD_GO functions. Identification of homologous genes was performed using their peptide sequence with the Blast function in Phytozome V13 (Goodstein *et al.*, 2012), and peptide sequences exhibiting an E≤1e-5 were considered homologous.

4.3.10 Prediction of Transcription Factors and Identification of Candidate Single Nucleotide Polymorphisms

Following the mapping of eQTL interactions and regulatory hotspots, we predicted putative transcription factors that could be regulators for the four candidate genes

(*GmPRR1a*, *GmMDE04*, *GmLHCA4a*, and *GmCDF3*) of the *E8-r3* locus, a region previously identified between the GM04:41,808,599 and GM04:42,376,23 flanking markers (Fig. 3.1H; Appendix 2.4) (Gélinas Bélanger *et al.*, 2024a). To do so, we generated a list of 4,611 putative TFs (herein named TF_list$_{4,611}$) (Appendix 2.5). The TF_list$_{4,611}$ was generated by merging the genes with annotated TF functions from PlantTFDB (Jin *et al.*, 2017) and Soybase (Grant *et al.*, 2009) databases. This corresponded to a total of 658 and 864 unique genes from the PlantTFDB and Soybase, respectively. In addition, another common 3,089 genes were identified in both databases. To identify the best candidate TFs present in the identified loci, we subsequently annotated each of them using the (i) differential transcript expression datasets from the parents; (ii) positive and negative co-expression network datasets; and (iii) Soybase gene ontology annotations.

In addition, we used a custom variant analysis pipeline similar to the one detailed in Gélinas Bélanger *et al.* (2024a) to identify putative causal mutations in the predicted TFs (Fig. 3.1I). Prediction of deleterious effects of the SNPs within these TFs was performed using Ensembl Variant Effect Predictor (VEP) with Glycine_max_v2.1 (McLaren *et al.*, 2016). Mutations predicted as having moderate or high consequences on the protein structure or located in the 3'UTR/5'UTR were retained, whereas the others were removed from the dataset. The putative effects of the identified missense mutations were then predicted using Sorting Intolerant From Tolerant 4G (SIFT4g) (Ng and Henikoff, 2003; Kumar *et al.*, 2009). To predict the effects of these mutations, we generated a database using the annotations of *G. max* Wm82.a2.v1 from EnsemblPlants and the SIFT4G_Create_Genomic_DB guidelines https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB. Single nucleotide polymorphisms with SIFT scores ≥ 0.05 were considered as tolerated and those < 0.05 were considered as deleterious. Following the identification of the candidate SNPs, we verified the genotypes associated with them using the GmHapMap dataset and retained only the SNPs that were present in a single parental line.

4.4 Results

4.4.1 Linkage Map Construction and Differential Gene Expression the Parental Lines

In our previous QTL study (Gélinas Bélanger *et al.*, 2024a), 541,106,451 (QS15524$_{F2:F3}$) and 286,844,986 (QS15544$_{RIL}$) unique single-end reads were generated during the sequencing step of the full mapping populations. After filtering, two linkage maps were generated from 1,613 (QS15524$_{F2:F3}$; Appendix 2.6, 2.7) and 2,746 (QS15544$_{RIL}$; Appendix 2.6, 2.8) high-quality GBS-

derived SNP markers. To validate our choice of experimental conditions for both of our populations (i.e. RNA collected from the middle leaflet of the trifoliate leaf 4 hours after sunrise at the V4 leaf stage), we performed a differential gene expression analysis in both pairs of parental lines. Based on this analysis, we identified 10,216 DEGs (4,953 up-regulated genes and 5,263 down-regulated genes in OV) in the $QS15524_{F2:F3}$ parents (Fig. 3.2A, 3.2B; Appendix 2.9) and 1,430 DEGs (438 upregulated genes and 992 down-regulated genes in MD) in the $QS15544_{RIL}$ parents (Fig. 3.2C, 3.2D; Appendix 2.10). To find an explanation to the large difference of DEGs between each population, we inspected the pedigrees of $QS15524_{F2:F3}$ and $QS15544_{RIL}$ but did not find any obvious factors (e.g., a cross using a very exotic line) that would have caused this discrepancy (data not shown). Overall, we found that two of our candidate genes, *GmCDF3* and *GmMDE04*, were upregulated in the OV parental line of the $QS15524_{F2:F3}$ population. In addition, we found that many FRSPD_GO terms were significantly enriched for both populations (e.g. 'Regulation of Flower Development' in the $QS15524_{F2:F3}$ parents, Bonferroni corrected p-value of 3.05E-76), thus indicating a large difference in the abundance of transcripts of FRSPD genes in the parental lines of both populations (Appendix 2.11).

4.4.2 Mapping of eQTL Interactions

Subsequently, the linkage maps were used to perform genome-wide mapping of eQTL interactions for the $QS15524_{F2:F3}$ (38,693 genes) and $QS15544_{RIL}$ (40,223 genes) populations using a combinatorial approach based on the IM, ICIM, and GCIM algorithms populations (Appendix 2.12, 2.13). In the $QS15524_{F2:F3}$ population, the ICIM (4,735 *trans*/17 *cis*), IM (1,714 *trans*/10 *cis*), and GCIM (10,906 *trans*/32 *cis*) methods identified a varying number of interactions (Table 3.1). The same analysis was performed with the $QS15544_{RIL}$ population, with IM (17,375 *trans*/5,337 *cis*) having the highest number of interactions followed by ICIM (7,941 *trans*/2,862 *cis*) and then GCIM (4,418 *trans*/2,375 *cis*) (Table 3.1). To reduce the number of regions for further analyses, we decided to retain only interactions that were identified by at least two algorithms and which were overlapping or within a 1,000,000 bp distance from each other (Table 3.1; Appendix 2.14, 2.15, 2.16). This merging step reduced the number of interactions to 2,218 *trans* (2,061 genes)/7 *cis* (7 genes) (Fig. 3.3A) and 4,073 *trans* (2,842 genes)/3,083 *cis* (2,418 genes) (Fig. 3.4A) for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations, respectively (Appendix 2.17). Using our combinatorial approach, we identified that

the *trans* interactions were regulated by a total of 280 regions covering a total of ≈ 212.19 Mbp in the QS15524$_{F2:F3}$ population and 1,213 regions covering a total of ≈ 588.03 Mbp in the QS15544$_{RIL}$ population. The number of interactions per region was between 1 and 507 with a density between 3.67E-5 and 1,481 interactions/kbp for the QS15524$_{F2:F3}$ population. For the QS15544$_{RIL}$ population, the number of interactions per region was between 1 and 450 with a density between 3.09E-7 and 100 interactions/kbp.

4.4.3 Identification and Characterization of the Hotspots and Regions Associated with *E8-r3*

To identify important regulatory regions controlled by or controlling the *E8-r3* region, we began by classifying our *trans* eQTL regions into minor regions or hotspots (either minor or major hotspots) to identify the most promising regions (Appendix 2.18). To do so, we retrieved all of the regions above the 95th (minor hotspot) or 99th (major hotspot) percentiles for the number of interactions and the 80th percentile for the eQTL interaction density to identify either minor or major hotspots (Table 3.2). For the QS15524$_{F2:F3}$ population, the thresholds corresponded to 29 to 259 interactions for minor hotspots or ≥ 260 interactions for major hotspots with densities ≥ 0.09 interactions/kbp. For the QS15544$_{RIL}$ population, the thresholds corresponded to 9 to 16 interactions for minor hotspots or ≥ 17 interactions for major hotspots with densities ≥ 0.17 interactions/kbp. Using these thresholds, 8 hotspots (2 major and 6 minor) were identified in the QS15524$_{F2:F3}$ population (i.e. 2.85% of the total number of identified regions) (Fig. 3.3B; Table 3.2). Similarly, 34 hotspots (9 major and 25 minor) were identified in the QS15544$_{RIL}$ population (i.e. 1.23% of the total number of identified regions), with a large number of hotspots located near each other on chromosomes GM01, GM03, GM04, GM05 and GM09 (Fig. 3.4B; Table 3.2). Following the identification of the hotspots, we noticed that several had markers in common, suggesting that these regions might be regulated by one or several common loci. This included the (i) F2_GM18:1,434,182-1,911,667 and F2_GM18:1,911,667-1,935,386 and the (ii) F2_GM15:49,385,092-49,442,075 and F2_GM15:49,442,075-49,442,237 regions of the QS15524$_{F2:F3}$ population. Due to their close location, we merged the neighboring loci into two merged regions named F2_GM18:1,434,182-1,935,386 and F2_GM15:49,385,092-49,442,237 for the subsequent analyses. To characterize the hotspots, we subsequently performed a GO enrichment analysis on each of them and observed that four (i.e. three in QS15524$_{F2:F3}$ and one in

QS15544$_{RIL}$) were significantly enriched with terms associated with FRSPD functions (Fig. 3.3B, 3.4B; Appendix 2.19).

As we were interested in understanding the role of the *E8-r3* region and its interactions, we investigated whether the identified eQTL minor regions and eQTL hotspots interacted in *trans* with our four candidate genes (*GmCDF3*, *GmPPR1a*, *GmLHCA4a*, and *GmMDE04*) (Table 3.3). On the whole, we identified that the F2_GM18:1,434,182-1,935,386 hotspot was involved in the regulation of *GmLHCA4a*. We also detected that *GmPPR1a* was regulated by the F2_GM15:49,385,092-49,442,237 hotspot as well as the RIL_GM04:17,227,512-20,251,662, RIL_GM04:31,408,946-31,525,671 and RIL_GM13:37,289,785-38,620,690 minor regions. For *GmMDE04*, we identified one interaction with the F2_GM15:49,385,092-49,442,237 hotspot and one interaction with the RIL_GM04:17,227,512-20,251,662 minor region. No interactions were observed for *GmCDF3*, and as such, this gene was not investigated further. In addition to the interactions with the *E8-r3* candidate genes, we also identified several *trans* regulatory events with five additional genes (*Glyma.04G168100*, *Glyma.04G168000*, *Glyma.04G169300*, *Glyma.04G168200*, and *Glyma.04G169100*) found in the *E8-r3* locus, including several loci found on GM04 (Appendix 2.17).

4.4.4 The F2_GM18:1,434,182-1,935,386 Hotspot Regulates *GmLHCA4a* and Several Homologous Genes

To further understand the role of the F2_GM18:1,434,182-1,935,386 hotspot, we investigated to understand the specific FRSPD_GO functions of the 91 interactions (90 genes[1]). In addition, we used the TF_list$_{4,611}$ to identify candidate transcription factors and found three (*Glyma.18G020900*, *Glyma.18G025600*, and *Glyma.18G025800*) that were located within or close to this hotspot (Fig. 3.5A). To understand the co-expression patterns between the 90 target genes and three candidate transcription factors, we generated a CEN using pairwise PCCs between these 93 genes (Fig. 3.5B).

By doing so, we observed that 79 of the target genes, including *GmLHCA4a*, exhibited a similar co-expression pattern and were grouped as such into the F2_GM18:1,434,182-1,935,386_C1 cluster, a group specifically enriched with terms related to photosynthesis and

[1] One gene was found to be regulated by both regions.

response to light stimulus. Another group comprising 11 target genes was grouped in the F2_GM18:1,434,182-1,935,386_C2 cluster, a group without significantly enriched functions. We found that the three candidate TFs that were identified for the F2_GM18:1,434,182-1,935,386 hotspot all clustered in the C1 group, with *Glyma.18G025600* exhibiting the highest co-expression values with *GmLHCA4a*. Interestingly, we also discovered that the C1 cluster comprised a total of six *LHCA* homologs annotated with photosynthesis, response to light, photosystem I, and chlorophyll-binding functions in Soybase: (i) *Glyma.04G167900*/*GmLHCA4a* (our candidate gene); (ii) *Glyma.02G064700*/*GmLHCA1*; (iii) *Glyma.02G309500*/*GmLHCA3a*; (iv) *Glyma.06G194900*/*GmLHCA4b*; (v) *Glyma.14G003400*/*GmLHCA3b*; and (vi) *Glyma.15G179400*/*GmLHCA6*.

4.4.5 Functional Investigation and Variant Analysis of the Candidate Transcription Factors Regulating the *LHCA* Homologs

After identifying the three candidate TFs, we found that these genes were not annotated with FRSPD functions in Soybase. To gain further insights about them, we investigated the TWCENs between these genes and the 38,692 genes dataset from the $QS15524_{F2:F3}$ population (Fig. 3.6A; Appendix 2.20). Using a PCC threshold of ≥ 0.85 (POS_{TWCEN}), we found that the POS_{TWCEN} of *Glyma.18G020900*, *Glyma.18G025600* and *Glyma.18G025800* respectively comprised 2,230, 527 and 136 genes. For these three candidate TFs, we also constructed the NEG_{TWCEN} using a PCC threshold of ≤ -0.85 and discovered that the NEG_{TWCEN} of *Glyma.18G020900*, *Glyma.18G025600* and *Glyma.18G025800* respectively comprised 444, 1,849 and 0 genes. The genes found in POS_{TWCEN} and NEG_{TWCEN} of *Glyma.18G025600* were significantly enriched with functions associated with flower development, photosynthesis, and chlorophyll-binding, whereas the POS_{TWCEN} and NEG_{TWCEN} of *Glyma.18G020900* were less strongly associated with these functions (Fig. 3.6B; Appendix 2.21). For *Glyma.18G025800*, we found that its POS_{TWCEN} and NEG_{TWCEN} were not significantly enriched with any GO terms.

To further understand the putative roles of the three candidate TFs located in the F2_GM18:1,434,182-1,935,386_C1 cluster, we investigated their expression profiles, as well as the presence of mutations, in the parental lines. Based on our observations, *Glyma.18G020900* (Fold change, 3.66; FDR adjusted p-value, 1.93E-08) and *Glyma.18G025800* (Fold change, 2.17; FDR adjusted p-value, 0.047), were found to be

significantly upregulated in 'OAC Vision', whereas *Glyma.18G025600* was not differentially expressed (Fold change, 1.45; FDR adjusted p-value, 0.07) (Fig. 3.7; Appendix 2.9). Deeper investigations using our candidate SNP identification pipeline led to the identification of two SNPs in *Glyma.18G025600*, but none of the other candidate TFs (Table 3.4). Overall, the presence of variants in the 3'UTR of *Glyma.18G025600* but not in the other candidates, the high co-expression values between this candidate and all of the *LHCA* homologs, and the FRSPD functions associated with its POS_{TWCEN} and NEG_{TWCEN} suggest that *Glyma.18G025600* is the most likely candidate for the F2_GM18:1,434,182-1,935,386 hotspot.

4.4.6 The F2_GM15:49,385,092-49,442,237 Hotspot Regulates the *GmPRR1* and *GmMDE* Homologs

In addition to the regulation of *GmLHCA4a*, we also identified several regions regulating two *E8-r3* candidate genes, *GmPRR1a*, previously identified in Gélinas Bélanger *et al.* (2024a), and *GmMDE04*, proposed by Escamilla *et al.* (2024). As previously mentioned, both genes were found to be regulated by the F2_GM15:49,385,092-49,442,237 hotspot (Fig. 3.8A). To further investigate the networks interacting with this hotspot, we generated a CEN comprising 285 genes based on 293 *trans* interactions (53 from F2_GM15:49,385,092-49,442,075 and 240 from F2_GM15:49,442,075-49,442,237[2]). Along, we used the $TF\_list_{4,611}$ to identify candidate TFs and found two, *Glyma.15G261300* and *Glyma.15G263700*, that were located within or close to the region. Although the *Glyma.15G263700* TF was not found within the F2_GM15:49,385,092-49,442,237 hotspot, it was still included in the analysis due to its close proximity with this region (< 300 kbp). Subsequently, we generated a CEN with the 285 target genes along with the 2 candidate TFs (Fig. 3.8B) and observed the formation of two separate co-expression clusters: F2_GM15:49,385,092-49,442,237_C1 (184 target genes and one candidate TF, *Glyma.15G261300*) and F2_GM15:49,385,092-49,442,237_C2 (101 target genes and one candidate TF, *Glyma.15G263700*). Although no specific function was found to be significantly associated with the 184 target genes of the F2_GM15:49,385,092-49,442,237_C1 cluster, this cluster contained *GmPRR1a* and two of its homologs (*Glyma.17G102200*/*PRR1d* and *Glyma.07G171200*/*RESPONSE REGULATOR 2*/*ARR2*). In addition, we identified *GmMDE04* and

[2] Four genes were found to be regulated by both regions.

two homologs, *Glyma.13G052100* (*GmMDE13*) and *Glyma.19G034600* (*GmMDE19*), in the F2_GM15:49,385,092-49,442,237_C2 cluster, a group enriched with an oxidative photosynthetic carbon pathway function. Interestingly, one additional *MDE* homolog, *GmMDE17* (*Glyma.17G081200*), was found to be regulated by the F2_GM15:49,385,092-49,442,237 hotspot with ICIM (LOD, 4.75; PVE, 8.56%), but none of the two other software (Appendix 2.12).

4.4.7 Functional Investigation and Variant Analysis of the Candidate Transcription Factors Regulating the *PRR* and *MDE* Homologs

Following the building of the CEN, we generated TWCENs for both candidate TFs of the F2_GM15:49,385,092-49,442,075 hotspot (Appendix 2.20). Using a PCC threshold of ≤ -0.85, we found that the NEG_{TWCEN} of *Glyma.15G263700* was large and comprised 1,284 genes, whereas *Glyma.15G261300* had none (Fig. 3.9A). Nothing conclusive was found for the POS_{TWCEN} of both candidates as the POS_{TWCEN} of *Glyma.15G263700* comprised only 21 genes and *Glyma.15G261300* had none. Subsequently, we performed a GO enrichment analysis on the NEG_{TWCEN} of *Glyma.15G263700* and discovered that it was significantly enriched with various FRSPD terms associated with flowering, floral organ formation, and photomorphogenesis (Fig. 3.9B; Appendix 2.21). This result is coherent with the fact that the *Glyma.15G263700* candidate TF is annotated with terms related to flower development in Soybase, whereas *Glyma.15G261300* is not annotated with FRSPD functions. To gain insights regarding the putative roles of these two candidate TFs, we investigated their expression profiles, as well as the presence of mutations, in the parental lines. In the $QS15524_{F2:F3}$ parents, *Glyma.15G263700* (Fold change, 2.60; FDR adjusted p-value, 5.15E-04) was differentially expressed, but not Glyma.15G261300 (Fold change, 1.90; FDR adjusted p-value, 0.03) (3.7; Appendix 2.9). Based on our variant analysis pipeline, we discovered mutations in both genes, with *Glyma.15G263700* displaying a missense mutation predicted to be deleterious by the SIFT algorithm and *Glyma.15G261300* having a 3'UTR variant at position GM15:49,385,259 (Table 3.4).

4.4.8 *GmPRR1* and *GmMDE* Homologs Are Regulated by the Same Minor Regions

Following these discoveries from the F2_GM15:49,385,092-49,442,237 hotspot (Fig. 3.10A), we investigated further to determine whether similar co-regulation events could be

observed for minor regions interacting in *trans* with *GmPRR* and *GmMDE* homologs. On the whole, we identified three different minor regions in the $QS15544_{RIL}$ population that were interacting with three *GmPRR* (*GmPRR1a*, *GmPPR1d*, and *Glyma.05G025000/GmPRR4*), and two *GmMDE* (*GmMDE04*, and *Glyma.06G205800*/*GmMDE06*) homologous genes: (i) RIL_GM04:17,227,512-20,251,662 (Fig. 3.10B); (ii) RIL_GM04:31,408,946-31,525,671 (Fig. 3.10C); and (iv) RIL_GM13:37,289,785-38,620,690 (Fig. 3.10D). Interestingly, *GmPPR1d* was found to be regulated by a region located between markers GM04:22,010,259 and GM04:26,441,718 which is adjacent to the minor region RIL_GM04:17,227,512-20,251,662 regulating *GmPRR1a* and *GmMDE04* (Appendix 2.17). For each of the regions, the number of candidate TFs ranged between 1 (for RIL_GM04:31,408,946-31,525,671) to 6 (for RIL_GM13:37,289,785-38,620,690) (Appendix 2.22). We performed the same analyses (i.e. expression analysis, TWCEN, and variant analysis pipeline) for these minor regions than for the F2_GM18:1,434,182-1,935,386 and F2_GM15:49,385,092-49,442,237 hotspots (Appendix 2.22). Overall, two (out of 11) candidates were annotated with FRSPD terms and none of them were found to be differentially expressed in the parental lines due to FDR-adjusted p-values that were above the threshold. We constructed TWCENs for all of the candidates but only *Glyma.04G135400* was found to have POS_{TWCEN} (446 genes) and NEG_{TWCEN} (445 genes) that were significantly enriched with GO terms including the 'Phytochrome binding' term (Appendix 2.21). Using our custom variant analysis pipeline, we found a total of five mutations in two different candidates (*Glyma.04G135400* and *Glyma.13G285400*) (Table 3.5). Based on these observations, we think that *Glyma.04G135400* is the best candidate for RIL_GM04:17,227,512-20,251,662. Still, we think that more research on neighboring candidate TFs outside of the RIL_GM04:31,408,946-31,525,671 and RIL_GM13:37,289,785-38,620,690 needs to be performed to identify better candidates.

4.5 Discussion

4.5.1 The F2_GM18:1,434,182-1,935,386 Hotspot is a Hub for the Coordinated Regulation of the Light Response and Photosynthetic Mechanisms

Photosystem I (PSI) is located in the thylakoid membrane and is a multiprotein complex that plays a crucial role in oxygenic photosynthesis by oxidizing plastocyanin and reducing ferredoxin (Sláma *et al.*, 2023). PSI is divided into the core complex and the outer antenna

complexes (also known as Light-Harvesting Complex I; LHCI). The role of the LHCI is to harvest light and transfer the excitation energy of the electrons to the reaction center. The antenna of the Light-Harvesting Complex I comprises four subunits which are the products of *Lhca1-4* genes in *Arabidopsis*. We previously demonstrated that the *E8-r3* region (GM04:41,808,599-42,376,237) regulates the number of days to maturity under a constant short-day photoperiod in both $QS15524_{F2:F3}$ and $QS15544_{RIL}$. Based on these observations, we proposed the *Glyma.04G167900* (*GmLHCA4a*) gene as a potential candidate for this region using a candidate SNP analysis (Gélinas Bélanger *et al.*, 2024a). In previous studies, *GmLHCA4a* has been identified as a candidate for the *q4–2* locus regulating leaf-related traits (i.e. leaf size and shape) and chlorophyll content (Yu *et al.*, 2020). Furthermore, *GmLHCA4a* has been suggested to be involved in the number of days to flowering as Liu *et al.* (2021) observed a 1.8-day difference between two *GmLHCA4a* haplotypes under short-day growth conditions. In the present study, we demonstrated that the F2_GM18:1,434,182-1,935,386 hotspot regulates the C1 cluster, a group of 79 genes regulating photosynthesis and light response mechanism in the $QS15524_{F2:F3}$ population. The F2_GM18:1,434,182-1,935,386_C1 cluster includes six Light-Harvesting Complex homologs and several genes associated with PSI (e.g., *Glyma.10G042100*/*PHOTOSYSTEM I SUBUNIT E-2*) and PSII (e.g., *Glyma.10G089300*, *Glyma.10G089500* and *Glyma.15G275600*; all *PHOTOSYSTEM II 5 KDA* proteins). In soybean, 62 proteins, including 34 *LHC A/B* proteins, have been predicted to be involved in the regulation of the Light-Harvesting Complex (Lan *et al.*, 2022). Consequently, the genes that were identified for the light response subcluster represent only a fraction of the LHC genes within the soybean genome.

4.5.2 The *Glyma.18G025600* Gene is the Best Candidate Regulator for the *LHC* Homologs

From these observations, we identified three candidate TFs (*Glyma.18G020900*, *Glyma.18G025600*, and *Glyma.18G025800*) located within the F2_GM18:1,434,182-1,935,386 hotspot. Based on our co-expression analysis, *Glyma.18G020900* and *Glyma.18G025600* were strongly co-expressed with the *LHC* homologs, but only *Glyma.18G025600* had POS_{TWCEN} and NEG_{TWCEN} associated with photosynthetic and photosystem I/II regulation functions. Interestingly, *Glyma.18G025600* harbored two mutations in its 3'UTR in OV, whereas none were found in *Glyma.18G020900*. To the best of our knowledge, this is the first

time *Glyma.18G025600* is proposed as a candidate for the transcriptional regulation of these six *LHC* homologs and more largely to the targets of the F2_GM18:1,434,182-1,935,386_C1 cluster in soybean. The *LOB DOMAIN-CONTAINING PROTEIN 21* gene (*Glyma.18G025600*; *GmLBD21*) is the ortholog of AT3G11090 (*AtASL12*/*AtLBD21*) in *Arabidopsis* which belongs to the class 1a of the AS2 protein family (Iwakawa *et al.*, 2002; Shuai *et al.*, 2002). The *AtAS2* gene (AT1G65620) encodes a domain that includes a leucine-zipper-like sequence in its amino-terminal half and a cysteine repeat (Matsumura *et al.*, 2009). On a functional level, *AtAS2* plays a role in the expansion of flat leaf lamina in *Arabidopsis* as *AtAS2* overexpressing and loss-of-function *Arabidopsis* mutants respectively exhibited upwardly and downwardly curled leaves (Iwakawa *et al.*, 2002). The gene *DOWN IN DARK AND AUXIN1* (AT3G27650; *AtDDA1/AtLBD25*/*AtASL3*), a gene closely associated with *LATERAL ORGAN BOUNDARIES* (*AtLOB*) and *ASYMMETRIC LEAVES2* (*AS2*), has been suggested to be implicated in photomorphogenesis and auxin response as *dda1-1* plants display aberrant hypocotyl elongation and reduced sensitivity to auxin phenotypes (Mangeon *et al.*, 2011). Overall, the current pieces of evidence suggest that *Glyma.18G025600* is the best candidate for the F2_GM18:1,434,182-1,935,386 hotspot.

4.5.3 *GmPRR1a* and *GmMDE04* are Co-Regulated by the Same Regions

The *GmPRR1a* and *GmMDE04* genes and their homologs are known to have critical impacts on photoperiodic flowering in *Arabidopsis* and soybean. The soybean PSEUDO RESPONSE REGULATORs 1a and 1d are orthologs of the *Arabidopsis* DNA-binding transcription factor TIMING OF CAB EXPRESSION 1/PSEUDO-RESPONSE REGULATOR 1 (AtTOC1/AtPRR1) which contains a CCT (CONSTANS, CO-like, TOC1) domain in the C terminus and a pseudo receiver domain in the N terminus (Gendron *et al.*, 2012). In *Arabidopsis*, this protein is known to be involved in the phytochrome regulation of circadian gene expression and photomorphogenic response (Más *et al.*, 2003) and thus acts as a molecular bridge between environmental cues and clock outputs. In soybean, *GmMDE04* (also named *GmFULb*) is involved in the *E1-GmMDEs-GmFT2a/5a-Dt1* signaling pathway and responds to photoperiod at the transcript level (Zhai *et al.*, 2022). Overexpression experiments have demonstrated that the *GmMDE06* homolog acts downstream of *E1* in the induction of the flowering process, increases the expression of *GmFT2a/GmFT5a* and promotes the termination of stem growth by

repressing *Dt1* (Zhai *et al.*, 2022). According to Zhai *et al.* (2022), *GmMDE04* is significantly expressed under short-day conditions versus long-day conditions in the 'Harosoy-E1', 'Zhonghuang 13' and 'Gaofeng1' backgrounds but not in 'Harosoy-e1', 'Kariyutaka', and 'Sidou 11'. This gene is the ortholog of the *Arabidopsis* gene *AGAMOUS-LIKE 8*/*FRUITFUL* which induces global proliferative arrest (i.e. the coordinated arrest of all active meristems) by repressing members of the *APETALA2* (*AtAP2*) clade involved in the maintenance of the shoot apical meristem (Martínez-Fernández *et al.*, 2020). As a whole, the suppression in *E1* expression has been demonstrated to be tightly associated with the photoperiod-insensitive expression of *GmMDE*s (Zhai *et al.*, 2022). Structurally, *GmMDE04* and *GmMDE06* exhibit a higher degree of similitude between each other than for the five other *MDE* genes (results not shown).

In the present study, we identified two co-regulation events between *GmPRR1a* and *GmMDE04*. The first was found in $QS15524_{F2:F3}$ (F2_GM15:49,385,092-49,442,237 hotspot), whereas the second was discovered in $QS15544_{RIL}$ (RIL_GM04:17,227,512-20,251,662). In addition, we also identified two other co-regulation events between *PRR* and *MDE* homologs (i.e., *GmPRR1d*, *GmARR2*, and *GmPRR4*) and MDE (*GmMDE06*, *GmMDE13*, *GmMDE17*, and *GmMDE19*) in the $QS15544_{RIL}$ population. Each of these regulation events were identified by at least two algorithms, except the interaction between *GmMDE17* and the F2_GM15:49,385,092-49,442,237 hotspot, thus indicating the robustness and reliability of these interactions. Still, we consider the interaction between *GmMDE17* and the F2_GM15:49,385,092-49,442,237 hotspot to be robust as we consider ICIM to be the one of the most reliable algorithms currently available to researchers. Regarding RIL_GM04:17,227,512-20,251,662, we discovered that this locus is located near the *E8-r1* locus (RIL_GM04:16,974,874-17,152,230), a locus discovered in the same study as for *E8-r3* and in the same population (Gélinas Bélanger *et al.*, 2024a); however, this locus was found for the pod-filling trait under field conditions and was not considered for the present study. Still, the data generated in Gélinas Bélanger *et al.* (2024a) demonstrate that a critical regulator is found within the same genomic region.

For the co-regulation events associated with the F2_GM15:49,385,092-49,442,237 hotspot, two genes (*Glyma.15G261300* and *Glyma.15G263700*) have been proposed as candidate regulators. At present, several lines of evidence (i.e. functional annotations, TWCEN functions, type of prevailing mutations, and more) suggest that *Glyma.15G263700* (*ALTERED PHLOEM DEVELOPMENT*/*GmAPL*; also called *GmFE*) is the best candidate. The *GmAPL* gene is the

ortholog of AT1G79430 in *Arabidopsis*, a phloem-specific Myb-related protein involved in the photoperiodic induction of flowering (Abe *et al.*, 2015). Abe *et al.* (2015) have demonstrated that a missense mutation causing a glycine (G) to glutamic acid (E) substitution causes a late-flowering phenotype in *Atfe* mutants. Using expression analysis, Abe *et al.* (2015) have shown that a fully functioning *AtFE* allele is required for the transcriptional activation of *FLOWERING LOCUS T INTERACTING PROTEIN 1* (*AtFTIP1*), a critical gene involved in the selective trafficking of *AtFT* protein from phloem companion cells to sieve elements (Liu *et al.*, 2012).

4.6 Conclusion

We developed a novel eQTL mapping pipeline that enabled us to identify hundreds of transient *cis* and *trans* interactions in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ soybean populations. From the *trans* interactions, we identified four hotspots involved in the regulation of FRSPD functions: (i) F2_GM06:39,892,719-43,437,125, F2_GM17:5,431,473-7,260,313 and F2_GM18:1,434,182-1,935,386 in $QS15524_{F2:F3;}$ and (ii) the RIL_GM04:10,812,813-10,985,437 in $QS15544_{RIL}$. Deeper investigations identified *trans* regulatory events between: (i) F2_GM18:1,434,182-1,935,386 and *GmLHCA4a*; and (ii) several regions identified in $QS15524_{F2:F3}$ and $QS15544_{RIL}$ and two candidate genes (*GmPRR1a* and *GmMDE04*) along with some homologs (*GmPRR1d*, *GmPRR4*, and *GmMDE06*). Using an approach combining the analysis of predicted TFs, TWCEN, annotated functions, and genomic variants, we identified several candidates for these regions of interest, with a focus on *GmLBD21* (*Glyma.18G025600*) and *GmAPL* (*Glyma.15G263700*). Overall, the discoveries regarding the loci regulating the three candidate genes for the *E8-r3* region (*GmLHCA4a*, *GmPRR1a*, and *GmMDE04*) represent only a small proportion of the *trans* and *cis* interactions captured with our combinatorial mapping approach. These findings demonstrate the potential of eQTL interactions and hotspot mapping combined with co-expression analyses to identify a large number of TF-related regulatory events and narrow the number of potential TF candidates.

4.7 Supplemental data

Appendix 2.1 Gene expression dataset for the $QS15524_{F2:F3}$ parental lines.

Appendix 2.2 Gene expression dataset for the $QS15544_{RIL}$ parental lines.

Appendix 2.3 FRSPD gene ontology annotations from Soybase.

Appendix 2.4 Single nucleotide polymorphisms found in the *E8-r3* region.

Appendix 2.5 Transcription factors from PlantTFDB and Soybase.

Appendix 2.6 Linkage maps for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Appendix 2.7 Genotypes for the 1,613 markers of the $QS15524_{F2:F3}$ population.

Appendix 2.8 Genotypes for the 2,746 markers of the $QS15544_{RIL}$ population.

Appendix 2.9 Differentially expressed genes for the $QS15524_{F2:F3}$ parental lines.

Appendix 2.10 Differentially expressed genes for the $QS15544_{RIL}$ parental lines.

Appendix 2.11 Gene ontology annotations for the differentially expressed genes of the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ parental lines.

Appendix 2.12 Mapping of the eQTL interactions in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations using three algorithms (ICIM, IM, and GCIM).

Appendix 2.13 General statistical relative to the cis and trans interactions mapping in both populations.

Appendix 2.14 Number of interactions per gene in both populations before merging.

Appendix 2.15 Number of interactions per gene in both populations after merging.

Appendix 2.16 General statistics for the number of eQTL interactions before and after merging.

Appendix 2.17 Expression QTL interactions after the merging in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Appendix 2.18 Statistics relevant to the identification of eQTL hotspots.

Appendix 2.19 Gene ontology annotations associated with the eQTL hotspots.

Appendix 2.20 Transcription-wide co-expression network statistics.

Appendix 2.21 Gene ontology annotations associated with the transcriptome-wide co-expression networks.

Appendix 2.22 Minor regions and candidate transcription factors associated with the regulation of the *E8-r3* genes.

4.8 Information

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, BioProject PRJNA1035514.

Author contribution

JGB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. TC: Conceptualization, Data curation, Investigation, Methodology, Software, Supervision, Validation, Writing – review & editing. VH-V: Supervision, Writing – review & editing. LO'D: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work presented here was supported by Génome Québec and Génome Canada with funds awarded to the SoyaGen Project by the Canadian Field Crop Research Alliance and Agriculture and Agri-Food Canada under the Agri-Innovation Program. JGB was supported by les Fonds de recherche du Québec volet Nature et Technologie, Centre SÈVE, Mitacs, the Natural Sciences and Engineering Research Council of Canada, and Seed World Group.

Acknowledgments

We would like to acknowledge the work of Éric Fortier for the extraction of RNA and phenotypic data collection. In addition, we would like to thank Joannie Berthon for her help with the phenotypic data collection.

Conflict of Interest

The authors JGB, TC, and LO'D were employed by the company Centre de recherche sur les grains (CÉROM) Inc. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any

product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

4.9 References

Abe, M., Kaya, H., Watanabe-Taneda, A., Shibuta, M., Yamaguchi, A., Sakamoto, T., *et al.* (2015). FE, a phloem-specific Myb-related protein, promotes flowering through transcriptional activation of FLOWERING LOCUS T and FLOWERING LOCUS T INTERACTING PROTEIN 1. *Plant J.* 83, 1059–1068. doi: 10.1111/tpj.12951

Abed, A., Légaré, G., Pomerleau, S., St-Cyr, J., Boyle, B., and Belzile, F. J. (2019). "Genotyping-by-sequencing on the ion torrent platform in barley," in *Barley*, (Springer), 233–252.

Ali, T., Zhou, B., Cleary, D., and Xie, W. (2022). The Impact of Climate Change on China and Brazil's Soybean Trade. *Land* 11, 2286.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169.

Bian, S., Jin, D., Li, R., Xie, X., Gao, G., Sun, W., *et al.* (2017). Genome-wide analysis of CCA1-like proteins in soybean and functional characterization of GmMYB138a. *Int. J. Mol. Sci.* 18, 2040. doi: 10.3390/ijms18102040

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Bylino, O. V., Ibragimov, A. N., and Shidlovskii, Y. V. (2020). Evolution of regulated transcription. *Cells* 9, 1–38. doi: 10.3390/cells9071675

Choi, J. H., Kim, T., Jung, J., and Joo, J. W. J. (2020). Fully automated web-based tool for identifying regulatory hotspots. *BMC Genomics* 21, 1–7. doi: 10.1186/s12864-020-07012-z

Corrales, A. R., Carrillo, L., Lasierra, P., Nebauer, S. G., Dominguez-Figueroa, J., Renau-Morata, B., *et al.* (2017). Multifaceted role of cycling DOF factor 3 (CDF3) in the regulation of flowering time and abiotic stress responses in *Arabidopsis*. *Plant Cell Environ.* 40, 748–764. doi: 10.1111/pce.12894

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., *et al.* (2011). The

variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dietz, N., Chan, Y. O., Scaboo, A., Graef, G., Hyten, D., Happ, M., *et al.* (2022). Candidate Genes Modulating Reproductive Timing in Elite US Soybean Lines Identified in Soybean Alleles of *Arabidopsis* Flowering Orthologs With Divergent Latitude Distribution. *Front. Plant Sci.* 13, 1–14. doi: 10.3389/fpls.2022.889066

Escamilla, D. M., Dietz, N., Bilyeu, K., Hudson, K., and Rainey, K. M. (2024). Genome-wide association study reveals GmFulb as candidate gene for maturity time and reproductive length in soybeans (Glycine max). *PLoS One* 19, e0294123. doi: 10.1371/journal.pone.0294123

Ge, S. X., Son, E. W., and Yao, R. (2018). iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* 19, 1–24. doi: 10.1186/s12859-018-2486-6

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., and O'Donoughue, L. (2024a). Dissection of the E8 locus in two early maturing Canadian soybean populations. *Front. Plant Sci.* 15, 1329065. doi: 10.3389/fpls.2024.1329065

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., O'Donoughue, L., Gélinas Bélanger, J., Copley, T. R., *et al.* (2024b). Identification of Quantitative Trait Loci Associated with Seed Quality Traits in Two Early-Maturing Soybean Populations. *Can. J. Plant Sci.* 0, null. doi: 10.1139/cjps-2024-0049

Gendron, J. M., Pruneda-Paz, J. L., Doherty, C. J., Gross, A. M., Kang, S. E., and Kay, S. A. (2012). *Arabidopsis* circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proc. Natl. Acad. Sci.* 109, 3167–3172. doi: 10.1073/pnas.1200355109

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24, 408–415. doi: 10.1016/j.tig.2008.06.001

Goedhart, J., and Luijsterburg, M. S. (2020). VolcaNoseR is a web app for creating, exploring, labeling and sharing volcano plots. *Sci. Rep.* 10, 20560.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., *et al.* (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* 40, 1178–1186. doi: 10.1093/nar/gkr944

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, 843–846. doi:

10.1093/nar/gkp798

Hotta, C. T. (2021). From crops to shops: How agriculture can use circadian clocks. *J. Exp. Bot.* 72, 7668–7679. doi: 10.1093/jxb/erab371

Huang, L., Liu, X., Pandey, M. K., Ren, X., Chen, H., Xue, X., *et al.* (2020). Genome-wide expression quantitative trait locus analysis in a recombinant inbred line population for trait dissection in peanut. *Plant Biotechnol. J.* 18, 779–790. doi: 10.1111/pbi.13246

Iwakawa, H., Ueno, Y., Semiarti, E., Onouchi, H., Kojima, S., Tsukaya, H., *et al.* (2002). The ASYMMETRIC LEAVES2 Gene of *Arabidopsis* thaliana, Required for Formation of a Symmetric Flat Leaf Lamina, Encodes a Member of a Novel Family of Proteins Characterized by Cysteine Repeats and a Leucine Zipper. *Plant Cell Physiol.* 43, 467–478. doi: 10.1093/pcp/pcf077

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., *et al.* (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14, R36. doi: 10.1186/gb-2013-14-4-r36

Kong, F., Liu, B., Xia, Z., Sato, S., Kim, B. M., Watanabe, S., *et al.* (2010). Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in Soybean. *Plant Physiol.* 154, 1220–1231. doi: 10.1104/pp.110.160796

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1082. doi: 10.1038/nprot.2009.86

Lan, Y., Song, Y., Zhao, F., Cao, Y., Luo, D., Qiao, D., *et al.* (2022). Phylogenetic, Structural and Functional Evolution of the LHC Gene Family in Plant Species. *Int. J. Mol. Sci.* 24, 488. doi: 10.3390/ijms24010488

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics* 175, 361–374. doi: 10.1534/genetics.106.066811

Lin, C.-S., and Poushinsky, G. (1985). a Modified Augmented Design (Type 2) for Rectangular Plots. *Can. J. Plant Sci.* 65, 743–749. doi: 10.4141/cjps85-094

Lin, X., Dong, L., Tang, Y., Li, H., Cheng, Q., Li, H., *et al.* (2022). Novel and multifaceted

regulations of photoperiodic flowering by phytochrome A in soybean. *Proc. Natl. Acad. Sci.* 119, e2208708119. doi: 10.1073/pnas.2208708119

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., and Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180, 995–1007. doi: 10.1534/genetics.108.092742

Liu, C., Chen, X., Wang, W., Hu, X., Han, W., He, Q., *et al.* (2021). Identifying wild versus cultivated gene-alleles conferring seed coat color and days to flowering in Soybean. *Int. J. Mol. Sci.* 22, 1–22. doi: 10.3390/ijms22041559

Liu, H., Wang, H., Gao, P., Xü, J., Xü, T., Wang, J., *et al.* (2009). Analysis of clock gene homologs using unifoliolates as target organs in soybean (Glycine max). *J. Plant Physiol.* 166, 278–289. doi: 10.1016/j.jplph.2008.06.003

Liu, L. feng, Gao, L., Zhang, L. xin, Cai, Y. peng, Song, W. wen, Chen, L., *et al.* (2022). Co-silencing E1 and its homologs in an extremely late-maturing soybean cultivar confers super-early maturity and adaptation to high-latitude short-season regions. *J. Integr. Agric.* 21, 326–335. doi: 10.1016/S2095-3119(20)63391-3

Liu, L., Liu, C., Hou, X., Xi, W., Shen, L., Tao, Z., *et al.* (2012). FTIP1 is an essential regulator required for florigen transport. *PLoS Biol.* 10, e1001313. doi: 10.1371/journal.pbio.1001313

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. doi: 10.1186/s13059-014-0550-8

Mangeon, A., Bell, E. M., Lin, W., Jablonska, B., and Springer, P. S. (2011). Misregulation of the LOB domain gene DDA1 suggests possible functions in auxin signalling and photomorphogenesis. *J. Exp. Bot.* 62, 221–233.

Martínez-Fernández, I., de Moura, S., Alves-Ferreira, M., Ferrándiz, C., and Balanzà, V. (2020). Identification of Players Controlling Meristem Arrest Downstream of the FRUITFULL-APETALA2 Pathway . *Plant Physiol.* 184, 945–959. doi: 10.1104/pp.20.00800

Más, P., Alabadí, D., Yanovsky, M. J., Oyama, T., and Kay, S. A. (2003). Dual role of TOC1 in the control of circadian and photomorphogenic responses in *Arabidopsis*. *Plant Cell* 15, 223–236. doi: 10.1105/tpc.006734

Matsumura, Y., Iwakawa, H., Machida, Y., and Machida, C. (2009). Characterization of genes in the ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES (AS2/LOB) family in *Arabidopsis* thaliana, and functional and molecular comparisons between AS2 and other

family members. *Plant J.* 58, 525–537.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., *et al.* (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0974-4

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Morales, A. M. A. P., O'Rourke, J. A., van de Mortel, M., Scheider, K. T., Bancroft, T. J., Borém, A., *et al.* (2013). Transcriptome analyses and virus induced gene silencing identify genes in the *Rpp4*-mediated Asian soybean rust resistance pathway. *Funct. Plant Biol.* 40, 1029–1047. Available at: https://doi.org/10.1071/FP12296

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., *et al.* (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918. doi: 10.1038/ng.3036

Schaalje, G. B., Lynch, D. R., and Kozub, G. C. (1987). Field evaluation of a modified augmented design for early stage selection involving a large number of test lines without replication. *Potato Res.* 30, 35–45. doi: 10.1007/BF02357682

Shuai, B., Reynaga-Peña, C. G., and Springer, P. S. (2002). The Lateral Organ Boundaries gene defines a novel, plant-specific gene family. *Plant Physiol.* 129, 747–761. doi: 10.1104/pp.010926

Sláma, V., Cupellini, L., Mascoli, V., Liguori, N., Croce, R., and Mennucci, B. (2023). Origin of Low-Lying Red States in the Lhca4 Light-Harvesting Complex of Photosystem I. *J. Phys. Chem. Lett.* 14, 8345–8352. doi: 10.1021/acs.jpclett.3c02091

Sun, H., Jia, Z., Cao, D., Jiang, B., Wu, C., Hou, W., *et al.* (2011). GmFT2a, a soybean homolog of flowering locus T, is involved in flowering transition and maintenance. *PLoS One* 6, 18–20. doi: 10.1371/journal.pone.0029238

The American Soybean Association (2023). SoyStats - International: World Soybean Production (2021/2022 year). *Am. Soybean Assoc.* Available at: http://soystats.com/ (Accessed December 11, 2023).

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., and Belzile, F. (2017). Fast-GBS: A new

pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics* 18, 1–7. doi: 10.1186/s12859-016-1431-9

Unc, A., Altdorff, D., Abakumov, E., Adl, S., Baldursson, S., Bechtold, M., *et al.* (2021). Expansion of Agriculture in Northern Cold-Climate Regions: A Cross-Sectoral Perspective on Opportunities and Challenges. *Front. Sustain. Food Syst.* 5, 663448. doi: 10.3389/fsufs.2021.663448

Venkat, A., and Muneer, S. (2022). Role of Circadian Rhythms in Major Plant Metabolic and Signaling Pathways. *Front. Plant Sci.* 13, 836244. doi: 10.3389/fpls.2022.836244

Wang, J., Yu, H., Weng, X., Xie, W., Xu, C., Li, X., *et al.* (2014). An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *J. Exp. Bot.* 65, 1069–1079. doi: 10.1093/jxb/ert464

Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., *et al.* (2010). A global analysis of *QTL* for expression variations in rice shoots at the early seedling stage. *Plant J.* 63, 1063–1074. doi: 10.1111/j.1365-313X.2010.04303.x

Wang, X., and Ma, L. (2013). Unraveling the circadian clock in *Arabidopsis*. *Plant Signal. Behav.* 8, 1–6. doi: 10.4161/psb.23014

Watanabe, S., Hideshima, R., Zhengjun, X., Tsubokura, Y., Sato, S., Nakamoto, Y., *et al.* (2009). Map-based cloning of the gene associated with the soybean maturity locus E3. *Genetics* 182, 1251–1262. doi: 10.1534/genetics.108.098772

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., *et al.* (2011). A map-based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering. *Genetics* 188, 395–407. doi: 10.1534/genetics.110.125062

West, M. A. L., Kim, K., Kliebenstein, D. J., Van Leeuwen, H., Michelmore, R. W., Doerge, R. W., *et al.* (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175, 1441–1450. doi: 10.1534/genetics.106.064972

Westra, H. J., and Franke, L. (2014). From genome to function by studying eQTL. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1896–1902. doi: 10.1016/j.bbadis.2014.04.024

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., *et al.* (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus

E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2155–E2164. doi: 10.1073/pnas.1117982109

Yu, K., Wang, J., Sun, C., Liu, X., Xu, H., Yang, Y., *et al.* (2020). High-density QTL mapping of leaf-related traits and chlorophyll content in three soybean RIL populations. *BMC Plant Biol.* 20, 470. doi: 10.1186/s12870-020-02684-x

Zhai, H., Wan, Z., Jiao, S., Zhou, J., Xu, K., Nan, H., *et al.* (2022). GmMDE genes bridge the maturity gene E1 and florigens in photoperiodic regulation of flowering in soybean. *Plant Physiol.* 189, 1021–1036. doi: 10.1093/plphys/kiac092

Zhang, Y., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2020). QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked *QTL* for quantitative traits in bi-parental segregation populations. *Comput. Struct. Biotechnol. J.* 18, 59–65. doi: 10.1016/j.csbj.2019.11.005

Zhu, J., Takeshima, R., Harigai, K., Xu, M., Kong, F., Liu, B., *et al.* (2019). Loss of function of the E1-like-B gene associates with early flowering under long-day conditions in Soybean. *Front. Plant Sci.* 9, 1–13. doi: 10.3389/fpls.2018.01867

Population	Method	Number of eQTL interactions		Merging of the eQTL regions using the genomic peak Venn function				
				Methods involved in the merging	Number of regions with duplicates[1]	Number of unique regions (without duplicates)[2]	Number of regions with duplicates[1]	Number of unique regions (without duplicates)[2]
		Trans	*Cis*		*Trans*		*Cis*	
$QS15524_{F2:F3}$	**ICIM**	4,735	17	**ICIM vs IM**	1,340		7	
	IM	1,714	10	**ICIM vs GCIM**	1,134		1	
	GCIM	10,906	32	**IM vs GCIM**	31		1	
	Total	**17,355**	**59**	**Total**	2,505	**2,218**	9	**7**
$QS15544_{RIL}$	**ICIM**	7,941	2,862	**ICIM vs IM**	2,058		1,046	
	IM	17,375	5,337	**ICIM vs GCIM**	1,796		2,302	
	GCIM	4,418	2,375	**IM vs GCIM**	1,649		1,046	
	Total	**29,734**	**10,574**	**Total**	5,503	**4,073**	4,394	**3,083**

[1] Number of eQTL regions after the merge with the genomic peak Venn function. The total number includes all of the duplicated regions between the methods.
[2] Number of eQTL regions after the merge with the genomic peak Venn function but without the duplicates found in each of the merging steps.

Table 3.1 Number of eQTL interactions and eQTL regions before and after the merge using the genomic peak Venn function.

Population	Chromosome	Start of region (bp)	End of region (bp)	Size of region (bp)	Nb of interactions	Density (interactions/ kbp)	Percentile
$QS15524_{F2:F3}$	GM06	39,892,719	43,437,125	3,544,406	507	0.14	Major (99^{th})
	GM17	5,431,473	7,260,313	1,828,840	342	0.19	
$QS15544_{RIL}$	GM01	39,404,966	39,405,971	1,005	17	16.92	Major (99^{th})
	GM02	46,648,011	46,714,310	66,299	80	1.21	
	GM04	10,812,813	10,985,437	172,624	77	0.45	
	GM05	3,769,727	3,884,649	114,922	38	0.33	
	GM07	6,889,969	6,890,075	106	35	330.19	
	GM09	34,116,171	34,117,683	1,512	18	11.91	
	GM14	35,854,652	35,899,383	44,731	450	10.06	
	GM15	6,116,797	6,146,533	29,736	20	0.67	
	GM16	3,627,910	3,667,696	39,786	17	0.43	
$QS15524_{F2:F3}$	GM04	1,404,047	1,408,537	4,490	48	10.70	Minor (95^{th})
	GM05	208,889	294,855	85,966	46	0.54	
	GM15	49,385,092	49,442,075	56,983	53	0.93	
	GM15[1]	49,442,075	49,442,237	162	240	1,480.00	
	GM18	1,434,182	1,911,667	477,485	51	0.11	
	GM18[2]	1,911,667	1,935,386	23,719	40	1.69	
$QS15544_{RIL}$	GM01	6,517,814	6,579,997	62,183	11	0.18	Minor (95^{th})
	GM01	6,580,209	6,580,306	97	10	103.09	
	GM01	39,112,506	39,154,217	41,711	10	0.24	
	GM01	39,399,889	39,404,966	5,077	10	1.97	
	GM01	40,907,974	40,908,162	188	12	63.83	
	GM02	8,822,628	8,834,850	12,222	9	0.74	
	GM03	2,411,109	2,411,229	120	10	83.33	
	GM03	2,411,229	2,450,947	39,718	16	0.40	
	GM03	3,038,816	3,039,628	812	12	14.78	
	GM03	6,809,980	6,810,064	84	9	107.14	
	GM03	10,811,705	10,834,554	22,849	14	0.61	
	GM03	10,834,651	10,836,440	1,789	15	8.39	
	GM04	4,093,911	4,101,799	7,888	10	1.27	
	GM04	4,158,205	4,185,355	27,150	16	0.59	
	GM04	22,705,951	22,709,551	3,600	15	4.17	
	GM04	23,570,549	23,642,207	71,658	14	0.20	
	GM04	24,217,772	24,253,427	35,655	9	0.25	
	GM05	3,081,540	3,124,214	42,674	11	0.26	
	GM05	3,167,692	3,167,836	144	15	104.17	
	GM05	3,583,876	3,584,733	857	12	14.00	
	GM08	16,115,503	16,148,760	33,257	10	0.30	
	GM09	33,937,720	33,971,428	33,708	9	0.27	
	GM09	33,971,428	33,990,878	19,450	16	0.82	
	GM09	34,060,804	34,116,171	55,367	9	0.16	
	GM12	35,211,026	35,246,755	35,729	16	0.45	

[1] Interacts with *GmMDE04* and *GmPRR1a*.
[2] Interacts with *GmLHCA4a*.

Table 3.2 Major and minor hotspots in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Region	Method	Gene Name	Linkage group	Marker		LOD	PVE (%)	Additive effect	Dominance effect
				Left marker	Right marker				
F2_GM15:49,442,075-49,442,237[1]	ICIM	*GmPRR1a*	15	49,442,075	49,442,237	4.3	10.7	2.1	-57.0
	IM	*GmPRR1a*	15	49,442,075	49,442,237	4.3	10.7	2.1	-57.0
	GCIM	*GmMDE04*	15	49,442,075	49,442,237	17.1	15.0	0	101.2
	ICIM	*GmMDE04*	15	49,442,075	49,442,237	4.2	8.7	1.1	96.6
F2_GM18:1,911,667-1,935,386[2]	GCIM	*GmLHCA4a*	18	1,911,667	1,911,667	9.8	5.3	0	6,454.1
	ICIM	*GmLHCA4a*	18	1,911,667	1,935,386	4.5	8.0	1,204.7	8,147.8
RIL_GM04:17,227,512-20,251,662[3]	ICIM	*GmMDE04*	4	17,227,512	17,230,775	6.6	14.7	-97.6	N/A
	IM	*GmMDE04*	4	17,227,512	17,230,775	5.3	0.8	-89.9	N/A
	GCIM	*GmPRR1a*	4	17,534,130	17,914,073	9.7	19.1	-42.6	N/A
	IM	*GmPRR1a*	4	18,383,138	20,251,662	6.3	1.4	38.8	N/A
RIL_GM04:31,408,946-31,525,671	ICIM	*GmPRR1a*	4	31,408,946	31,525,671	10.1	19.9	42.6	N/A
	IM	*GmPRR1a*	4	31,408,946	31,525,671	6.6	1.5	39.4	N/A
RIL_GM13:37,289,785-38,620,690[4]	GCIM	*GmPRR1a*	13b	37,289,785	37,516,022	6.3	13.6	36.0	N/A
	ICIM	*GmPRR1a*	13b	37,790,482	37,795,923	5.2	9.5	-31.0	N/A
	IM	*GmPRR1a*	13b	38,027,686	38,620,690	5.8	1.4	-40.0	N/A

[1] The region identified here is F2_GM15:49,442,075-49,442,237 but has been merged to the F2_GM15:49,385,092-49,442,075 region to generate the F2_GM15:49,385,092-49,442,237 hotspot.
[2] The region identified here is the F2_GM18:1,911,667-1,935,386 but has been merged to the F2_GM18:1,434,182-1,911,667 region to generate the F2_GM18:1,434,182-1,935,386 hotspot.
[3] The RIL_GM04:17,227,512-20,251,662 region was obtained by merging the farthest left (GM04: 17,227,512) and right (GM04:20,251,662) markers.
[4] The RIL_GM13:37,289,785-38,620,690 region was obtained by merging the farthest left (GM04: 37,289,785) and right (GM04:38,620,690) markers.
N/A, not available.

Table 3.3 Expression quantitative trait loci for the *GmLHCA4a*, *GmPRR1a*, and *GmMDE04* genes.

Region/hotspot	Gene	SNP	REF allele	ALT allele	W82/ MA/OV[1]	Location	SIFT Consequence
F2_GM15:49,385,092-49,442,237	*Glyma.15G261300*	GM15:49,385,259	A	C	A/C/A	3'UTR	N/A
F2_GM15:49,385,092-49,442,237	*Glyma.15G263700*	GM15:49,734,668	A	G	A/G/A	CDS	Missense (Deleterious)
F2_GM15:49,385,092-49,442,237	*Glyma.15G263700*	GM15:49,736,375	T	G	T/G/T	5'UTR	N/A
F2_GM18:1,434,182-1,935,386	*Glyma.18G025600*	GM18:1,893,844	A	G	A/G/A	3'UTR	N/A
F2_GM18:1,434,182-1,935,386	*Glyma.18G025600*	GM18:1,894,078	C	A	C/A/C	3'UTR	N/A

[1] W82, 'William 82'; MA, 'Maple Arrow'; OV, 'OAC Vision'.
N/A, not available.

Table 3.4 Single nucleotide polymorphisms for the candidate transcription factors of the QS15524$_{F2:F3}$ population.

Region/hotspot	Gene	SNP	REF allele	ALT allele	W82/ MD/90[1]	Location	SIFT Consequence
RIL_GM04:17,227,512-20,251,662	*Glyma.04G135400*	GM04:19,964,773	A	T	A/T/A	CDS	Missense (Deleterious)
RIL_GM13:37,790,482-38,620,690	*Glyma.13G285400*	GM13:38,627,139	C	G	C/C/G	5'UTR	N/A
RIL_GM13:37,790,482-38,620,690	*Glyma.13G285400*	GM13:38,627,196	A	T	A/A/T	5'UTR	N/A
RIL_GM13:37,790,482-38,620,690	*Glyma.13G285400*	GM13:38,627,262	T	A	T/*/A[2]	5'UTR	N/A
RIL_GM13:37,790,482-38,620,690	*Glyma.13G285400*	GM13:38,627,374	T	A	T/T/A	5'UTR	N/A

[1] W82, 'William 82'; MD, 'AAC Mandor'; 90, '9004'.
[2] An asterisk (*) indicates a heterozygote genotype for that SNP.
N/A, not available.

Table 3.5 Single nucleotide polymorphisms for the candidate transcription factors of the QS15544$_{RIL}$ population.

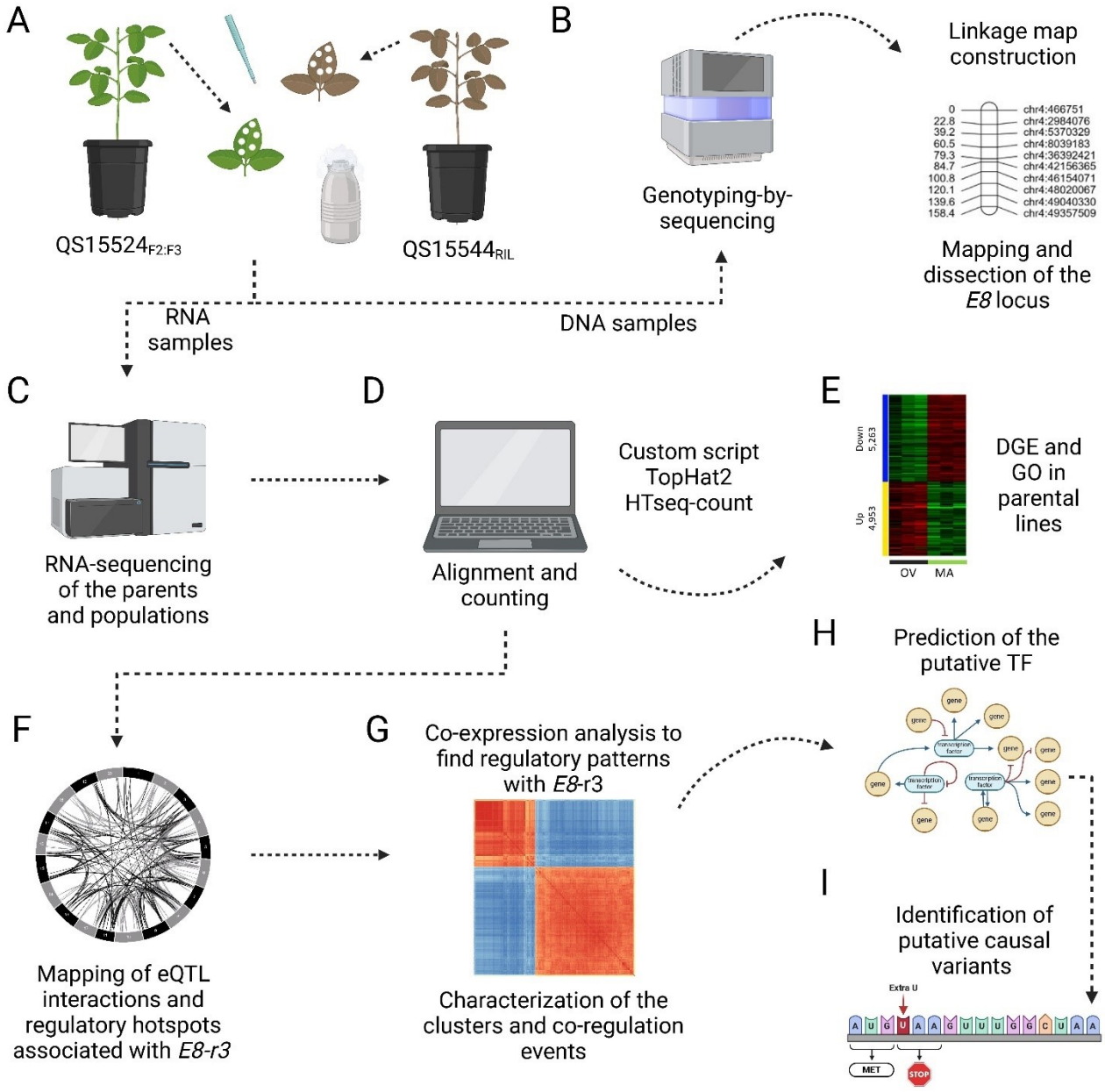

Figure 3.1 Experimental pipeline to identify candidate TFs involved in the regulation of *E8-r3* genes. (A) Leaf tissue collection of the middle leaflet at the V4 stage followed by DNA and RNA extraction. **(B)** Generation of the linkage map with the genotyping-by-sequencing datasets and mapping of the *E8-r3* candidate genes with ICIM and GCIM as detailed in Gélinas Bélanger *et al.* (2024a). **(C)** Isolation of mRNA and sequencing on the Illumina NovaSeq6000 platform. **(D)** Alignment and counting of the RNA-seq datasets using various bioinformatic scripts. **(E)** Identification of differentially expressed genes in the parental lines to validate the experimental conditions. **(F)** Mapping of the overlapping eQTL interactions using three algorithms (ICIM, IM,

and GCIM) and identification of the trans interactions associated with *GmLHCA4a*, *GmPRR1a*, and *GmMDE04* candidate genes found in the *E8-r3* region. **(G)** Building of CENs between the genes to identify homologous genes and closely associated candidate TFs. **(H)** Deeper investigations of the candidate TFs using TWCENs, gene expression, and FRSPD_GO annotations. **(I)** Identification of putative causal variants in the candidate TFs. Created with BioRender.com.

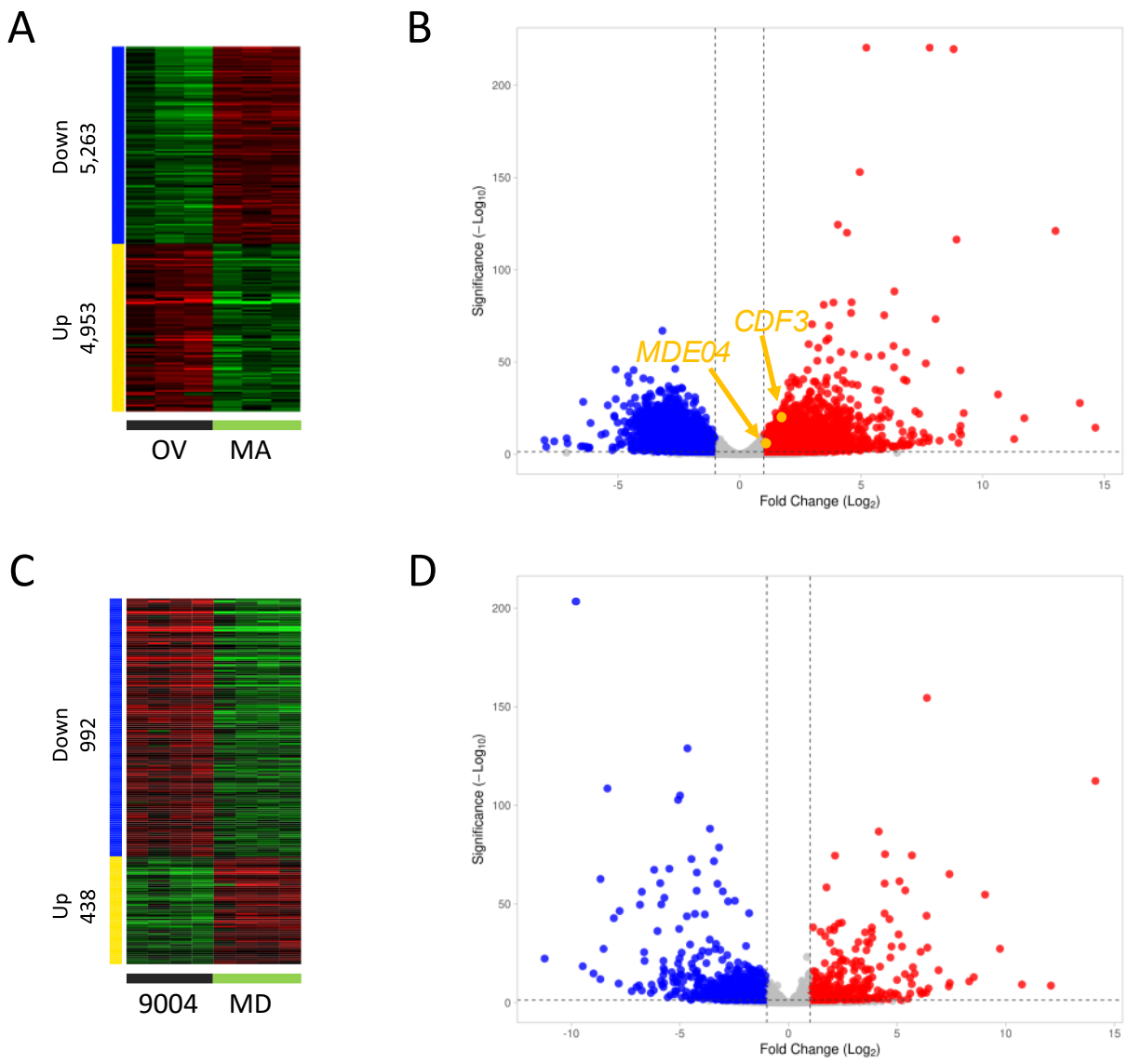

Figure 3.2 Differentially expressed candidate genes for the *E8-r3* locus in the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ parental lines. Heatmaps showing the number of DEGs in the QS15524$_{F2:F3}$ **(A)** and QS15544$_{RIL}$ **(C)** parental lines. Volcano plots showing the differentially expressed candidate genes in the QS15524$_{F2:F3}$ **(B)** and QS15544$_{RIL}$ **(D)** parental lines. Two candidate *E8-r3* genes (*GmCDF3* and *GmMDE04*) have been found to be upregulated in the QS15524$_{F2:F3}$ parents, whereas none of the four candidates were found to be differentially expressed in the QS15544$_{RIL}$ parents.

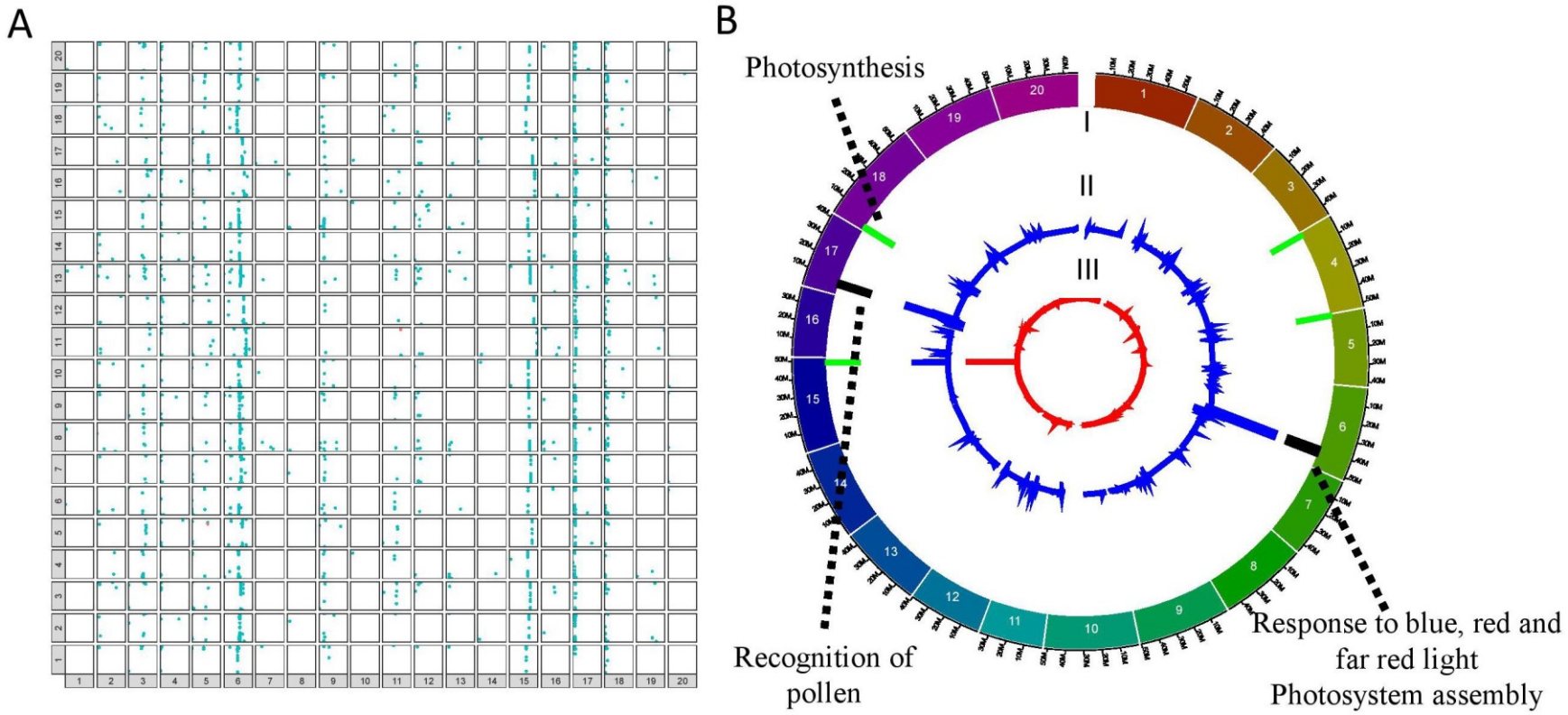

Figure 3.3 Mapping of the eQTL interactions and regulatory major hotspots using the combinatorial approach in the QS15524$_{F2:F3}$ population. (A) Identification of the eQTL interactions found with at least two mapping algorithms in the QS15524$_{F2:F3}$ population. The X-axis represents regulating regions, whereas the Y-axis represents the locations of the target genes. *Cis* interactions and *trans* interactions are respectively illustrated as the orange and light blue dots. **(B)** Mapping of the regulatory hotspots. Level I, locations of the hotspots. Major and minor hotspots are respectively indicated using black and green rectangles. Level II, number of eQTL interactions per marker. Level III, eQTL density per marker. The dotted lines indicate the FRSPD_GO functions significantly associated with the hotspots (Bonferroni p-value, 0.01).

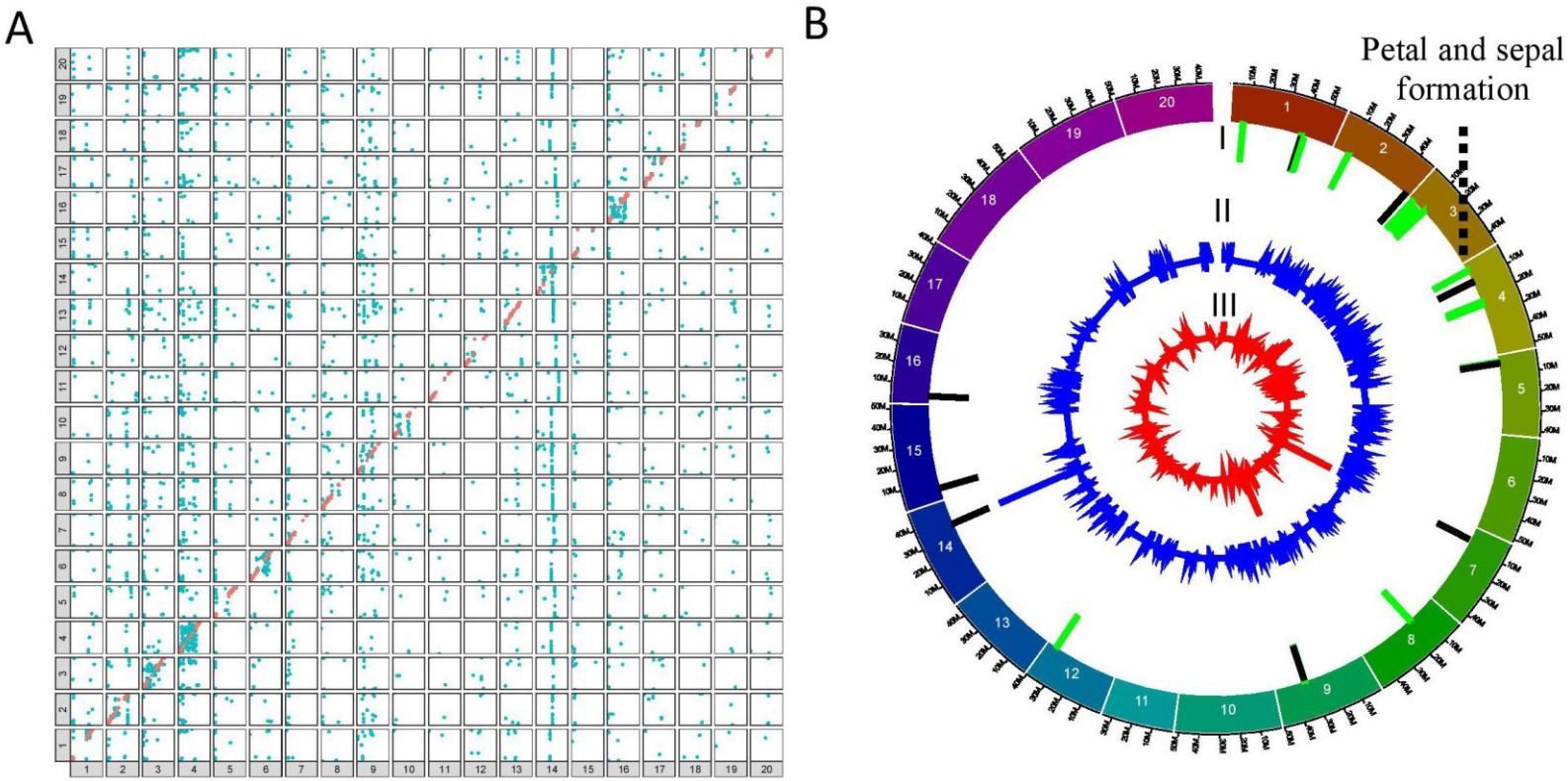

Figure 3.4 Mapping of the eQTL interactions and regulatory major hotspots using the combinatorial approach in the QS15544$_{RIL}$ population. (A) Identification of the eQTL interactions found with at least two mapping algorithms in the QS15544$_{RIL}$ population. The X-axis represents regulating regions, whereas the Y-axis represents the locations of the target genes. *Cis* interactions and *trans* interactions are respectively illustrated as the orange and light blue dots. **(B)** Mapping of the regulatory hotspots. Level I, locations of the hotspots. Major and minor hotspots are respectively indicated using black and green rectangles. Level II, number of eQTL interactions per marker. Level III, eQTL density per marker. The dotted lines indicate the FRSPD_GO functions significantly associated with the hotspots (Bonferroni p-value, 0.01).

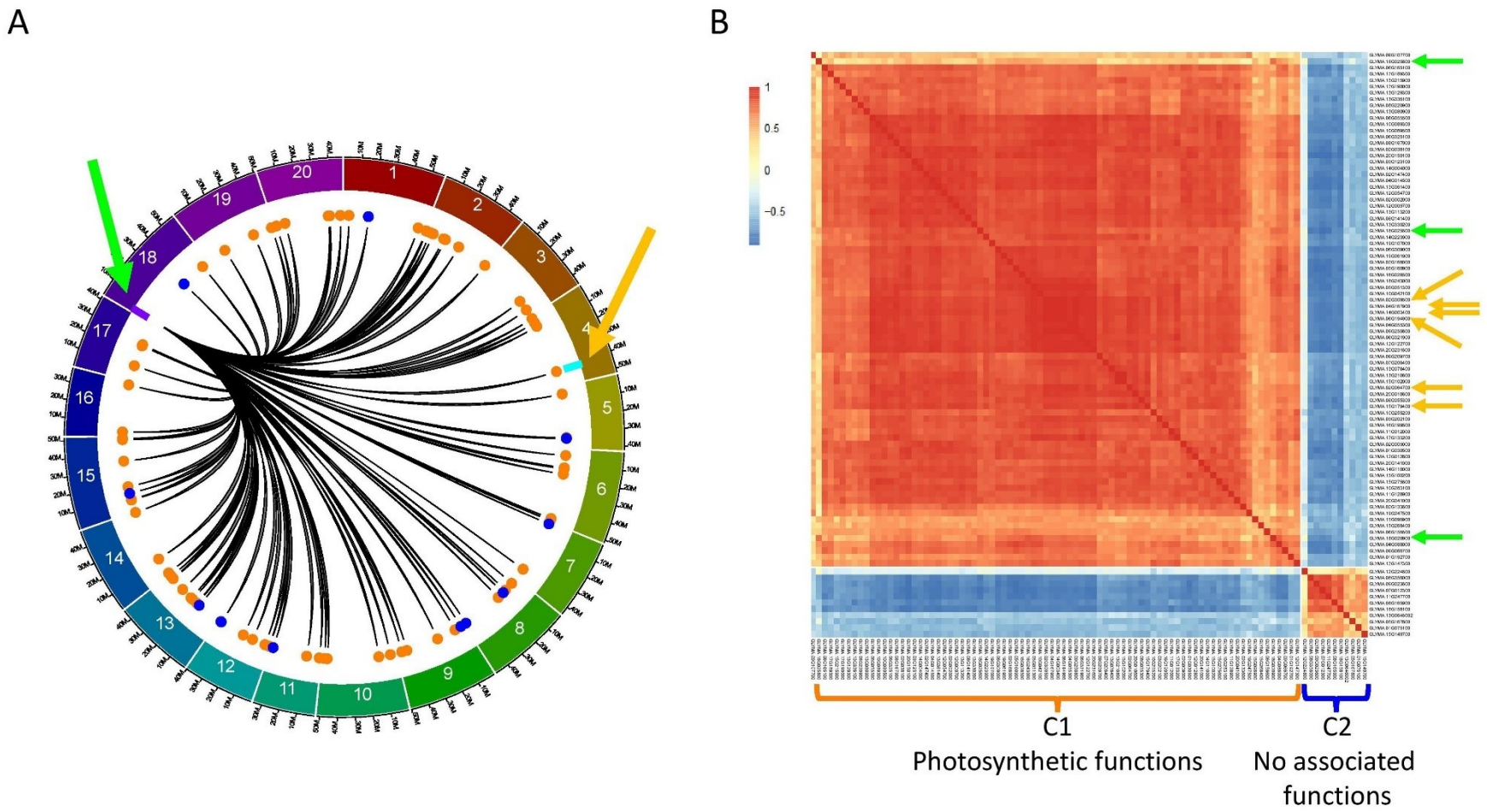

Figure 3.5 Characterization of the F2_GM18:1,434,182-1,935,386 hotspot and its interaction with *GmLHCA4a* in the QS15524$_{F2:F3}$ population. (A) Identification of the 91 *trans* interactions (90 genes) associated with the F2_GM18:1,434,182-1,935,386 hotspot. Colored dots represent the C1 (orange color; 79 genes) and C2 (royal blue color; 11 genes) clusters depicted in panel (B). Orange arrow, location of the candidate gene *GmLHCA4a* and the *E8-r3* locus (light blue rectangle). Green arrow, location of the three candidate TFs (*Glyma.18G020900*, *Glyma.18G025600*, and *Glyma.18G025800*) and the F2_GM18:1,434,182-1,935,386 hotspot (purple

rectangle). **(B)** Co-expression network with the 90 target genes and three candidate TFs. Orange arrows, location of the six *LHCA* genes (*GmLHCA4a*, *Glyma.02G064700*/*GmLHCA1*, *Glyma.02G309500*/*GmLHCA3a*, *Glyma.06G194900*/*GmLHCA4b*, *Glyma.14G003400*/*GmLHCA3b*, and *Glyma.15G179400*/*GmLHCA6*). Green arrows, location of the three candidate TFs. The C1 cluster is significantly associated with photosynthetic functions, whereas C2 is not enriched with any FRSPD_GO terms.

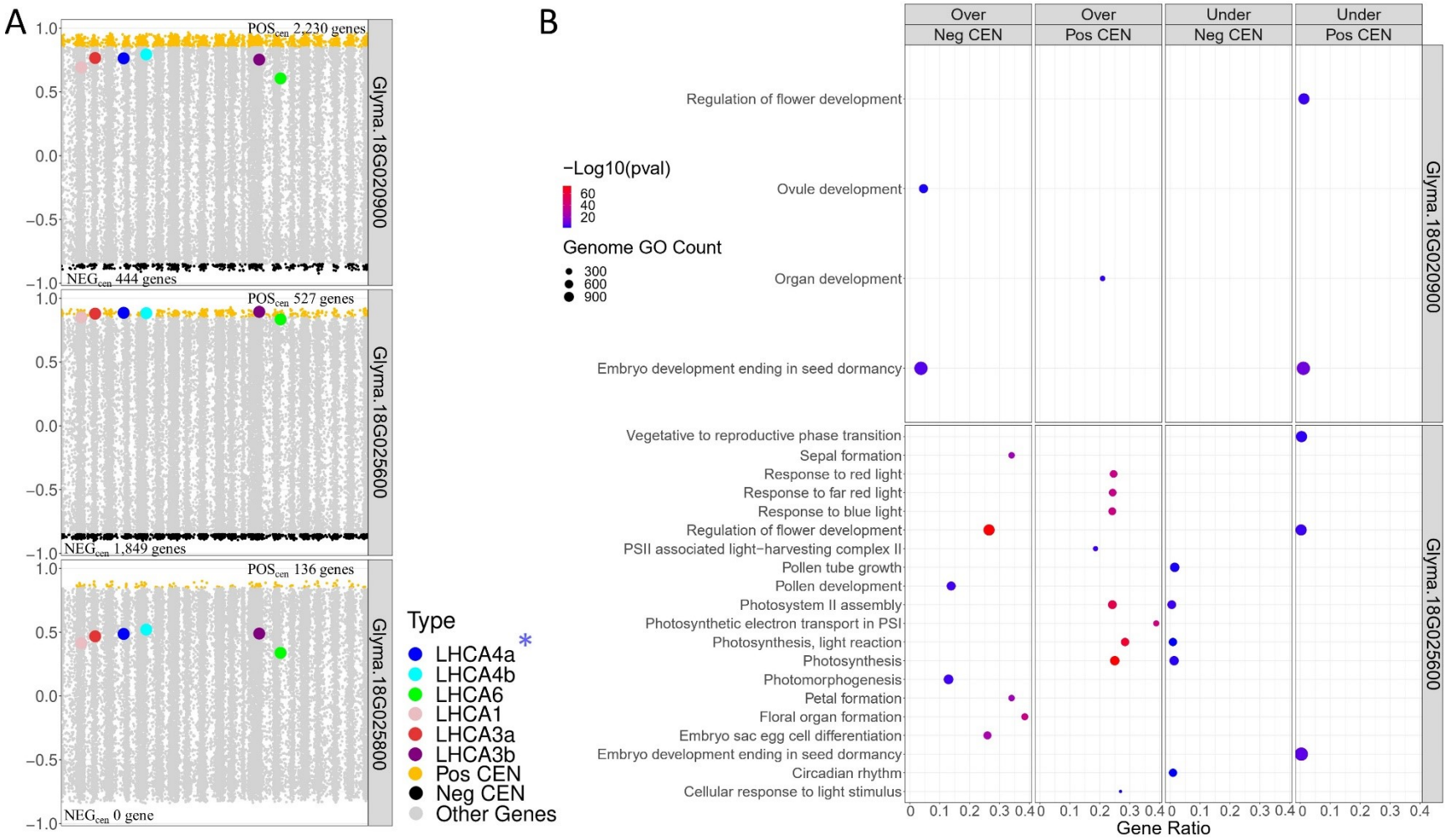

Figure 3.6 Transcriptome-wide co-expression network for the three candidate TFs of the F2_GM18_1,434,182-1,935,386 hotspot. **(A)** Positive and negative TWCENs for the three candidate TFs using PCC thresholds of ≥ 0.85 (POS_{TWCEN}) and ≤ -0.85 (NEG_{TWCEN}). As shown in the panel, *Glyma.18G020900* exhibits the largest POS_{TWCEN} (2,230 genes) followed by *Glyma.18G025600* (527 genes), and *Glyma.18G025800* (136 genes). For the NEG_{TWCEN}, *Glyma.18G025600* (1,849 genes) displays the largest network, whereas the

network of *Glyma.18G020900* is smaller (444 genes). No gene was found for the NEG_{TWCEN} of *Glyma.18G025800*. The highest level of co-expression between the *LHCA* homologs and a candidate TF was achieved with *Glyma.18G025600* with a mean PCC of 0.87 for the six homologs. In comparison, the mean PCC of *Glyma.18G020900* and *Glyma.18G025800* for these six homologs were 0.73 and 0.45, respectively. *GmLHCA4a*, the candidate target gene for the *E8-r3* locus, is highlighted with an asterisk. **(B)** Functional annotation of the POS_{TWCEN} and NEG_{TWCEN} of each candidate TF. The POS_{TWCEN} and NEG_{TWCEN} of the *Glyma.18G025600* gene were significantly enriched with a large number of FRSPD genes associated with photosynthetic properties such as light response. Only gene annotations that are either over-represented (i.e., "Over" facet) or under-represented (i.e., "Under" facet) are displayed in the figure. Non-FRSPD annotations were not displayed for visualization purposes, but are available in Appendix 2.21.

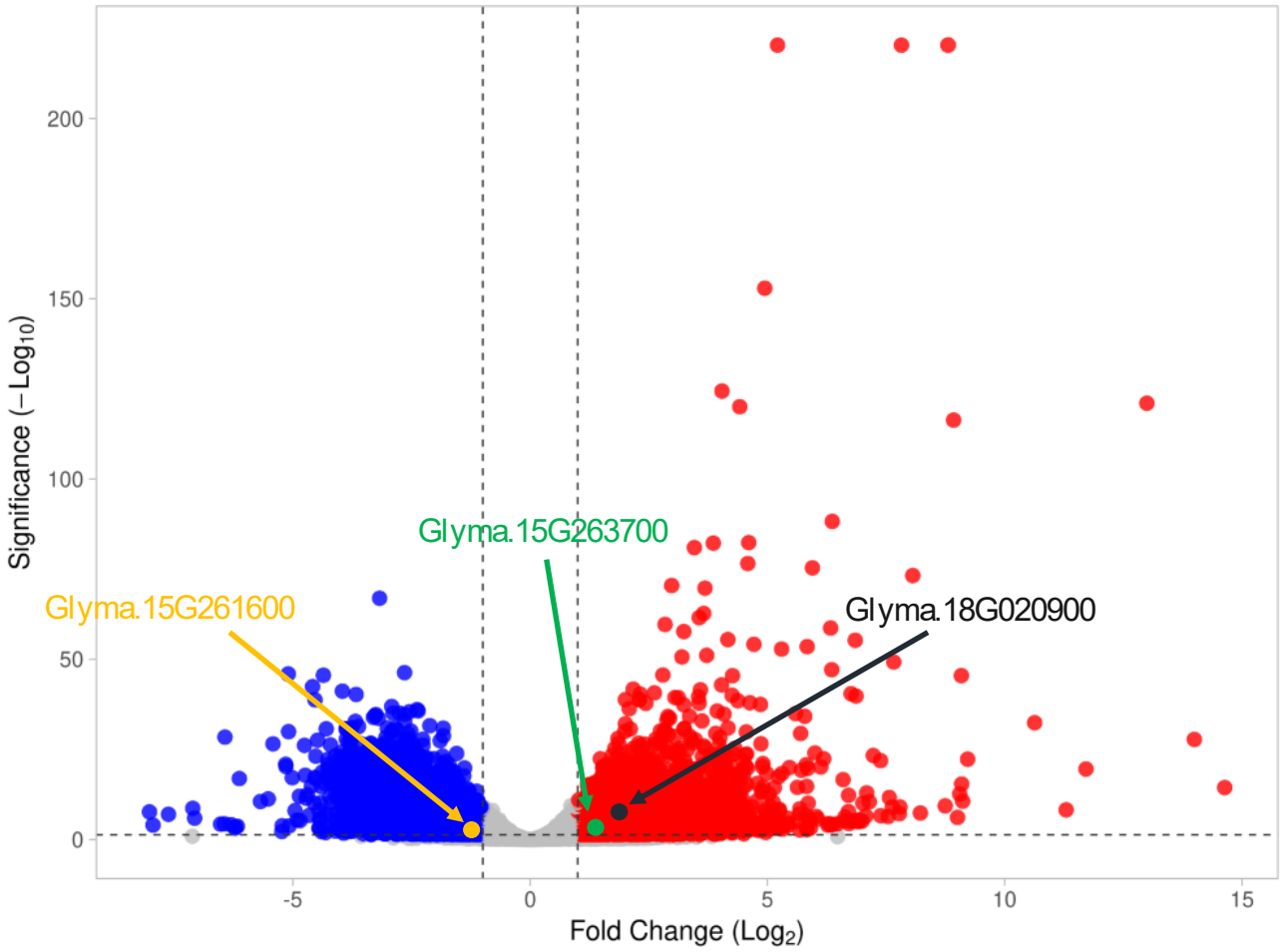

Figure 3.7 Differentially expressed candidate transcription factors in the QS15524$_{F2:F3}$ parents. Three candidate TFs (*Glyma.15G263700*, *Glyma.18G020900*, and *Glyma.18G025800*), have been found to be differentially expressed in the QS15524$_{F2:F3}$ parental lines.

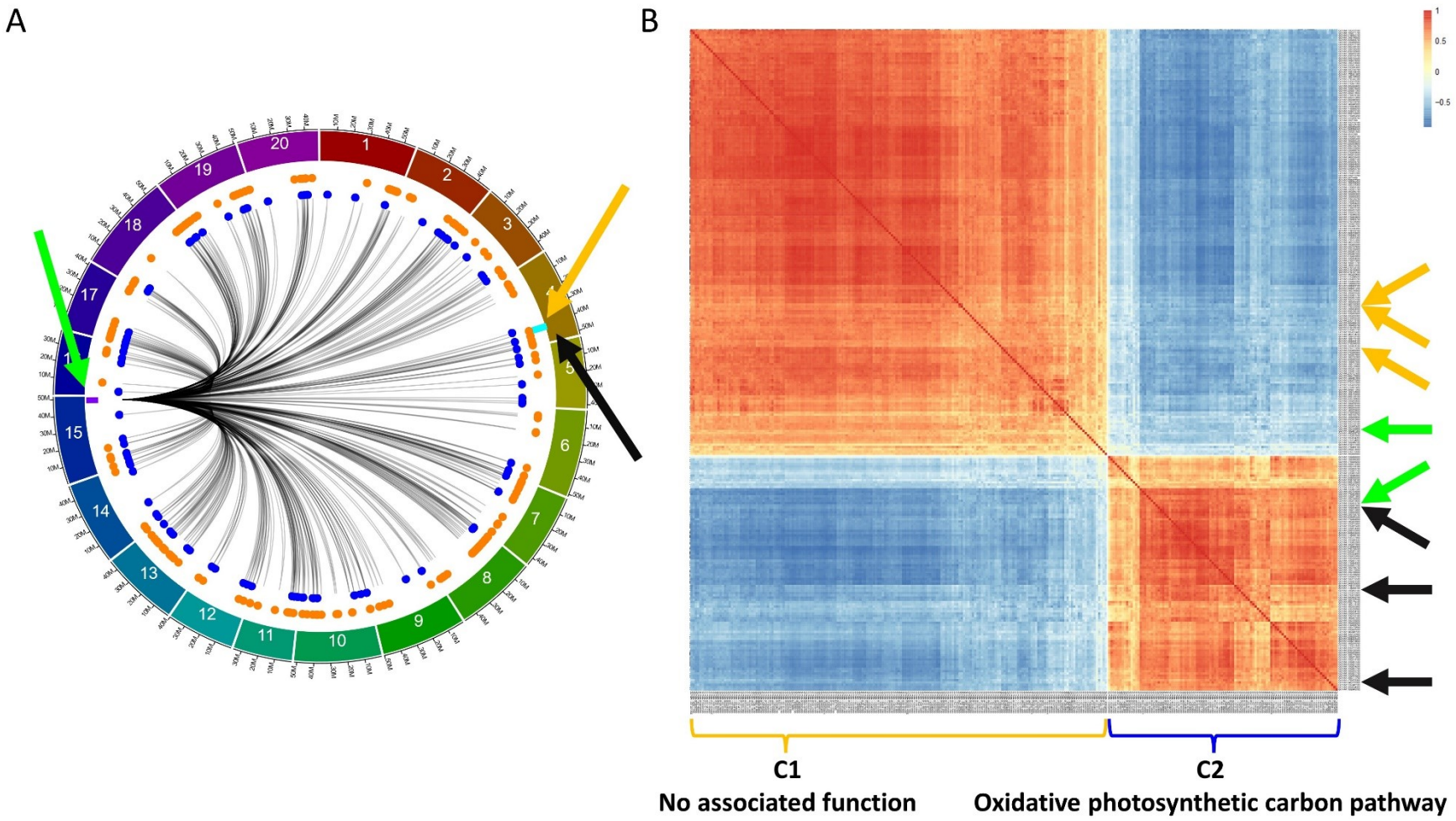

Figure 3.8 Characterization of the F2_GM15:49,385,092-49,442,237 hotspot and its interaction with *GmPRR1a*, *GmMDE04,* and their homologs in the QS15524$_{F2:F3}$ population. (A) Identification of 293 *trans* interactions (285 genes) with the F2_GM15:49,385,092-49,442,237 hotspot. Green arrow, location of the two candidate TFs (*Glyma.15G261300* and *Glyma.15G263700*) and F2_GM15:49,385,092-49,442,237 hotspot (purple rectangle). The locations of *GmPRR1a*, *GmMDE04*, and the *E8-r3* locus are

respectively indicated by the orange arrow, black arrow, and light blue rectangle. Orange and royal blue dots respectively represent the genes located in the C1 and C2 clusters. **(B)** Co-expression network with the 285 genes and the two candidate TFs. Orange arrows, location of *GmPRR1a*, and two *PRR* homologs (*GmPRR1d* and *GmARR2*). Black arrows, location of *GmMDE04*, and two *MDE* homologs (*GmMDE13* and *GmMDE19*). Green arrows, location of the candidate TFs. The target genes and candidate TF found in the C1 cluster are indicated with the orange bracket, whereas those found in the C2 cluster are indicated with the royal blue bracket. Based on a functional enrichment analysis, the C2 cluster is associated with the term 'oxidative photosynthetic carbon pathway', whereas C1 is not associated with any terms.

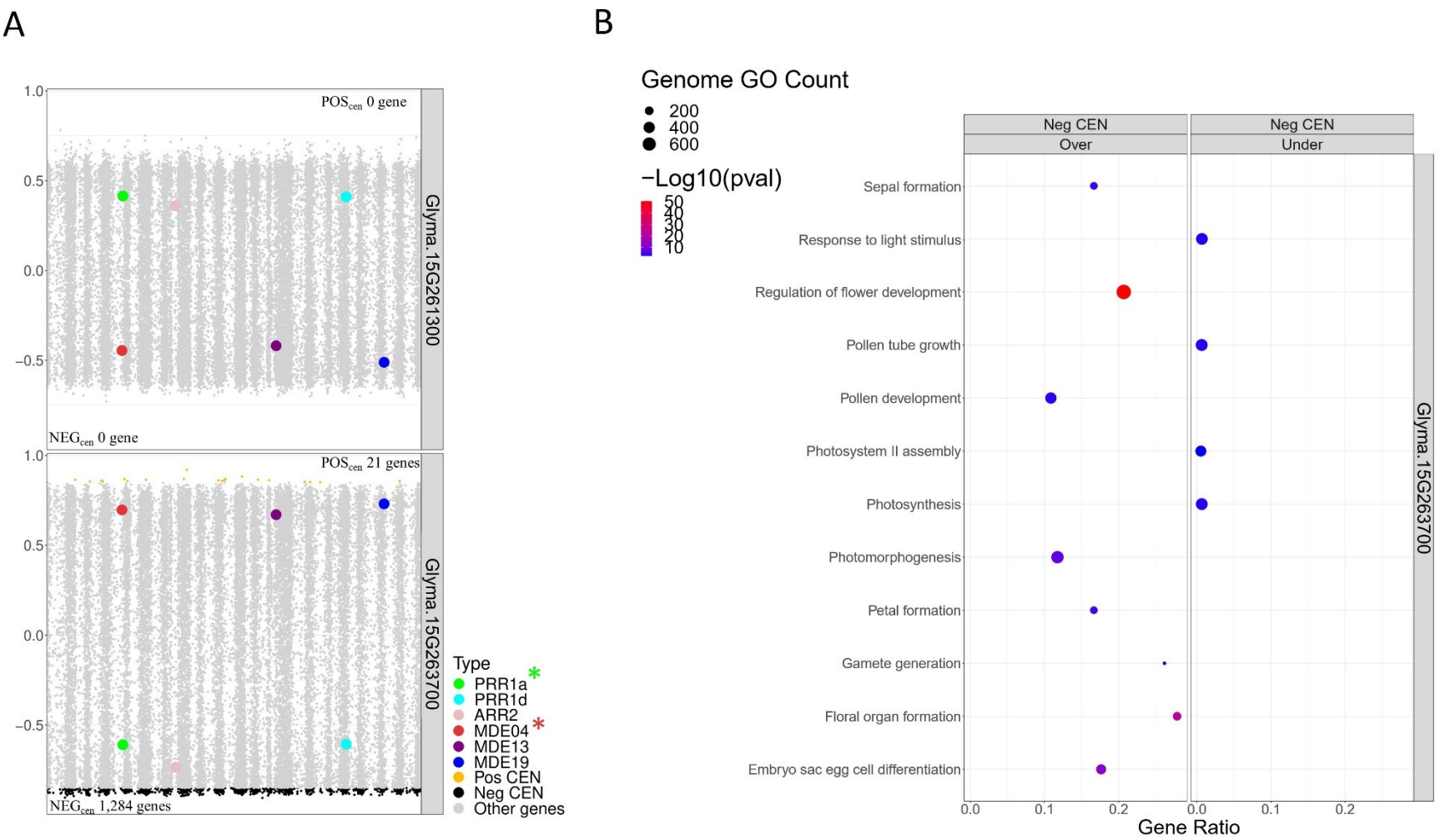

Figure 3.9 Transcriptome-wide co-expression network for the candidate TFs of the F2_GM15:49,385,092-49,442,237 hotspot. **(A)** Positive and negative TWCENs for the two candidate TFs using PCC thresholds of ≥ 0.85 (POS_{TWCEN}) and ≤ -0.85 (NEG_{TWCEN}). As shown in the panel, only *Glyma.15G263700* displayed a large NEG_{TWCEN} (1,284 genes). The NEG_{TWCEN} of *Glyma.15G261300* (0 genes) and the POS_{TWCEN} of both candidates (*Glyma.15G263700*, 21 genes; *Glyma.15G261300*, 0 genes) were small. The candidate target

genes *GmPRR1a* and *GmMDE04* are highlighted with asterisks. **(B)** Functional annotation of the NEG_{TWCEN} of *Glyma.15G263700*. This NEG_{TWCEN} is strongly enriched with terms associated with flowering and response to light functions. Only gene annotations that are either over-represented (i.e., "Over" facet) or under-represented (i.e., "Under" facet) are displayed in the Figure. Non-FRSPD annotations were not displayed for visualization purposes, but are available in Appendix 2.21.

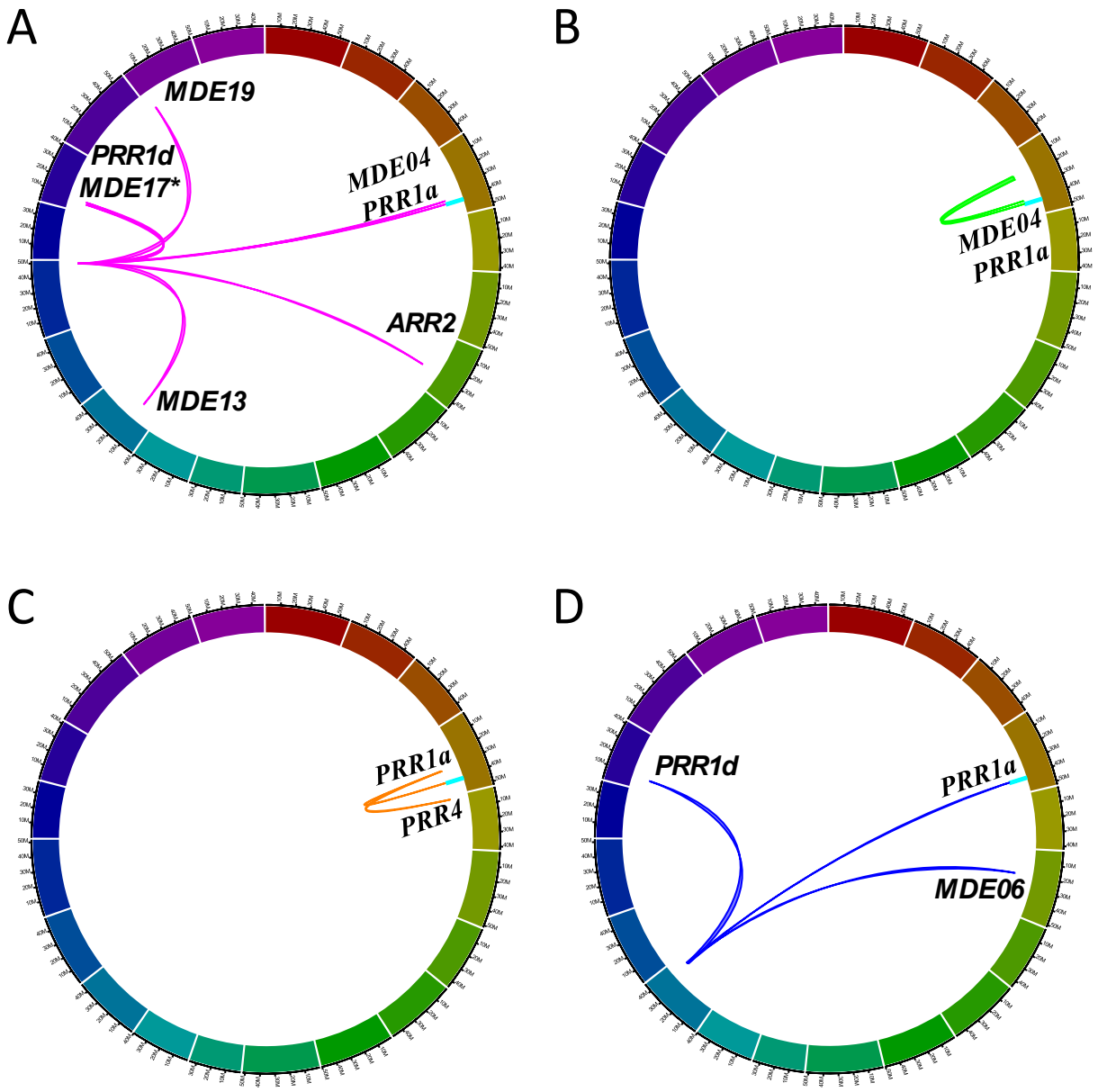

Figure 3.10 Minor regions regulating the *PRR* and *MDE* homologs in the QS1544$_{RIL}$ population. Circos plots illustrating the interactions between the F2_GM15:49,385,092-49,442,237 **(A)**, RIL_GM04:17,227,512-20,251,662 **(B)**, RIL_GM04:31,389,583-31,525,671 **(C)**, and RIL_GM13:37,289,785-38,620,690 **(D)** regions and the different *PRR* and *MDE* homologs, including the candidate target genes *GmPRR1a* and *GmMDE04* that are located in the *E8-r3* locus (light blue rectangle). The F2_GM15:49,385,092-49,442,237 is regulating an additional *MDE* homolog, *GmMDE17*. The asterisk denotes that this additional homolog has been mapped with only one algorithm (ICIM) instead of two like all the other interactions.

4.10 Connecting text

Chapter 4 demonstrates that multiple gene regulatory networks regulate the level of expression of hundreds of interactions in *trans* at the V4 stage in soybean, just before the initiation of floral meristematic transition. As revealed in this chapter, a large number of interactions is regulated by hotspots which are specific genomic locations regulating the expression of several genes. In addition, this chapter shows that multiple regions regulate three genes (*GmLHCA4a*, *GmPRR1a*, and *GmMDE04*) that were proposed as candidates for *E8-r3*. Our investigations led to the discovery of specific *trans* interactions from hotspots and minor regions that regulate numerous homologs of these candidate genes. Two transcription factors, *ALTERED PHLOEM DEVELOPMENT* (*Glyma.15G263700*) and *DOMAIN-CONTAINING PROTEIN 21* (*Glyma.18G025600*), were identified as potential candidates for regulating the F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386 hotspots which are respectively associated with *GmPRR1a*/ *GmMDE04* and *GmLHCA4a*. Hence, these candidates are good targets for breeding to influence the gene regulatory networks guiding reproductive traits. Still, breeding remains a challenging multifaceted act due to the parallel selection of crucial key traits, including reproductive but also seed quality traits. Selection for marginal areas, such as the MGs 000 and 00 cultivation areas, faces additional roadblocks due to the limited germplasm available. Consequently, sources of variation for these maturity groups are scarce and must be properly understood. Chapter 5 investigates novel sources of variation for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations for six different quality traits. Overall, the following study demonstrates that three major pod-filling and maturity-regulating loci identified in Chapter 3 (*E8-r1*, *E8-r2,* and GM16:5,680,173-5,730,237) are in close linkage with loci that regulate traits such as seed weight and oleic acid.

5. Identification of Quantitative Trait Loci Associated with Seed Quality Traits in Two Early-Maturing Soybean Populations

Jérôme Gélinas Bélanger[1,2], Tanya Rose Copley[1], Valerio Hoyos-Villegas[2] & Louise O'Donoughue[1]

[1]CÉROM, Centre de recherche sur les grains Inc., St-Mathieu-de-Beloeil, Québec, Canada

[2]Department of Plant Science, McGill University, Montréal, Québec, Canada

* Correspondence:

Dr. Louise O'Donoughue

louise.odonoughue@cerom.qc.ca

Reproduced from the Canadian Journal of Plant Science

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., O'Donoughue, L. (2024). Identification of Quantitative Trait Loci Associated with Seed Quality Traits in Two Early-Maturing Soybean Populations. ***Can. J. Plant Sci.*** 0, null. doi: 10.1139/cjps-2024-0049

Minor modifications were made to conform to the McGill University thesis guidelines.

5.1 Abstract

English version

In Canada, soybean (*Glycine max* (L.) Merr.) is primarily cultivated in three provinces (Ontario, Quebec, and Manitoba). Canadian breeders want to expand the current cultivation range to more northern agro-environments by developing early-maturing elite lines while maintaining good seed quality traits. To examine quantitative trait loci involved in 100-seed weight and seed protein, oil, and fatty acid (oleic, linolenic, and linoleic acids) contents, we generated an early-maturing recombinant inbred line population ($QS15544_{RIL}$) and an F2:F3 ($QS15524_{F2:F3}$) population adapted to cultivation zones MGs 00 and 000, and phenotyped them for 3 years and 1 year, respectively. Using two mapping algorithms (Inclusive composite interval mapping and Genome-wide composite interval mapping), we identified a total of 12 major regions that were either associated with $QS15544_{RIL}$ (five loci), $QS15524_{F2:F3}$ (four loci), or both (three loci) populations. Of the 12 identified regions, three (RIL_GM12, RIL_GM16, and F2_GM04.2) were not previously identified and might, respectively, serve as novel sources of regulation for oil content, seed weight, and oleic acid. For the RIL_GM05 locus, we identified two novel variants in *Glyma.05G244100*/*MOTHER OF FT AND TFL1*, a gene with a confirmed role in the regulation of oleic and linoleic acid contents. Two of the major loci (RIL_GM04 and RIL_GM16) associated with the 100-seed weight trait and one locus (F2_GM04.2) associated with oleic acid were found to be overlapping three loci (*E8-r1*, GM16:5,680,173–5,730,237, and *E8-r2*) involved in early-maturity and/or shorter pod-filling that were previously identified by our group, suggesting possible breeding bottlenecks due to linkage drag or pleiotropic effects.

Version française

Au Canada, le soja (*Glycine max* (L.) Merr.) est principalement cultivé dans trois provinces (Ontario, Québec et Manitoba). Les sélectionneurs canadiens veulent augmenter l'étendue de la culture du soja en créant des cultivars hâtifs adaptés aux agroenvironnements nordiques tout en maintenant une bonne qualité du grain. Pour identifier des loci de traits quantitatifs (QTL) associés avec le poids 100 grains, la teneur en protéines, le taux d'huile et la teneur en acides gras (acides oléique, linolénique et linoléique), nous avons généré deux populations, une consanguine recombinante ($QS15544_{RIL}$) et une F2:F3 ($QS15524_{F2:F3}$), adaptées aux zones de culture MGs 00 et 000, et avons phenotypé ces populations pour respectivement trois et une années. En utilisant

deux algorithmes de cartographie (Inclusive composite interval mapping et Genome-wide composite interval mapping), nous avons identifié un total de 12 régions majeures étant associées soit à la population $QS15544_{RIL}$ (cinq loci), la population $QS15524_{F2:F3}$ (quatre loci) ou les deux (trois loci) populations. Des 12 régions identifiées, trois (RIL_GM12, RIL_GM16 et F2_GM04.2) n'avaient pas été précédemment découvertes et se présentent ainsi comme des nouvelles sources de régulation pour la teneur en huile, le poids 100 grains et la teneur en acide oléique. Pour le locus RIL_GM05, nous avons identifié deux nouveaux variants pour le gène *Glyma.05G244100*/*MOTHER OF FT AND TFL1*, connu pour réguler les teneurs en acide oléique et linoléique. Trois autres loci (RIL_GM04, RIL_GM16 et F2_GM04.2) ont été identifiés à proximité ou superposant trois loci (*E8-r1*, GM16:5,680,173–5,730,237 et *E8-r2*) liés à la maturité hâtive et/ou la durée du remplissage des gousses précédemment découverts par notre groupe, suggérant ainsi de possibles limitations de sélection dues à la pléiotropie ou une forte liaison génétique.

5.2 Introduction

Soybean (*Glycine max* (L.) Merr.) is the most important leguminous oilseed crop and significantly contributes to maintaining food security on a global scale due to its high levels of unsaturated fat, high protein content, and high protein quality (Pagano and Miransari, 2016). Furthermore, its ability to fix nitrogen contributes to reducing the overreliance on chemical fertilizers and the pollution associated with their use (Li *et al.*, 2020). The bulk of soybean production is currently in Brazil (37%), the United States of America (28%), and Argentina (16%), whereas China (5%), Paraguay (3%), India (3%), and Canada (2%) are considered minor producers (The American Soybean Association, 2023). On a global scale, soybean is a major contributor to the world's oilseed output (59%) and plant protein meal consumption (70%) (The American Soybean Association, 2023). The bulk of food-grade Canadian soybean production is found in Quebec and Ontario, whereas Prairies' production is mainly destined for the animal-feed market due to a lower protein content (Isaacs, 2020). Soybean seeds are composed of five main components: (i) protein; (ii) oil; (iii) carbohydrates; (iv) ash; and (v) water (i.e., moisture content) (Singer *et al.*, 2023). Protein content is a crucial feature of soybean seed quality components and export potential due to its importance in the diets of livestock, with soybean meal, and humans, with food-grade soybeans (The American Soybean Association, 2023). Soybean proteins are highly

valuable in human and livestock diets because they contain all nine essential amino acids (Qin *et al.*, 2022). The amino acid profile of soybean seeds is considered well-balanced for all essential amino acids, except for sulfur-containing ones such as methionine (Qin *et al.*, 2022). The high content and excellent composition of soybean's oil are also important features contributing to its rising role as a leader in seed oil production. Of the five major fatty acids found in soybean seeds, linoleic acid (C18:2) corresponds to the bulk (56.3% of the total oil content) of the oil composition, whereas oleic (C18:1; 18.7%), palmitic (C16:0; 10.4%), α-linolenic (C18:3; 9.2%), and stearic (C18:0; 3.7%) are found in lower quantities (Canadian Grain Commission, 2022). The fatty acid profile of soybean seeds is a major determinant in the commercialization of soybean seeds due to its impact on the quality and functionality of soybean. An increased soybean oil's shelf life is associated with a reduction in the content of polyunsaturated fatty acids such as linolenic and linoleic acid due to enhanced oxidative stability (Clemente and Cahoon, 2009). Cultivars with high oleic acid (>70%) content exhibit enhanced oxidative stability that contributes to a longer shelf life (Bilyeu *et al.*, 2018). In Canada, mean oil and protein contents for soybean samples correspond to 21.7% and 38.6% on a dry basis, respectively (Canadian Grain Commission, 2022). In the country, the negative correlation between protein content and yield is considered a major roadblock to the expansion of soybean cultivation, with a decrease in yield between 45.3 kg ha^{-1} (Eastern Canada) to 78.4 kg ha^{-1} (Western Prairie) per 1% increase in the protein content (Cober *et al.*, 2023). Seed weight, typically expressed as 100-seed weight, is a critical component of yield and food quality in soybean (Liu *et al.*, 2018; Qi *et al.*, 2020). As such, this trait remains one of the most sought-after avenues to increase yield in newly developed cultivars.

To supply the growing demand for soybean-related food, feed, and industrial goods, global soybean production needs to increase (Foyer *et al.*, 2019). One approach to partly solve this problem is to identify key regions controlling seed quality traits in short-season cultivars that exhibit good potential for expansion beyond their actual northern growing limits. However, developing commercial cultivars with good quality features (e.g., protein, oil, and amino and fatty acid profiles) for such extreme conditions is challenging due to the limited germplasm available for MG00 and MG000 cultivation areas (Iquira *et al.*, 2010), and increasing problems with pests and diseases such as the soybean cyst nematode (SCN) (Tylka and Marett, 2021). Similarly, it is common knowledge that linkage drag between desirable and undesirable alleles, including those that affect yield, is a significant obstacle for breeders, especially when the gene pool available is

limited. In the literature, some studies using early-maturing genetic material have identified linkage drag between loci of interest (e.g., SCN resistance) and quality traits (e.g., decrease in protein concentration, reduction in 100-seed weight, and increase in seed oil content), thus illustrating that bottlenecks can arise from these events (St-Amour *et al.*, 2020). For breeders, knowledge of these possible linkage drag associations is important to limit ineffective selection efforts, particularly in a context of limited germplasm (Torkamaneh *et al.*, 2021).

In a previous research article, we identified novel shorter pod-filling and early maturity-related quantitative trait loci (QTL) regions on chromosomes GM04 and GM08 in two biparental soybean populations, named QS15544$_{RIL}$ and QS15524$_{F2:F3}$ (Gélinas Bélanger *et al.*, 2024). On chromosome GM04, these regions included three major loci for pod-filling and maturity traits: *E8-r1* (GM04:16,974,874-17,152,230), *E8-r2* (GM04:35,168,111-37,664,017), and *E8-r3 (*GM04:41,808,599-42,376,237 (GM04:41,808,599-42,376,237). Moreover, several additional regions for the same traits were identified on GM04, GM07, and GM16 in the QS15544$_{RIL}$ population, including the major GM16:5,680,173-5,730,237 region which regulates the number of days to maturity. Both of these mapping populations were developed using Canadian germplasm that was fixed for their alleles at the common early maturity loci *E1* to *E4* alleles, and bred for growing areas suitable for MG00 and MG000 cultivars. Through our experiments, we also observed that both populations exhibit a high level of variation for several key seed quality traits and we were interested in determining if selecting for early maturity would influence these traits. In this study, we aimed to identify loci involved in the regulation of six seed quality traits in the QS15544$_{RIL}$ and QS15524$_{F2:F3}$ biparental populations. The main objectives of this study were to (i) uncover QTL for six seed quality traits in two biparental populations using two mapping algorithms (Inclusive composite interval mapping and Genome-wide composite interval mapping); (ii) identify loci that could be linked to regions associated with early reproductive traits; and (iii) identify candidate single nucleotide polymorphisms and genes regulating these quality traits.

5.3 Materials and methods

5.3.1 Generation of the Mapping Populations

The mapping populations were generated as detailed in Gélinas Bélanger *et al.* (2024). Briefly, the full mapping population of 162 $F_{5:8}$ individuals of the QS15544 population (recombinant inbred lines; herein named QS15544$_{RIL}$) was derived from a single biparental cross

between '9004' ♀ (PI 592534; MG000, earlier-maturing accession) × 'AAC Mandor' ♂ (MG00, later-maturing accession). The full mapping population of 176 $F_{2:3}$ individuals of the QS15524 population (herein named QS15524$_{F2:F3}$) was derived from a single biparental cross between 'OAC Vision' ♀ (PI 567787; MG000, earlier-maturing accession) × 'Maple Arrow' ♂ (PI 548593; MG00, later-maturing accession). Both populations used in this study were developed at the Centre de recherche sur les grains (CÉROM) Inc. in Saint-Mathieu-de-Beloeil (QC, Canada).

5.3.2 Growing Conditions, Phenotyping, and Statistical Analyses

Both populations were grown at Saint-Mathieu-de-Beloeil (QC, Canada) using a Modified Augmented Design with Method 3 (adjustment by regression) defined as the following: $Y'_{ij(k)} = Y_{ij(k)} - b(X_{ij(A)} - \bar{X}_A)$ (Lin and Poushinsky, 1985; Schaalje *et al.*, 1987). The three variables of this model were defined as: (i) $Y_{ij(k)}$, as the observed value of the *k*th test line of the whole plot of the *i*th row and *j*th column; (ii) $X_{ij(A)}$, as the observed value of the control plot in the *ij*th whole plot; and (iii) *b*, as the regression coefficient of the mean of two control subplots (Lin and Poushinsky, 1985; Schaalje *et al.*, 1987). The F_6:F_8 generations of the QS15544$_{RIL}$ population were grown over three summers: summers 2020 (one-row plots), 2021 (two-row plots) and 2022 (two-row plots). The F_3 generation of the QS15524$_{F2:F3}$ population was grown in single-row plots in the summer of 2021. Six seed quality traits were phenotyped for both populations and included: (i) 100-seed weight; (ii) oil; (iii) oleic acid; (iv) linolenic acid; (v) linoleic acid; and (vi) protein content. All of the seed quality traits were phenotyped using Near Infrared Reflectance (NIR) with a Perten DA7250 Analyzer (PerkinElmer, Wellesley, MA, USA), except for the 100-seed weight phenotype which was analyzed using a microscale. In QS15544$_{RIL}$, the three years of data were then averaged for QTL mapping. Statistical analysis for the Modified Augmented Design was performed in Agrobase Generation II® (Agronomix Software Inc., 2009). Pearson correlations between the phenotypes, which included all six aforementioned phenotypes and three reproductive traits (days to flowering, pod-filling, and maturity) evaluated in Gélinas Bélanger *et al.* (2024), were calculated using R version 4.0.4 (R Core Team, 2010). Phenotypic data distributions and quantile-quantile (Q-Q) plots were generated in R and correlation matrices between the phenotypes were built using the pheatmap version 1.0.12 package. The broad-sense heritability values for the quality phenotypes were estimated using a linear mixed model with the est_herit function and the kinship matrices with the calc_kinship function implemented in R/qtl2 (Broman *et al.*, 2019).

5.3.3 Tissue Collection, Nucleic Acid Extraction, and Sequencing

The tissue collection and sequencing experimental procedures were performed as detailed in Gélinas Bélanger *et al.* (2024). Briefly, genomic DNA was extracted from the leaf tissue of the offspring and parental plants grown in the greenhouse 25 days after sowing (V4 stage) for both populations (Fehr *et al.*, 1971). Total DNA was extracted from tissue using the Omega Bio-Tek Mag-bind Plant Kit (Omega Biotek, Norcross, GA, USA) with further purification using the Mag-Bind Total Pure NGS (Omega Biotek, Norcross, GA, USA). Sampling of the $QS15524_{F2:F3}$ parental lines for the whole genome sequencing (WGS) was performed by pooling the samples from the five pots used to grow each of the parental lines, extracting total DNA, and having the libraries prepared at the Génome Québec Innovation Centre (Montréal, QC, Canada) using the NxSeq® AmpFREE Library Preparation kit (Lucigen, Middleton, WI, U.S.A.). To do so, the two parental libraries were barcoded, combined, and sequenced to a depth of 15X on the Illumina HiSeq X platform with 150 base pair paired-end reads. The WGS data for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ parental lines were also retrieved from the GmHapMap as available and detailed in Torkamaneh *et al.* (2021). To generate the genotyping-by-sequencing (GBS) datasets of the $QS15524_{F2:F3}$ (F_2 generation) and $QS15544_{RIL}$ mapping populations, the libraries were prepared at the Institute of Integrative Biology and Systems (Laval University, QC, Canada) using the *PstI/MspI* enzymes as detailed in Abed *et al.*, (2019). Samples were randomly divided into two sets of 91 individuals, which were barcoded and pooled to form two libraries per population. Sequencing of the $QS15524_{F2:F3}$ GBS libraries was done by combining a total of 91 barcoded samples per library. Sequencing of each library was done on four Ion PI V3 Chips per library with sequencing performed on the Ion Proton Sequencer and HiQ chemistry at the Institute of Integrative Biology and Systems, for a total of eight sequenced chips. For the $QS15544_{RIL}$ population, samples were randomly divided into two sets of 91 samples and sequenced using the same technologies, with two chips per library.

5.3.4 Bioinformatics

The bioinformatic procedures were performed as detailed in Gélinas Bélanger *et al.* (2024). Briefly, WGS data were processed using the fast-WGS pipeline (Torkamaneh *et al.*, 2018) for the $QS15524_{F2:F3}$ parental lines, whereas GBS data for both populations were processed using the fast-GBS pipeline (Torkamaneh *et al.*, 2017). All alignment procedures were performed using version

2 of the *Glycine max* reference genome (Gmax_275_v2.0). For the GBS datasets, genotypes were filtered using vcftools version 0.1.16 (Danecek *et al.*, 2011) to: i) maintain only biallelic sites; ii) remove InDels; iii) keep polymorphisms located only on chromosomes and not scaffolds; and iv) filter allele frequency and count with the –maxmissing 0.2, –maf 0.3 and –mac 4 commands. Missing genotypes for the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations were then self-imputed using Beagle version 4.1.0 with 12 iterations (Browning and Browning, 2016). Genotypes were phased with Convert2map (https://bitbucket.org/jerlar73/convert-genotypes-to-mapping-files/src/master, accessed December 12th, 2022) using the parental data from the GmHapMap for the $QS15544_{RIL}$ population and the fast-WGS resequenced data for the $QS15524_{F2:F3}$ parental lines. Subsequently, correction of the genotype calls for the $QS15524_{F2:F3}$ population was performed using Genotype Corrector (Miao *et al.*, 2018) using the software default options (sliding window size of 11 and error rates for homo1 and homo2 of 0.03 and 0.01, respectively) and all the implemented quality checks. For the $QS15544_{RIL}$ population, the removal of the double crossovers was performed using Convert2map. For the $QS15524_{F2:F3}$ population, binning was performed with the binning option implemented in Genotype Corrector.

5.3.5 Building of the Linkage Maps and QTL Mapping

The linkage maps of the $QS15544_{RIL}$ and $QS15524_{F2:F3}$ populations were generated using QTL IciMapping version 4.2 (Meng *et al.*, 2015) with the Kosambi mapping function to convert the recombination frequency into centimorgans (cM) as described in Gélinas Bélanger *et al.* (2024). Linkage groups (LG) were split when gaps exceeded 30 cM and the markers were anchored to their physical positions. The quality of these maps was confirmed by plotting (i) the genetic distance vs. the physical position and (ii) the pairwise recombination fraction and logarithm of the odds (LOD) score. To identify loci associated with the six seed traits of interest, QTL mapping was performed using two standard mapping algorithms that have complementary features: (i) Inclusive composite interval mapping (ICIM) approach implemented in QTL IciMapping and (ii) Genome-wide composite interval mapping (GCIM) method implemented in the QTL.gCIMapping.GUI package (Zhang *et al.*, 2020b) as detailed in Gélinas Bélanger *et al.* (2024) with minor modifications.

Genome-wide composite interval mapping was performed using the fixed model and a walking speed of 1 cM for both populations. In addition, the restricted maximum likelihood (REML) function was chosen to perform mapping in the $QS15524_{F2:F3}$ population. For ICIM,

mapping was performed using the following mapping parameters: i) deletion of the missing phenotypes; ii) a scanning interval step of 1 cm and a PIN of 0.001; and iii) a LOD threshold determined with 1,000 permutations and α of 0.05. The quality of the phenotypic data was verified by: (i) assessing the normality of the distribution; (ii) identifying obvious outliers present in the dataset; and (iii) plotting Q-Q plots. To find the LOD threshold using the permutation approach (1,000 permutations per trait), we ran individual analyses for each phenotype in the QTL IciMapping software for both populations and obtained thresholds ranging from LOD scores of 3.41 to 3.85 for $QS15544_{RIL}$ and 4.00 to 4.18 for $QS15524_{F2:F3}$. Based on these results, we decided to choose the upper limit (i.e., 3.85 for $QS15544_{RIL}$ and 4.18 for $QS15524_{F2:F3}$) as the final threshold for the ICIM results for all traits (Appendix 3.1). To remove minor QTL from the GCIM analysis, the LOD threshold was increased from the default value of 2.5 to the same threshold as for the ICIM algorithm (Zhang *et al.*, 2020b). To facilitate the understanding of the paper, all of the QTL presented in this article are reported based on the alleles of the early-maturing parents '9004' ($QS15544_{RIL}$ population) and 'OAC Vision' ($QS15524_{F2:F3}$ population).

5.3.6 QTL Filtering and Nomenclature

Following the identification of the regions, we classified them either as of higher interest or minor interest based on: (i) meeting the minimal LOD thresholds of 3.85 for $QS15544_{RIL}$ and 4.18 for $QS15524_{F2:F3}$; and (ii) QTL identified in the same genomic regions from both populations for the same trait; or (iii) simultaneous regulations of multiple seed quality traits (e.g., oil content and 100-seed weight) from neighboring QTL in the same population; or (iv) two neighboring regions with LOD scores ≥ 7 that were identified in the same population and for the same trait; or (v) their presence in a pod-filling and/or maturity QTL previously reported in Gélinas Bélanger *et al.* (2024). Regions that fulfilled one or more of these criteria were considered of higher interest and are presented in the following text, while the others (i.e., those meeting only the minimal LOD thresholds of 3.85 for $QS15544_{RIL}$ and 4.18 for $QS15524_{F2:F3}$) were considered of minor interest and compiled as such in Appendix 3.2. Following their classification, all regions were then investigated to identify those that were previously reported in the literature. To do so, we extended the regions by adding 1 Mbp both upstream and downstream to their flanking markers. Subsequently, the extended regions were then compared with all the QTL regions (biparental and

GWAS) available in Soybase (Grant *et al.*, 2009) and all regions with overlapping loci were considered previously reported in the literature.

Following QTL identification and filtering, each QTL was named with the algorithm, population, and trait information using the following nomenclature: Algorithm_Population_Trait. For example, the GCIM_F2_weight_9 QTL was identified in the $QS15524_{F2:F3}$ population using GCIM for the 100-seed weight trait hit #9. Regions that were found with both algorithms within 2 Mbp of each other were merged into a single locus using the following nomenclature: Trait_Population_GCIMhit_ICIMhit. For example, the Weight_F2_G9_I5 was identified both with GCIM (hit #9) and ICIM (hit #5) in the $QS15524_{F2:F3}$ population for the 100-seed weight trait. Regions considered as being of higher interest (i.e., fulfilling at least one of the aforementioned criteria) were merged using the following nomenclature: Population_Chromosome. For example, Bothpop_GM06 was identified on chromosome GM06 within both populations. All the details regarding the merging of the QTL regions are available in Appendix 3.2.

5.3.7 Identification of Candidate SNPs and Genes

The candidate SNPs and genes were identified using a five-step custom pipeline as described in Gélinas Bélanger *et al.* (2024) with slight modifications. First, we generated a database of genes based on Soybase's (Grant *et al.*, 2009) protein, fatty acid, and seed weight annotations (herein named PFASW terms) to limit the number of candidates investigated. To do so, we screened all of the Soybase annotations and retrieved all terms related to (i) protein synthesis (11 terms); (ii) fatty acid synthesis (44 terms); and (iii) seed weight (11 terms). Subsequently, we retrieved 1,084 genes from Soybase which were annotated with at least one of these terms. To complement this list, we also added six candidates from Derbyshire *et al.* (2023), one candidate from Duan *et al.* (2022), and one candidate from Zhang *et al.* (2018) that were not present in the list of 1,084 candidates from Soybase. Second, this list of 1,092 candidates (Appendix 3.3) was shortlisted by retaining only the genes that were located within 1 Mbp both upstream or downstream of the flanking markers of the merged regions. Third, we retained only variants that were predicted to have moderate or high consequences on the protein conformation. Fourth, putative effects of identified non-synonymous missense mutations were then predicted using Sorting Intolerant From Tolerant 4G (SIFT4g) (Ng and Henikoff, 2003; Kumar *et al.*, 2009). To do so, we generated a soybean database using the annotations of *G. max* Wm82.a2.v1 from EnsemblPlants and by

following the SIFT4G_Create_Genomic_DB guidelines (https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB, accessed December 12th, 2022). Single nucleotide polymorphisms with SIFT scores < 0.05 were classified as putatively deleterious and the ones ≥ 0.05 were considered as tolerated. Fifth, variants were retained based on the parental alleles generating the phenotype, and variants with similar allele patterns in the other population but without an effect on the trait were removed.

5.4 Results

5.4.1 Linkage Map Construction and Phenotypic Analysis

In our previous study (Gélinas Bélanger *et al.*, 2024), 286,844,986 ($QS15544_{RIL}$) and 541,106,451 ($QS15524_{F2:F3}$) unique single-end reads were generated from the sequencing steps of the $QS15544_{RIL}$ and $QS15524_{F2:F3}$ populations. Using these two datasets, we were able to generate two high-quality linkage maps from 2,746 ($QS15544_{RIL}$) and 1,613 ($QS15524_{F2:F3}$) GBS-derived SNP markers (Appendix 3.4, 3.5) that were subsequently used in the present study to perform linkage mapping for six quality traits.

In this study, we phenotyped the field-grown seeds of both populations using NIR and generated the data used for the present analysis. Each of the traits was normally distributed in the $QS15544_{RIL}$ (Fig. 4.1, 4.2) and $QS15524_{F2:F3}$ (Fig. 4.3, 4.4) populations based on the shape of histogram distributions and Q-Q plots. Moreover, both populations exhibited an adequate range for all the studied traits. In the $QS15544_{RIL}$ population, large differences were observed in the 100-seed weight trait (mean value 20.2 g, min value 15.8 g, and max value 24.1 g), protein content (mean value 40.7%, min value 38.5%, and max value 43.0%), and oil content (mean value 20.4%, min value 18.8%, and max value 21.7%). Similar observations were noted in the $QS15524_{F2:F3}$ population, with an oil content ranging from 17.9 to 20.6% (mean value of 19.3%), a protein content between 39.1 and 44.0% (mean value of 41.5%), and a 100-seed weight between 15.7 and 23.3 g (mean value of 19.4 g). Transgressive segregation was observed for the linolenic acid trait in the $QS15544_{RIL}$ population and the 100-seed weight trait in the $QS15524_{F2:F3}$ population.

Additional descriptive statistics (e.g. standard deviation, kurtosis, skewness, and heritability values) and phenotypic data for the six different traits are available in Appendix 3.6 and Appendix 3.7, respectively. The broad-sense heritability values for each of the traits were high (i.e., $H^2 \geq 0.5$) for most traits (i.e., 100-seed weight, oil, protein, oleic acid, and linoleic acid contents) and

moderate (i.e., $0.5 < H^2 \geq 0.3$) for the linolenic and linoleic traits. In the $QS15544_{RIL}$ population, oleic acid content was strongly negatively correlated (PCC of -0.73) with the content of linoleic acid. In contrast, the 100-seed weight was strongly positively correlated (PCC of 0.65) with the number of days to maturity (Fig. 4.5). In addition, the 100-seed weight was also moderately positively correlated with the pod-filling (PCC of 0.46) and flowering (PCC of 0.45) traits. In the $QS15524_{F2:F3}$ population, the oleic acid content was strongly negatively correlated (PCC of -0.50) with the linolenic acid content (Fig. 4.6). In both populations, the protein content was strongly negatively correlated with the oil content (Fig. 4.5, 4.6).

5.4.2 QTL Mapping

Our combinatorial mapping approach using GCIM and ICIM yielded a total of 36 (18 regions both for $QS15544_{RIL}$ and $QS15524_{F2:F3}$) and 35 (23 for $QS15544_{RIL}$ and 12 for $QS15524_{F2:F3}$) QTL, respectively (Appendix 3.2). We subsequently merged the loci from each trait that were identified using both GCIM and ICIM in each of the populations and identified 14 and 9 regions that were identified with both algorithms in $QS15544_{RIL}$ and $QS15524_{F2:F3}$, respectively. Conversely, we also found that 13 and 12 QTL were found with only one algorithm in $QS15544_{RIL}$ and $QS15524_{F2:F3}$, respectively. In $QS15544_{RIL}$, the LOD scores ranged from 3.9 to 18.2 for the QTL identified with GCIM and 4.1 to 21.4 for those detected with ICIM. In $QS15524_{F2:F3}$, the LOD scores ranged from 4.3 to 21.7 for GCIM and 4.3 to 14.1 for ICIM. As a general rule, we found that the regions identified with both algorithms displayed higher LOD scores and phenotypic variation explained (PVE). For instance, the average LOD and PVE values were respectively 5.7 and 5.2% for the regions identified with only one algorithm, whereas the LOD and PVE were respectively 10.0 and 10.5% for those identified with both GCIM and ICIM. Of the 71 QTL that were identified in this study, the largest number (i.e., 27 regions) was identified for the 100-seed weight trait, whereas the lowest was identified for the linolenic trait (i.e., 3 regions).

5.4.3 Identification of Overlapping QTL Signals in Both Populations

Three regions with overlapping loci were identified in both populations and had contrasting effects on multiple traits (Fig. 4.7, 4.8; Table 4.1). The first region (Bothpop_GM06) was identified between GM06:13,440,263 and GM06:16,257,054 and regulates the content of oleic and linoleic fatty acids in the $QS15544_{RIL}$ population as well as the 100-seed weight trait in both populations.

This region is associated with three loci in QS15544$_{RIL}$ (Oleic_RIL_G2_I2, ICIM_RIL_linoleic_2, and Weight_RIL_G3_I2) and one locus in QS15524$_{F2:F3}$ (GCIM_F2_weight_4). Amongst these four loci, the highest LOD scores (9.3-17.5) were obtained for the 100-seed weight trait in the QS15544$_{RIL}$ population (Weight_RIL_G3_I2 locus), whereas the lowest LOD (4.8) was obtained for the oleic acid trait in the QS15544$_{RIL}$ population (GCIM_RIL_oleic_2). The additive effects obtained for these loci indicate that the allele from '9004' decreases the oleic acid and 100-seed weight traits and increases the content of linoleic acid. In the QS15524$_{F2:F3}$ population, the increase in 100-seed weight is associated with the allele from 'OAC Vision'. The LOD and PVE associated with the 100-seed weight trait in the QS15544$_{RIL}$ are higher than those from the QS15524$_{F2:F3}$.

The second region (Bothpop_GM13) was detected on chromosome GM13 between positions GM13:34,528,497 and GM13:38,686,988. The Bothpop_GM13 locus regulates the 100-seed weight, linolenic acid, linoleic acid, and oil content traits in QS15544$_{RIL}$. In QS15524$_{F2:F3}$, the region regulates the 100-seed weight trait. In QS15544$_{RIL}$, the allele from '9004' decreases the 100-seed weight and linolenic acid content, whereas it increases the contents in oil and linoleic acid. In QS15524$_{F2:F3}$, the allele from 'OAC Vision' increases the 100-seed weight. The LOD scores (7.3 with GCIM and 14.9 with ICIM) and PVE (11.4% with GCIM and 13.8% with ICIM) associated with the oil content trait in the QS15544$_{RIL}$ are high, whereas the LOD and PVE associated with the other traits are lower.

The third region (Bothpop_GM19) was identified between GM19:41,471,856 and GM19:43,990,450 and regulates the 100-seed weight in both populations. In addition, this region controls the oil content in QS15544$_{RIL}$. In the QS15544$_{RIL}$ population, the allele from '9004' decreases the 100-seed weight trait and increases the oil content. In QS15524$_{F2:F3}$, the increase in 100-seed weight content is associated with the allele from 'OAC Vision' with an additive effect estimated at 0.84 g with GCIM and 0.98 g with ICIM. The LOD scores and PVE associated with the 100-seed weight trait in both populations are high.

5.4.4 Mapping of the QTL Unique to the QS15544$_{RIL}$ Population

In total, we observed five regions that were unique to the QS15544$_{RIL}$ population (Fig. 4.7, 4.9; Table 4.2). For these five regions, the LOD scores ranged from 7.1 to 21.4 and the PVE were between 2.6 and 28.5 %. Two of these regions had contrasting effects on more than one trait. The first region (RIL_GM04) is associated with the 100-seed weight trait and was identified between

the GM04:16,853,028 and GM04:18,312,993 flanking markers (i.e., Weight_RIL_G2_I1 locus) on LG04. The additive effects of -0.64 (GCIM) and -0.72 (ICIM) demonstrate that the allele from the early-maturity parent '9004' decreases the seed weight. Interestingly, this region overlaps the *E8-r1* pod-filling region (GM04:16,974,874 and GM04:17,152,230) identified in the same population (Gélinas Bélanger *et al.*, 2024). In the case of the *E8-r1* locus, the '9004' allele is associated with a decrease in the pod-filling period by 1.8 days.

The second region (RIL_GM05) is a small locus located on chromosome GM05 between markers GM05:41,123,772 and GM05:41,415,752 and which simultaneously regulates the levels of oleic and linoleic fatty acids. The additive effects associated with the RIL_GM05 region demonstrate that the allele from '9004' decreases the content of oleic acid, but increases the content of linoleic acid. The third region (RIL_GM10) is another small locus involved in the regulation of the 100-seed weight trait. The region is located between flanking markers GM10:2,629,415 and GM10:2,705,636 on LG10a and the additive effects of -0.34 (GCIM) and -0.31 (ICIM) demonstrate that the '9004' allele decreases the seed weight.

The fourth region (RIL_GM12) regulates the content of oil and is located between the flanking markers GM12:2,974,123 and GM12:4,503,160. The Oil_RIL_I4_I5 locus comprises two QTL detected with ICIM (ICIM_RIL_oil_4 and ICIM_RIL_oil_5) that are located ≈ 1.5 Mbp from each other. For the Oil_RIL_I4_I5 region only, the candidate variants are probably not the same due to the opposite effects on the final phenotype of each region detected with ICIM. As such, the additive effects associated with the ICIM_RIL_oil_4 and ICIM_RIL_oil_5 regions suggest that the alleles from '9004' contribute to decreasing and increasing the oil content, respectively. The fifth region (RIL_GM16) is associated with the Weight_RIL_G5_I6 locus and is located between the GM16:5,841,864 and GM16:5,861,155 flanking markers. The region is located near the GM16:5,680,173-5,730,237 locus previously identified in the same population for the number of days to maturity trait (Gélinas Bélanger *et al.*, 2024). In the present study, the '9004' allele contributes to an increase in the seed weight by 0.45 (ICIM) to 0.71 (GCIM) grams. In the case of the GM16:5,680,173-5,730,237 locus, the '9004' allele is associated with an increase of the number of days to maturity by 1.5-1.6 days.

5.4.5 Detection of the QTL Unique to the QS15524$_{F2:F3}$ Population

Using our combinatorial mapping approach, we identified four loci that were unique to the QS15524$_{F2:F3}$ population (Fig. 4.7, 4.10; Table 4.3). The LOD scores ranged from 4.3 to 15.6, whereas the PVE were between 1.1 and 20.5 %. Three of these regions had contrasting effects on multiple traits. The first region (F2_GM04.1) is associated with two QTL (Weight_F2_G3_I1 and Oil_F2_G1_I1) regulating the 100-seed weight and oil content traits. These loci are located in neighboring regions located at ≈ 1 Mbp from each other and have moderate LOD scores (4.3-8.8) and PVE (2.6-11.3%). The additive effects associated with these regions demonstrate that the allele from the early-maturing parent 'OAC Vision' decreases the 100-seed weight (i.e., additive effects of -0.34 with GCIM and -0.46 with ICIM), but increases the oil content (i.e., additive effects of 0.21 with GCIM and 0.31 with ICIM). The second region (F2_GM04.2) regulates the oleic fatty acid content and is flanked by the GM04:36,499,381 and GM04:40,206,770 markers associated with the Oleic_F2_G1_I1 locus. This region overlaps the *E8-r2* maturity locus located between the GM04:35,168,111 and GM04:37,664,017 markers that was previously identified in the same population (Gélinas Bélanger *et al.*, 2024). The allele from 'OAC Vision' increases the content of oleic acid, whereas it decreases the number of days to maturity.

The third region (F2_GM15) was identified between the GM15:3,593,701 and GM15:3,686,829 flanking markers and is associated with three different loci (Weight_F2_G7_I3, Oil_F2_G2_I2, and Prot_F2_G3_I1) regulating the 100-seed weight, oil content, and protein content traits. The LOD scores and PVE for these three loci are high and range from 7.3 to 15.6 % and 8.9 to 20.5 %, respectively. The region is small as the three regions co-localize between the GM15:3,593,701 and GM15:3,686,829 flanking markers. For this QTL, the allele from 'OAC Vision' caused a decrease in the oil content and seed weight, but an increase in the protein content. The fourth region (F2_GM17) is flanked by the GM17:40,269,918 and GM17:41,149,771 markers and regulates the levels of oleic acid (ICIM_F2_oleic_2 locus), seed weight (Weight_F2_G8_I4) and the content of protein (GCIM_F2_prot_4). The additive effects obtained for these loci indicate that the allele from 'OAC Vision' increases the oleic acid content and seed weight, and decreases the protein content.

5.4.6 Identification of Candidate SNPs and Genes

Following the detection of 12 major QTL in both populations, we identified a total of 38 (24 SIFT-Tolerated and 14 SIFT-Deleterious) candidate SNPs located in the coding sequences (CDS) of 25 genes located in eight major QTL regions (Table 4.4). Several of these mutations were located in candidate genes annotated with PFASW functions such as *Glyma.05G245000* (*3-OXO-5-α-STEROID 4-DEHYDROGENASE*), *Glyma.06G168000* (*FATTY ACYL-ACP THIOESTERASES B*; *GmFATB4B*), *Glyma.13G262500*/*FATTY-ACID-BINDING PROTEIN 3* (*GmFAP3*), *Glyma.06G166800* (*GmSWEET15*), and *Glyma.15G049200* (*GmSWEET39*).

In both populations, non-synonymous missense mutations in PFASW candidate genes were identified for the Bothpop_GM06, Bothpop_GM13, and Bothpop_GM19 loci. Each of these regions influences the 100-seed weight trait for both populations, but also additional traits related to fatty acid synthesis in the QS15544$_{RIL}$ population. For the 100-seed weight trait, we observed that the early-maturing parents (i.e., '9004' and 'OAC Vision') exhibited contrasting phenotypes for each of these three regions. For the Bothpop_GM06 and Bothpop_GM13 loci, we respectively identified six candidate genes comprising a total of nine missense mutations and three candidate genes comprising a total of 5 mutations. In almost all cases (12 out of 14), these mutations were found in a single parent, suggesting that the candidate gene might be different for both populations. Interestingly, many candidate genes located in both of these regions are orthologs of well-characterized PFASW genes (e.g., *GmFATB4B*, *GmSWEET15*, and *GmFAP3*). For the Bothpop_GM19 region, we identified an A→T (pos GM19:42,560,224) missense mutation in *Glyma.19G164800*/*CRUCIFERIN 3* (*GmCRU3*) that was present in 'Maple Arrow' and '9004'. This genotype pattern is in concordance with the respective additive effects of the Weight_RIL_G6_I7 (i.e., decrease of 0.80 for GCIM and 0.69 for ICIM in '9004) and Weight_F2_G9_I5 (i.e., increase of 0.84 for GCIM and 0.98 for ICIM in 'OAC Vision') regions. Another interesting variant was identified in *Glyma.19G161300* (pos GM19:42,207,842), a gene annotated with the 'Seed Development' PFASW term. In addition to these candidates, we also identified a mutation in another *TRANSPARENT TESTA8* gene (*Glyma.19G155300*/*GmTT8*), a gene with validated fatty acid and 100-seed weight functions in *Arabidopsis* (Chen *et al.*, 2014).

In the QS15544$_{RIL}$ population, non-synonymous missense mutations were identified for the RIL_GM05, RIL_GM10, and RIL_GM12 regions. We detected three mutations in two candidate genes (*Glyma.05G245000*, 1 mutation; *Glyma.05G244100*, 2 mutations) associated with the

RIL_GM05 locus. In the RIL_GM10 region, we identified one missense mutation for *Glyma.10G026000*/*TRANSPARENT TESTA8* (*GmTT8b*) and three missense mutations for *Glyma.10G032000* (*F BOX-LIKE17*). For the RIL_GM12 locus, we identified an A→C (pos GM12:1,971,068) missense mutation in *Glyma.12G027300* (*ENOYL-ACP REDUCTASE 1*; *GmENR1*). In all but one case (mutation pos GM10:2,270,537 located in the CDS of *GmTT8b*), the mutations were found only in 'AAC Mandor' and none of the three other parental lines. In the case of the mutation located at pos GM10:2,270,537 of the *GmTT8b* gene, it was found only in '9004' and none of the three other parental lines.

In the $QS15524_{F2:F3}$ population, we also discovered variants in PFASW candidate genes in two regions (F2_GM15 and F2_GM17). For the F2_GM15 locus, the only identified mutation was a T→A (pos GM15:3,875,093) missense mutation in *Glyma.15G049200* (*GmSWEET39*). In addition, we identified missense mutations in five different genes located in the F2_GM17 locus, including four in *Glyma.17G236700* (*ACYL-COA-BINDING DOMAIN 3*; *GmACBP3*) and one in *Glyma.17G247700* (*DA*; *GmDA1*). Each of the mutations associated with the F2_GM15 and F2_GM17 loci was either found only in 'OAC Vision' or 'Maple Arrow' and none of the three other parental lines.

5.5 Discussion

5.5.1 Identification of Three Novel Loci

For breeders, novel loci are essential to increase the allelic diversity of their programs and build a SNP catalog for a wide variety of commercial traits. In the present study, we identified a total of 12 major regions that were either associated with $QS15544_{RIL}$ (5 loci), $QS15524_{F2:F3}$ (4 loci), or both (3 loci) and successfully identified candidate SNPs in eight of them (Fig. 4.7). To identify these regions, we developed a pipeline using two complementary mapping algorithms, ICIM and GCIM. The first approach, ICIM, has been shown to have a reduced false discovery rate, increased detection power, and less biased estimates of QTL effects, thus making it a robust method for the identification of major regions (Li *et al.*, 2007). On the other hand, it has been demonstrated that GCIM has a great ability to identify small-effect and linked QTL, particularly in F_2 populations (Zhang *et al.*, 2020b). Based on our trials, these algorithms can be efficiently used in tandem to validate each other and find the regions of higher interest while at the same time identifying novel regions that generally have a smaller impact on the studied phenotype. To the best of our

knowledge, three regions (out of 12) were not previously identified and therefore might serve as novel sources of regulation for the content of oleic acid (F2_GM04.2), oil (RIL_GM12) and seed weight (RIL_GM16).

For these novel loci, we were able to identify one candidate variant for the RIL_GM12 region (i.e., GM12:1,971,068), but none for the two others. As previously mentioned, two closely located QTL with opposite effects on oil content were detected for the RIL_GM12 region (ICIM_RIL_oil_4 and ICIM_RIL_oil_5) (Table 4.2). Based on our five-step variant analysis pipeline, we identified only one Sift-Tolerated missense variant for *Glyma.12G027300* (*ENOYL-ACP REDUCTASE 1*; *GmENR1*), meaning that one of these regions remains without candidate genes (Table 4.4). In *Arabidopsis*, the *mosaic death1* (AT2G05990; *AtMOD1*) mutant, an ortholog of the soybean *Glyma.12G027300* gene, encodes an enoyl-acyl carrier protein reductase which is involved in *de novo* synthesis of fatty acids.

5.5.2 Detection of Two Major Regions Involved in Seed Quality

In soybean, oil and protein contents are quantitative traits controlled by two major loci located on GM15 and GM20 (cqSeed protein-001 and cqSeed protein-003), and a large number of small-effect QTL across the genome (Grant *et al.*, 2009; Liu *et al.*, 2023). In many cases, these loci exhibit a pleiotropic action on both traits and, as such, a decrease in the content of one trait generally leads to an increase in the content of the other (Grant *et al.*, 2009; Liu *et al.*, 2023). Most often, this negative correlation between both traits is associated with the same natural variant [e.g., *GmSWEET39* (Zhang *et al.*, 2020a)] and cannot be segregated. In the present study, the F2_GM15 locus was found to be overlapping or close to many previously identified major QTL associated with protein, seed weight, and oil content on GM15 (Table 4.3). Over the years, numerous studies have brought to light the role of *Glyma.15G049200* (*GmSWEET39*) in the regulation of protein and oil content and seed weight (Yang *et al.*, 2019; Zhang *et al.*, 2020a, 2021) (Table 4.4). Using Variant effect predictor, we identified an arginine to serine missense mutation located at amino acid position 246 in *GmSWEET39*. The same mutation was identified in Kumar *et al.* (2023) and is associated with a decrease of ≈ 8% in oil content with no significant changes in protein content in comparison to the wild-type. In our analysis, we found that a decrease in oil content was associated with the allele from 'OAC Vision' and explained between 17.2 and 20.3 % of the phenotypic variation observed for this trait (Table 4.3). Concomitantly, the F2_GM15 locus was

also associated with an increase in protein in 'OAC Vision' corresponding to about 13.3-20.5% of the PVE for this trait.

Furthermore, we also identified significant statistical associations with the oleic and linoleic traits for the RIL_GM05 locus (Table 4.2). Using our five-step variant analysis pipeline, we identified candidate variants in two genes, *Glyma.05G245000* (*3-OXO-5-α-STEROID 4-DEHYDROGENASE*) and *Glyma.05G244100* (*MOTHER OF FT AND TFL1*; *GmMFT*) (Table 4.4). The *3-OXO-5-α-STEROID 4-DEHYDROGENASE* gene has been identified as the main candidate for the regulation of oleic acid in a large association panel comprising 621 soybean accessions and might be involved in the elongation of very-long-chain fatty acids (Sung *et al.*, 2021). A recent study has confirmed the role of *GmMFT* in the regulation of seed size as well as linoleic acid, oleic acid, protein, and oil contents in soybean (Cai *et al.*, 2023). Cai *et al.* (2023) demonstrated that the content of linoleic acid in soybean seed is significantly decreased in *Gmmft* knockout lines, but significantly increased in overexpression lines. They also demonstrated that *GmMFT* regulates the expression of several *GmSWEET* genes, including *GmSWEET39*, a candidate gene for the F2_GM15 locus. Interestingly, the variants found in the present study for 'AAC Mandor' do not correspond to the eight *GmMFT* haplotypes studied in Cai *et al.* (2023). We screened the GmHapMap lines (Torkamaneh *et al.*, 2021) to identify the prevalence of the mutations identified in the present study and found that the homozygous forms of the G→A (pos GM05:41,853,377) and A→C (pos GM05:41,854,422) missense variants were present in a low proportion of the dataset, with respectively 5.8 (58 lines out of 1,007) and 21.7 % (218 lines out of 1,007) of the lines harboring these variants. As such, these variants might serve as novel sources of variation for the regulation of linoleic and oleic acid seed content.

5.5.3 Breeding Considerations

Developing early-maturing soybean cultivars for MG00 and MG000 growing zones with high seed quality is challenging due to the lower genetic diversity available and possible linkage drag/pleiotropic effects between desirable and undesirable alleles. An optimal understanding of the possible bottlenecks associated with linkage drags/pleiotropic effects is thus necessary to optimize breeding efforts. Three regions (Bothpop_GM06, Bothpop_GM13, and Bothpop_GM19) exhibiting QTL for the 100-seed weight were identified in both populations. As demonstrated in the results section, several other traits (e.g., oil and various fatty acid contents) were identified as

co-localizing in these regions (Fig. 4.7), suggesting a close linkage between those traits and 100-seed weight that could potentially limit options for breeders due to linkage drag. Linkage drag is a well-known phenomenon that can potentially impart negative effects on essential breeding traits and thus slow the selection process. To overcome linkage drag, breeders must identify rare recombinants among desirable allelic variants which requires resources and time (Voss-Fels *et al.*, 2017). In the case of the Bothpop_GM06, Bothpop_GM13, and Bothpop_GM19 regions, further investigation is required to confirm the potential bottlenecks associated with linkage drag for the different traits.

In the present study, we also demonstrated that two critical loci (RIL_GM04 and RIL_GM16) associated with the 100-seed weight trait are overlapping or very close to two important regions regulating the number of days to pod-filling (i.e., region GM04:16,974,874-17,152,230, also named *E8-r1*) and the number of days to maturity (i.e., region GM16:5,680,173-5,730,237) in the $QS15544_{RIL}$ population (Gélinas Bélanger *et al.*, 2024). As demonstrated, the 100-seed trait is strongly positively correlated with the flowering, pod-filling, and maturity traits in the $QS15544_{RIL}$ population (Fig. 4.5). Consequently, these observations suggest that possible linkage drag events or pleiotropic effects might limit breeding opportunities for the 100-seed weight trait in these particular sources of early maturity. Nevertheless, the range of seed weight (15.70-24.10 g) observed in our two populations remains acceptable, thus indicating that these limitations can be overcome with a proper selection of parents as other loci controlling the seed weight can be prioritized in breeding programs. To develop cultivars with earlier maturity/shorter pod-filling and higher seed weight, we suggest relying on other pod-filling and maturity loci identified in this population (e.g., *E8-r2*, GM16:22,756,017-23,154,638 and GM07:5,256,305-5,4049,71) that are not linked with the 100-seed weight trait. Interestingly, 'OAC Vision' seems to be a more valuable parent than '9004' to generate early-maturing material with increased grain size using the Bothpop_GM06, Bothpop_GM13, and Bothpop_GM19 loci. Moreover, the linkage drag/pleiotropic effects of RIL_GM04 and RIL_GM16 loci on seed weight in '9004' also suggests that 'OAC Vision' might be better for breeding purposes. On a final note, most of the traits, exceptions made for 100-seed weight/oil content, protein content/oil content, and oleic acid/linoleic acid, were weakly correlated, meaning that most traits can be individually selected by breeders. Overall, our results demonstrate that breeding for early maturity can be achieved without significantly impacting seed quality traits.

5.6 Supplemental data

Appendix 3.1 Threshold values for the logarithm of the odds obtained for each trait in both populations.

Appendix 3.2 Major and minor *QTL* identified in each of the populations.

Appendix 3.3 Lists of gene ontology annotations and candidate genes associated with PFASW functions.

Appendix 3.4 Linkage map for the $QS15544_{RIL}$ population.

Appendix 3.5 Linkage map for the $QS15524_{F2:F3}$ population.

Appendix 3.6 Descriptive statistics for each of the studied traits.

Appendix 3.7 Phenotypic data for each of the studied traits.

5.7 Information

Acknowledgments

We would like to thank Dr. Martine Jean and Vincent-Thomas Boucher St-Amour for their advice on QTL mapping and linkage map construction. We would like to acknowledge Maxime Carrier, Florence Vachon Laberge, Rebecca Lacroix, Marie-Ève Lachance Foisy and Daphnée Paré for the phenotypic data collection.

Competing interests statement

The authors declare there are no competing interests.

Author contribution statement

Conceptualization: JGB, LO

Data curation: JGB

Formal analysis: JGB

Funding acquisition: LO

Investigation: JGB

Methodology: JGB, LO

Project administration: LO

Resources: LO

Supervision: TRC, VH, LO

Validation: JGB

Visualization: JGB

Writing – original draft: JGB

Writing – review & editing: TRC, VH, LO

Funding statement

This work was supported by Génome Québec and Genome Canada with funds awarded to the SoyaGen Project and by the Canadian Field Crop Research Alliance and Agriculture and Agri-Food Canada under the Agri-Innovation Program. JGB was supported by the Natural Sciences and Engineering Research Council of Canada, Mitacs, les Fonds de recherche du Québec volet Nature et Technologie, Centre SÈVE and Seed World Group.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: Bioproject #PRJNA1035514, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1035514.

5.8 References

Abed, A., Légaré, G., Pomerleau, S., St-Cyr, J., Boyle, B., and Belzile, F. J. (2019). "Genotyping-by-sequencing on the ion torrent platform in barley," in *Barley*, (Springer), 233–252.

Agronomix Software Inc. (2009). Agrobase Generation II. Available at: https://www.agronomix.com/

Bilyeu, K., Škrabišová, M., Allen, D., Rajcan, I., Palmquist, D. E., Gillen, A., *et al.* (2018). The Interaction of the Soybean Seed High Oleic Acid Oil Trait With Other Fatty Acid Modifications. *JAOCS, J. Am. Oil Chem. Soc.* 95, 39–49. doi: 10.1002/aocs.12025

Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., *et al.* (2019). R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics* 211, 495–502. doi: 10.1534/genetics.118.301595

Browning, B. L., and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Cai, Z., Xian, P., Cheng, Y., Zhong, Y., Yang, Y., Zhou, Q., *et al.* (2023). MOTHER-OF-FT-AND-TFL1 regulates the seed oil and protein content in soybean. *New Phytol.* 239, 905–919. doi: 10.1111/nph.18792

Canadian Grain Commission (2022). Quality of Canadian oilseed-type soybeans 2022. *Grain*

Harvest Export Qual. - Can. Grain Comm. Available at: https://www.grainscanada.gc.ca/en/grain-research/export-quality/oilseeds/soybean-oil/2022/04-oil-protein.html (Accessed January 17, 2024).

Chen, M., Xuan, L., Wang, Z., Zhou, L., Li, Z., Du, X., *et al.* (2014). TRANSPARENT TESTA8 Inhibits Seed Fatty Acid Accumulation by Targeting Several Seed Development Regulators in *Arabidopsis*. *Plant Physiol.* 165, 905–916. doi: 10.1104/pp.114.235507

Clemente, T. E., and Cahoon, E. B. (2009). Soybean Oil: Genetic Approaches for Modification of Functionality and Total Content. *Plant Physiol.* 151, 1030–1040. doi: 10.1104/pp.109.146282

Cober, E. R., Daba, K. A., Warkentin, T. D., Tomasiewicz, D. J., Mooleki, P. S., Karppinen, E. M., *et al.* (2023). Soybean seed protein content is lower but protein quality is higher in Western Canada compared with Eastern Canada. *Can. J. Plant Sci.* 103, 411–421. doi: 10.1139/CJPS-2022-0147

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Derbyshire, M. C., Marsh, J., Tirnaz, S., Nguyen, H. T., Batley, J., Bayer, P. E., *et al.* (2023). Diversity of fatty acid biosynthesis genes across the soybean pangenome. *Plant Genome* 16, e20334. doi: 10.1002/tpg2.20334

Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., *et al.* (2022). Natural allelic variation of GmST05 controlling seed size and quality in soybean. *Plant Biotechnol. J.* 20, 1807–1818.

Fehr, W. R., Caviness, C. E., Burmood, D. T., and Pennington, J. S. (1971). Stage of Development Descriptions for Soybeans, Glycine Max (L.) Merrill. *Crop Sci.* 11, 929–931. doi: 10.2135/cropsci1971.0011183x001100060051x

Foyer, C. H., Siddique, K. H. M., Tai, A. P. K., Anders, S., Fodor, N., Wong, F. L., *et al.* (2019). Modelling predicts that soybean is poised to dominate crop production across Africa. *Plant Cell Environ.* 42, 373–385. doi: 10.1111/pce.13466

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., and O'Donoughue, L. (2024). Dissection of the E8 locus in two early maturing Canadian soybean populations. *Front. Plant Sci.* 15, 1329065. doi: 10.3389/fpls.2024.1329065

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, 843–846. doi: 10.1093/nar/gkp798

Iquira, E., Gagnon, É., and Belzile, F. (2010). Comparison of genetic diversity between canadian adapted genotypes and exotic germplasm of soybean. *Genome* 53, 337–345. doi: 10.1139/G10-009

Isaacs, J. (2020). Is High-Protein, High-Yielding Soy Possible? *Germination Mag.* Available at: https://germination.ca/is-high-protein-high-yielding-soy-possible/ (Accessed January 17, 2024).

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous

variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1082. doi: 10.1038/nprot.2009.86

Kumar, V., Goyal, V., Mandlik, R., Kumawat, S., Sudhakaran, S., Padalkar, G., *et al.* (2023). Pinpointing Genomic Regions and Candidate Genes Associated with Seed Oil and Protein Content in Soybean through an Integrative Transcriptomic and QTL Meta-Analysis. *Cells* 12, 1–20. doi: 10.3390/cells12010097

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics* 175, 361–374. doi: 10.1534/genetics.106.066811

Li, R., Chen, H., Yang, Z., Yuan, S., and Zhou, X. (2020). Research status of soybean symbiosis nitrogen fixation. *Oil Crop Sci.* 5, 6–10. doi: 10.1016/j.ocsci.2020.03.005

Lin, C.-S., and Poushinsky, G. (1985). a Modified Augmented Design (Type 2) for Rectangular Plots. *Can. J. Plant Sci.* 65, 743–749. doi: 10.4141/cjps85-094

Liu, D., Yan, Y., Fujita, Y., and Xu, D. (2018). Identification and validation of *QTL* for 100-seed weight using chromosome segment substitution lines in soybean. *Breed. Sci.* 68, 442–448. doi: 10.1270/jsbbs.17127

Liu, S., Liu, Z., Hou, X., and Li, X. (2023). Genetic mapping and functional genomics of soybean seed protein. *Mol. Breed.* 43, 29. doi: 10.1007/s11032-023-01373-5

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Miao, C., Fang, J., Li, D., Liang, P., Zhang, X., Yang, J., *et al.* (2018). Genotype-Corrector: Improved genotype calls for genetic mapping in F2 and RIL populations. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-28294-0

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Pagano, M. C., and Miransari, M. (2016). "The importance of soybean production worldwide," in *Abiotic and Biotic Stresses in Soybean Production*, (Elsevier Inc.), 1–26. doi: 10.1016/B978-0-12-801536-0.00001-3

Qi, Z., Song, J., Zhang, K., Liu, S., Tian, X., Wang, Y., *et al.* (2020). Identification of QTNs Controlling 100-Seed Weight in Soybean Using Multilocus Genome-Wide Association Studies. *Front. Genet.* 11, 1–12. doi: 10.3389/fgene.2020.00689

Qin, P., Wang, T., and Luo, Y. (2022). A review on plant-based proteins from soybean: Health benefits and soy product development. *J. Agric. Food Res.* 7, 100265. doi: 10.1016/j.jafr.2021.100265

R Core Team (2010). R: A Language and Environment for Statistical Computing. *Vienna, Austria R Found. Stat. Comput.* Available at: http://www.gnu.org/copyleft/gpl.html.

Schaalje, G. B., Lynch, D. R., and Kozub, G. C. (1987). Field evaluation of a modified augmented design for early stage selection involving a large number of test lines without replication. *Potato Res.* 30, 35–45. doi: 10.1007/BF02357682

Singer, W. M., Lee, Y. C., Shea, Z., Vieira, C. C., Lee, D., Li, X., *et al.* (2023). Soybean genetics, genomics, and breeding for improving nutritional value and reducing antinutritional traits in food and feed. *Plant Genome* 16, e20415. doi: 10.1002/tpg2.20415

St-Amour, V. T. B., Mimee, B., Torkamaneh, D., Jean, M., Belzile, F., and O'Donoughue, L. S. (2020). Characterizing resistance to soybean cyst nematode in PI 494182, an early maturing soybean accession. *Crop Sci.* 60, 2053–2069. doi: 10.1002/csc2.20162

Sung, M., Van, K., Lee, S., Nelson, R., LaMantia, J., Taliercio, E., *et al.* (2021). Identification of SNP markers associated with soybean fatty acids contents by genome-wide association analyses. *Mol. Breed.* 41, 1–16. doi: 10.1007/s11032-021-01216-1

The American Soybean Association (2023). SoyStats - International: World Soybean Production (2021/2022 year). *Am. Soybean Assoc.* Available at: http://soystats.com/ (Accessed December 11, 2023).

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., and Belzile, F. (2017). Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics* 18, 1–7. doi: 10.1186/s12859-016-1431-9

Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., *et al.* (2018). Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnol. J.* 16, 749–759. doi: 10.1111/pbi.12825

Torkamaneh, D., Laroche, J., Valliyodan, B., O'Donoughue, L., Cober, E., Rajcan, I., *et al.* (2021). Soybean (Glycine max) Haplotype Map (GmHapMap): a universal resource for soybean translational and functional genomics. *Plant Biotechnol. J.* 19, 324–334. doi: 10.1111/pbi.13466

Tylka, G. L., and Marett, C. C. (2021). Known Distribution of the Soybean Cyst Nematode, Heterodera glycines, in the United States and Canada in 2020. *Plant Heal. Prog.* 22, 72–74. doi: 10.1094/PHP-10-20-0094-BR

Voss-Fels, K. P., Qian, L., Parra-Londono, S., Uptmoor, R., Frisch, M., Keeble-Gagnère, G., *et al.* (2017). Linkage drag constrains the roots of modern wheat. *Plant Cell Environ.* 40, 717–725. doi: 10.1111/pce.12888

Yang, H., Wang, W., He, Q., Xiang, S., Tian, D., Zhao, T., *et al.* (2019). Identifying a wild allele conferring small seed size, high protein content and low oil content using chromosome segment substitution lines in soybean. *Theor. Appl. Genet.* 132, 2793–2807. doi: 10.1007/s00122-019-03388-z

Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M. L., *et al.* (2020a). Selection of GmSWEET39 for oil and protein improvement in soybean. *PLOS Genet.* 16, e1009114. Available at: https://doi.org/10.1371/journal.pgen.1009114

Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., *et al.* (2018). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Mol. Plant* 11, 460–472. doi: 10.1016/j.molp.2017.12.016

Zhang, S., Hao, D., Zhang, S., Zhang, D., Wang, H., Du, H., *et al.* (2021). Genome-wide

association mapping for protein, oil and water-soluble protein contents in soybean. *Mol. Genet. Genomics* 296, 91–102. doi: 10.1007/s00438-020-01704-7

Zhang, Y., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2020b). QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked *QTL* for quantitative traits in bi-parental segregation populations. *Comput. Struct. Biotechnol. J.* 18, 59–65. doi: 10.1016/j.csbj.2019.11.005

Region	QTL	Population	Trait	LG	Position of the QTL peak (cM)		Marker		LOD		PVE (%)		Additive effect		Effect of the allele from the early-maturing parent ('9004' or 'OAC Vision')	Soybase hits* and notes
					GCIM	ICIM	Left	Right	GCIM	ICIM	GCIM	ICIM	GCIM	ICIM		
Bothpop_GM06 (GM06:13,440, 263-16,257,054)	Oleic_RIL_G2_I2	RIL	Oleic	6b	5.74	30.00	GM06:13,440,263	GM06:14,883,459	4.80	8.76	6.58	10.93	−0.72	−0.68	↓ (9004)	n.d.†
	ICIM_RIL_linoleic_2	RIL	Linoleic	6b	n.d.†	30.00	GM06:14,871,473	GM06:14,883,459	n.d.†	5.79	n.d.†	7.86	n.d.†	0.65	↑ (9004)	n.d.†
	Weight_RIL_G3_I2	RIL	100-seed weight	6b	36.06	48.00	GM06:15,271,680	GM06:16,257,054	9.27	17.5	3.06	15.61	−0.37	−0.63	↓ (9004)	≥3 regions, see Table S1
	GCIM_F2_weight_4	F2	100-seed weight	6	113.00	n.d.†	GM06:14,645,419	GM06:15,830,779	5.07	n.d.†	2.38	n.d.†	0.33	n.d.†	↑ (OV)	≥3 regions, see Table S1
Bothpop_GM13 (GM13:34,528, 497-38,686,988)	Linolenic_RIL_G1_I1	RIL	Linolenic	13b	1.08	1.00	GM13:34,528,497	GM13:35,164,031	3.85	4.19	10.93	9.21	−0.24	−0.24	↓ (9004)	n.d.†
	Oil_RIL_G3_I6	RIL	Oil	13b	4.74	12.00	GM13:35,164,031	GM13:36,629,203	7.29	14.87	11.43	13.84	0.25	0.26	↑ (9004)	Seed oil 37-8; Seed oil 2-g1
	ICIM_RIL_weight_5	RIL	100-seed weight	13b	n.d.†	19.00	GM13:37,276,161	GM13:37,284,883	n.d.†	5.95	n.d.†	4.61	n.d.†	−0.36	↓ (9004)	Seed weight 13-g9
	ICIM_RIL_linoleic_3	RIL	Linoleic	13b	n.d.†	42.00	GM13:38,027,686	GM13:38,620,690	n.d.†	4.45	n.d.†	5.72	n.d.†	0.57	↑ (9004)	n.d.†
	Weight_F2_G6_I2	F2	100-seed weight	13b	51.00	52.00	GM13:38,627,374	GM13:38,686,988	8.41	4.32	4.37	5.19	0.44	0.47	↑ (OV)	n.d.†
Bothpop_GM19 (GM19:41,471, 856-43,990,450)	Weight_RIL_G6_I7	RIL	100-seed weight	19	122.31	122.00	GM19:43,240,106	GM19:43,736,913	14.59	19.47	14.25	18.35	−0.80	−0.69	↓ (9004)	≥3 regions, see Table S1
	Oil_RIL_G4_I7	RIL	Oil	19	124.81	124.00	GM19:43,736,913	GM19:43,990,450	5.73	7.11	6.26	6.16	0.19	0.16	↑ (9004)	Seed oil 43-28
	Weight_F2_G9_I5	F2	100-seed weight	19	77.00	70.00	GM19:41,471,856	GM19:43,990,450	21.65	12.34	16.21	17.01	0.84	0.98	↑ (OV)	≥3 regions, see Table S1

Note: QTL, quantitative trait loci; LG, linkage group; GCIM, Genome-wide composite interval mapping; ICIM, Inclusive composite interval mapping; LOD, logarithm of the odds; PVE, phenotypic variation explained. *Soybase hits indicate loci previously reported in soybean.
†n.d., not detected.

Table 4.1 Major quantitative trait loci identified in both populations for six seed quality traits.

Region	QTL	Trait	LG	Position of the QTL peak (cM)		Marker		LOD		PVE (%)		Additive effect		Effect of the allele from the early-maturing parent ('9004')	Soybase hits* and notes
				GCIM	ICIM	Left	Right	GCIM	ICIM	GCIM	ICIM	GCIM	ICIM		
RIL_GM04 (GM04:16,853,028–18,312,993)	Weight_RIL_G2_I1	100-seed weight	4	221.8	242.0	GM04:16,853,028	GM04:18,312,993	13.6	21.4	9.0	20.4	−0.64	−0.72	↓	Seed weight 36-15 + Overlaps the maturity locus *E8-r1*
RIL_GM05 (GM05:41,123,772–41,415,752)	Linoleic_RIL_G1_I1	Linoleic	5	248.6	247.0	GM05:41,123,772	GM05:41,415,752	7.5	11.4	10.6	16.3	0.8	0.94	↑	Seed linoleic 6-5; Seed linolenic 4-g5
	Oleic_RIL_G1_I1	Oleic	5	248.6	248.0	GM05:41,392,627	GM05:41,415,752	18.2	19.1	17.4	28.3	−1.17	−1.10	↓	n.d.†
RIL_GM10 (GM10:2,629,415–2,705,636)	Weight_RIL_G4_I3	100-seed weight	10a	28.8	29.0	GM10:2,629,415	GM10:2,705,636	8.1	8.1	2.6	6.4	−0.34	−0.31	↓	Seed weight 4-g9
RIL_GM12 (GM12:2,974,123–4,503,160)‡	Oil_RIL_I4_I5	Oil	12a	n.d.†	16.0 & 32.0	GM12:2,974,123	GM12:4,503,160	n.d.†	13.3 & 7.1	n.d.†	12.3 & 6.0	n.d.†	0.23 & −0.16	↑ & ↓	Two ICIM regions near each other
RIL_GM16 (GM16:5,841,864–5,861,155)	Weight_RIL_G5_I6	100-seed weight	16	57.2	57.0	GM16:5,841,864	GM16:5,861,155	10.4	9.3	11.1	7.3	0.71	0.45	↑	Overlaps the maturity locus GM16:5,680,173–5,730,237

Note: QTL, quantitative trait loci; LG, linkage group; GCIM, Genome-wide composite interval mapping; ICIM, Inclusive composite interval mapping; LOD, logarithm of the odds; PVE, phenotypic variation explained. *Soybase hits indicate loci previously reported in soybean.
†n.d., not detected using the method.
‡The Oil_RIL_I4_I5 region associated with this region comprises two QTL detected with ICIM (ICIM_RIL_oil_4 and ICIM_RIL_oil_5) with opposite effects that are located ≈1.5 Mbp from each other.

Table 4.2 Major quantitative trait loci specific to the QS15544$_{RIL}$ population for six seed quality traits.

Region	QTL	Trait	LG	Position of the QTL peak (cM)		Marker		LOD		PVE (%)		Additive effect		Effect of the allele from the early-maturing parent ('OAC Vision')	Soybase hits* and notes
				GCIM	ICIM	Left	Right	GCIM	ICIM	GCIM	ICIM	GCIM	ICIM		
F2_GM04.1 (GM04:8,039,183–9,230,671)	Weight_F2_G3_I1	100-seed weight	4	61.1	61.0	GM04:8,039,183	GM04:8,192,408	6.5	4.3	2.6	5.1	−0.34	−0.46	↓	≥3 regions, see Table S1
	Oil_F2_G1_I1	Oil	4	65.7	65.0	GM04:9,115,728	GM04:9,230,671	7.6	8.8	6.3	11.3	0.21	0.31	↑	n.d.†
F2_GM04.2 (GM04:36,499,381–40,206,770)	Oleic_F2_G1_I1	Oleic	4	82.2	80.0	GM04:36,499,381	GM04:40,206,770	6.7	8.9	7.2	15.9	0.72	1.05	↑	Overlaps the maturity locus *E8-r2*
F2_GM15 (GM15:3,593,701–3,686,829)	Prot_F2_G3_I1	Protein	15	60.0	59.0	GM15:3,593,701	GM15:3,657,456	14.8	8.4	13.3	20.5	0.53	0.6	↑	≥3 regions, see Table S1
	Oil_F2_G2_I2	Oil	15	60.3	62.0	GM15:3,657,456	GM15:3,686,829	15.6	14.1	17.2	20.3	−0.35	−0.43	↓	≥3 regions, see Table S1
	Weight_F2_G7_I3	100-seed weight	15	63.0	63.0	GM15:3,657,456	GM15:3,686,829	13.3	7.3	8.9	9.8	−0.62	−0.7	↓	n.d.†
F2_GM17 (GM17:39,433,946–41,149,771)	Weight_F2_G8_I4	100-seed weight	17	136.0	140.0	GM17:39,433,946	GM17:39,603,730	4.3	8.6	1.1	11.1	0.23	0.71	↑	≥3 regions, see Table S1
	ICIM_F2_oleic_2	Oleic	17	n.d.†	143.0	GM17:40,269,918	GM17:40,701,207	n.d.†	7.8	n.d.†	13.5	n.d.†	0.97	↑	Seed oleic 1-g11
	GCIM_F2_prot_4	Protein	17	n.d.†	145.4	GM17:41,149,771	GM17:41,149,771	n.d.†	4.5	n.d.†	2.5	n.d.†	−0.23	↓	Seed protein 39-3

Note: QTL, quantitative trait loci; LG, linkage group; GCIM, Genome-wide composite interval mapping; ICIM, Inclusive composite interval mapping; LOD, logarithm of the odds; PVE, phenotypic variation explained.
*Soybase hits indicate loci previously reported in soybean.
†n.d., not detected.

Table 4.3 Major quantitative trait loci specific to the QS15524$_{F2:F3}$ population for six seed quality traits.

Region	Gene	Nucleotide variant		Effect of SNP on trait(s)[‡]	Amino acid variant		SIFT Prediction	Soybase annotations[†]	Gene name	*Arabidopsis* ortholog	References[§]
		Position	W82/MA/OV/MD/90[$]		Position	W82/MA/OV/MD/90[$]					
RIL_GM05 (GM05:41,123,772–41,415,752)	Glyma.05G245000	GM05:41,912,828	T/T/T/**C**/T	↓ Oleic ↑ linoleic	295	N/N/N/**S**/N	Tolerated	Literature	*3-OXO-5-α-STEROID 4-DEHYDROGENASE*	AT1G18180	Gm (Zhang et al. 2018)
	Glyma.05G244100	GM05:41,853,377	G/G/G/**A**/G	↓ Oleic ↑ linoleic	123	S/S/S/**I**/S	Deleterious	Literature	*MOTHER OF FT AND TFL1*	AT1G18100	Gm (Li et al. 2015, 2019; Niu et al. 2020; Duan et al. 2022; Cai et al. 2023; Yang et al. 2024)
	Glyma.05G244100	GM05:41,854,422	A/A/A/**C**/A	↓ Oleic ↑ linoleic	106	L/L/L/**V**/L	Tolerated	Literature	*MOTHER OF FT AND TFL1*	AT1G18100	Gm (Li et al. 2015, 2019; Niu et al. 2020; Duan et al. 2022; Cai et al. 2023; Yang et al. 2024)
Bothpop_GM06 (GM06:13,440,263–16,257,054)	Glyma.06G166800	GM06:13,904,650	G/**A**/G/G/G	↓ Seed weight	12	V/**I**/V/V/V	Tolerated	Seed maturation	*SWEET15*	AT5G23660	At (Chen et al. 2015)
	Glyma.06G166800	GM06:13,906,051	A/**G**/A/A/A	↓ Seed weight	217	R/**G**/R/R/R	Deleterious	Seed maturation	*SWEET15*	AT5G23660	At (Chen et al. 2015)
	Glyma.06G168000	GM06:14,022,744	T/**A**/T/T/T	↓ Oleic ↑ palmitic ↑ linoleic	103	V/**E**/V/V/V	Deleterious	Fatty acid biosynthetic process	*FATTY ACYL-ACP THIOESTERASES B*	AT1G08510	At (Bonaventure et al. 2003; Branham et al. 2016)
	Glyma.06G170700	GM06:14,291,340	G/**T**/T/T/G	↓ Oleic ↑ palmitic ↑ linoleic	165	F/**L**/L/L/F	Tolerated	Fatty acid biosynthetic process	*SNF1-RELATED PROTEIN KINASE REGULATORY SUBUNIT GAMMA 1*	AT3G48530	n.d.
	Glyma.06G178700	GM06:15,164,955	T/**C**/C/**C**/T	↑ Seed weight	133	F/**S**/**S**/**S**/F	Tolerated	Seed maturation	*HISTONE DEACETYLASE 6*	AT5G63110	At (Tanaka, Kikuchi and Kamada 2008; Willmann et al. 2011)
	Glyma.06G178800	GM06:15,166,155	C/**T**/T/T/C	↑ Seed weight	49	R/**C**/C/**C**/R	Deleterious	Seed maturation	*HISTONE DEACETYLASE 6*	AT5G63110	At (Tanaka, Kikuchi and Kamada 2008; Willmann et al. 2011)
	Glyma.06G178800	GM06:15,166,263	G/**T**/T/T/G	↑ Seed weight	85	V/**L**/L/L/V	Tolerated	Seed maturation	*HISTONE DEACETYLASE 6*	AT5G63110	At (Tanaka, Kikuchi and Kamada 2008; Willmann et al. 2011)
	Glyma.06G186000	GM06:16,179,690	A/**T**/T/T/A	↑ Seed weight	35	Q/**L**/L/L/Q	Tolerated	Seed development	*CYCLOPS 1*	AT5G13690	At (Ronceret et al. 2008)
	Glyma.06G186000	GM06:16,183,298	A/**G**/G/G/A	↑ Seed weight	285	H/**R**/R/R/H	Tolerated	Seed development	*CYCLOPS 1*	AT5G13690	At (Ronceret et al. 2008)
RIL_GM10 (GM10:2,629,415–2,705,636)	Glyma.10G026000	GM10:2,270,537	T/T/T/T/**A**	↓ Seed weight	386	S/S/S/S/**T**	Tolerated	Seed coat development/seed development	*TRANSPARENT TESTA 8*	AT4G09820	At (Chen et al. 2014)
	Glyma.10G032000	GM10:2,793,136	T/T/T/**A**/T	↑ Seed weight	548	H/H/H/**L**/H	Tolerated	Seed development	*F BOX-LIKE17*	AT3G54650	n.d.
	Glyma.10G032000	GM10:2,795,327	C/C/C/**G**/C	↑ Seed weight	241	R/R/R/**P**/R	Deleterious	Seed development	*F BOX-LIKE17*	AT3G54650	n.d.
	Glyma.10G032000	GM10:2,796,990	T/T/T/**C**/T	↑ Seed weight	31	Y/Y/Y/**C**/Y	Deleterious	Seed development	*F BOX-LIKE17*	AT3G54650	n.d.
RIL_GM12 (GM12:2,974,123–4,503,160)	Glyma.12G027300	GM12:1,971,068	A/A/A/**C**/A	↓ or ↑ oil[‖]	93	I/I/I/**M**/I	Tolerated	Fatty acid biosynthetic process/fatty acid synthase complex	*ENOYL-ACP REDUCTASE 1*	AT2G05990	At (Mou et al. 2000)

Bothpop_GM13 (GM13:34,528, 497–38,686,988)	Glyma.13G262500	GM13:36,604,101	C/C/*/**T**/C	↑ Linolenic ↓ linoleic ↓ oil ↑ seed weight¶	29	S/S/*/**L**/S	Deleterious	Fatty acid binding/fatty acid metabolic process	*FATTY-ACID-BINDING PROTEIN 3*	AT1G53520	n.d.
	Glyma.13G262500	GM13:36,607,034	**T**/**T**/**T**/**G**/**T**	↑ Linolenic ↓ linoleic ↓ oil	152	L/L/L/**R**/L	Tolerated	Fatty acid binding/fatty acid metabolic process	*FATTY-ACID-BINDING PROTEIN 3*	AT1G53520	n.d.
	Glyma.13G262500	GM13:36,607,308	A/A/A/**C**/A	↑ Linolenic ↓ linoleic ↓ oil	210	Q/Q/Q/**P**/Q	Deleterious	Fatty acid binding/fatty acid metabolic process	*FATTY-ACID-BINDING PROTEIN 3*	AT1G53520	n.d.
	Glyma.13G276200	GM13:37,777,304	**T**/**T**/**T**/**A**/**T**	↑ Seed weight	94	C/C/C/**S**/C	Deleterious	Seed development	*GRIM REAPER*	AT1G53130	n.d.
	Glyma.13G291300	GM13:39,154,000	A/**T**/A/A/**T**	↑ Oil ↓ seed weight¶	7	S/**T**/S/S/**T**	Tolerated	Branched-chain amino acid biosynthetic process	*KETOL-ACID REDUCTOISOMERASE*	AT3G58610	Gm (Liu et al. 2020; Niu et al. 2020)
F2_GM15 (GM15:3,593, 701–3,686,829)	Glyma.15G049200	GM15:3,875,093	**T**/**T**/**A**/**T**/**T**	↓ Seed weight ↓ oil ↑ protein	246	R/R/**S**/R/R	Tolerated	Seed maturation	*SWEET39*	AT5G13170	Gm (Yang et al. 2019; H. Zhang et al. 2020; Zhang et al. 2021)
F2_GM17 (GM17:39,433, 946–41,149,771)	Glyma.17G236700	GM17:39,189,878	G/G/**A**/G/G	↑ Oleic	142	E/E/**K**/E/E	Tolerated	Fatty acid transport	*ACYL-COA-BINDING DOMAIN 3*	AT4G24230	n.d.
	Glyma.17G236700	GM17:39,189,902	A/A/**C**/A/A	↑ Oleic	150	K/K/**Q**/K/K	Tolerated	Fatty acid transport	*ACYL-COA-BINDING DOMAIN 3*	AT4G24230	n.d.
	Glyma.17G236700	GM17:39,189,965	G/G/**C**/G/G	↑ Oleic	171	D/D/**H**/D/D	Tolerated	Fatty acid transport	*ACYL-COA-BINDING DOMAIN 3*	AT4G24230	n.d.
	Glyma.17G236700	GM17:39,190,180	A/A/**T**/A/A	↑ Oleic	242	E/E/**D**/E/E	Tolerated	Fatty acid transport	*ACYL-COA-BINDING DOMAIN 3*	AT4G24230	n.d.
	Glyma.17G238400	GM17:39,389,410	A/**C**/A/A/A	↓ Seed weight	180	L/**V**/L/L/L	Tolerated	Seed development	*CYCLIN D3;1*	AT4G34160	n.d.
	Glyma.17G247700	GM17:40,308,933	G/**A**/G/G/G	↓ Seed weight	67	R/**K**/R/R/R	Tolerated	Plant ovule morphogenesis/regulation of seed growth/seed morphogenesis	*DA*	AT1G19270	At (Mora-Ramirez et al. 2021) / At (Li et al. 2008)
	Glyma.17G255400	GM17:40,939,386	**T**/**G**/**T**/**T**/**T**	↓ Oleic	325	E/**A**/E/E/E	Deleterious	Fatty acid homeostasis	*alpha/beta-Hydrolases superfamily protein*	AT4G24160	n.d.
	Glyma.17G262200	GM17:41,574,857	C/**A**/C/C/C	↑ Protein	96	P/**T**/P/P/P	Deleterious	Cellular amino acid catabolic process	*PHYLLO*	AT1G68890	n.d.
Bothpop_GM19 (GM19:41,471, 856–43,990,450)	Glyma.19G150000	GM19:41,018,818	**T**/**T**/**T**/**T**/*	↓ Seed weight	25	F/F/F/F/*	Tolerated	Seed development	*MTP8*	AT3G58060	n.d.
	Glyma.19G150000	GM19:41,024,050	G/G/G/G/**A**	↓ Seed weight	219	A/A/A/A/**T**	Deleterious	Seed development	*MTP8*	AT3G58060	n.d.
	Glyma.19G150000	GM19:41,024,888	A/A/A/A/**G**	↓ Seed weight	313	H/H/H/H/**R**	Deleterious	Seed development	*MTP8*	AT3G58060	n.d.
	Glyma.19G155300	GM19:41,564,721	C/**G**/**G**/**G**/C	↑ Seed weight ↓ oil	468	H/**D**/**D**/**D**/H	Tolerated	Seed coat development/seed development	*TRANSPARENT TESTA 8*	AT4G09820	At (Chen et al. 2014)
	Glyma.19G161300	GM19:42,207,842	G/G/*/**A**/G	↑ Seed weight	123	P/P/*/**S**/P	Tolerated	Seed development	*F BOX-LIKE17*	AT3G54650	n.d.
	Glyma.19G164800	GM19:42,560,224	A/A/**T**/**T**/A	↑ Seed weight	344	V/V/E/E/V	Tolerated	Seed maturation	*CRUCIFERIN 3*	AT4G28520	n.d.
	Glyma.19G170100	GM19:43,083,145	C/C/C/C/**A**	↓ Seed weight	224	G/G/G/G/**C**	Deleterious	Seed development	*SENSITIVE TO HOT TEMPERATURES 5*	AT5G43940	n.d.

Note: SIFT, Sorting Intolerant From Tolerant. [§]Nucleotides or amino acids differing from the wild-type Williams 82 are indicated by bold letters. An asterisk (*) indicates a heterozygote for the SNP of interest. W82, 'William 82'; MA, 'Maple Arrow,' OV, 'OAC Vision'; MD, 'AAC Mandor'; 90, '9004'.

[†]Soybase annotations associated with the investigated candidate genes based on the five-step variant analysis pipeline. Two candidate genes (Glyma.05G245000 and Glyma.05G244100) were identified using the literature as indicated in Table S1c.

[‡]Possible effect of SNP on traits identified in the QTL region. Traits were associated with Soybase function annotations unless otherwise stated. Single nucleotide polymorphisms are considered as nucleotides differing from the 'Williams 82' genotype.

[§]References associated with *Glycine max* (Gm) or *Arabidopsis thaliana* (At). Candidate genes without references are indicated as not detected (n.d.).

[‖]Two QTL (ICIM_RIL_oil_4 and ICIM_RIL_oil_5) with opposite effects were mapped ≈1.5 Mbp from each other, but only one candidate variant was identified using the five-step variant analysis pipeline.

¶Possible role in seed weight. The QTL results indicated contrasting effects on seed weight for the $QS15544_{RIL}$ and $QS15524_{F2:3}$ populations suggesting that the SNPs in these genes may have a role in seed weight.

Table 4.4 Candidate genes for eight major quantitative trait loci regions.

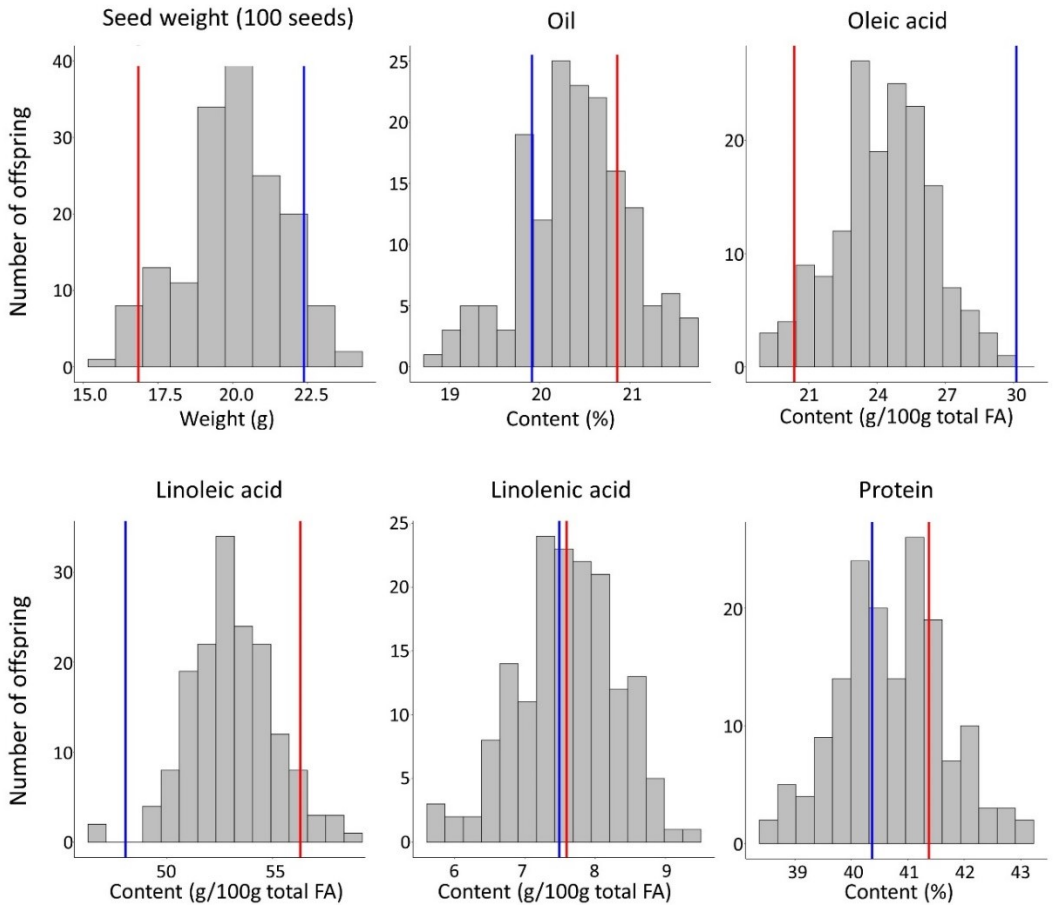

Figure 4.1 Phenotypic trait data distribution for the QS15544$_{RIL}$ population. Parental lines are indicated with vertical-coloured lines. Red lines, ‘9004’; Blue lines, 'AAC Mandor'.

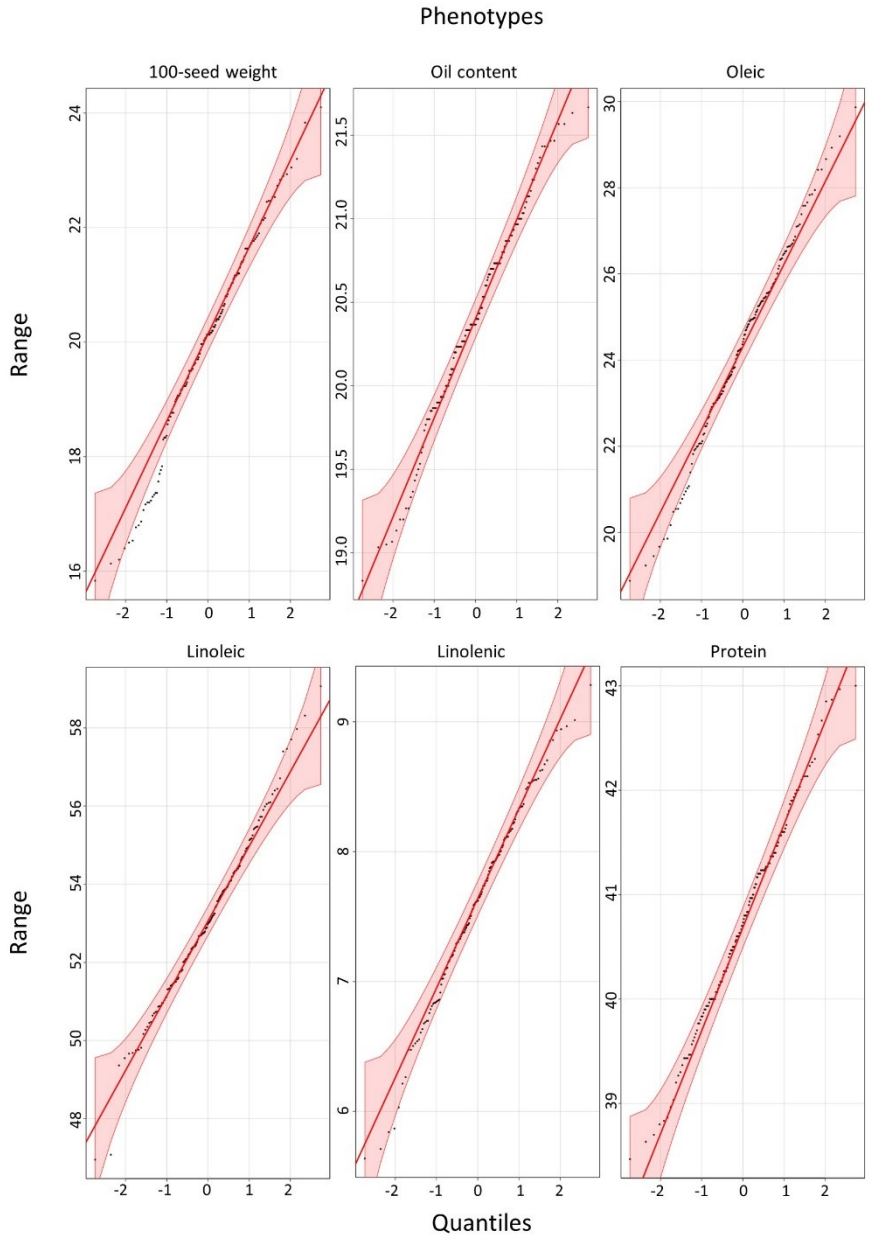

Figure 4.2 Quantile-quantile (Q-Q) plot demonstrating normal distribution for the seed quality phenotypes in the QS15544$_{RIL}$ population.

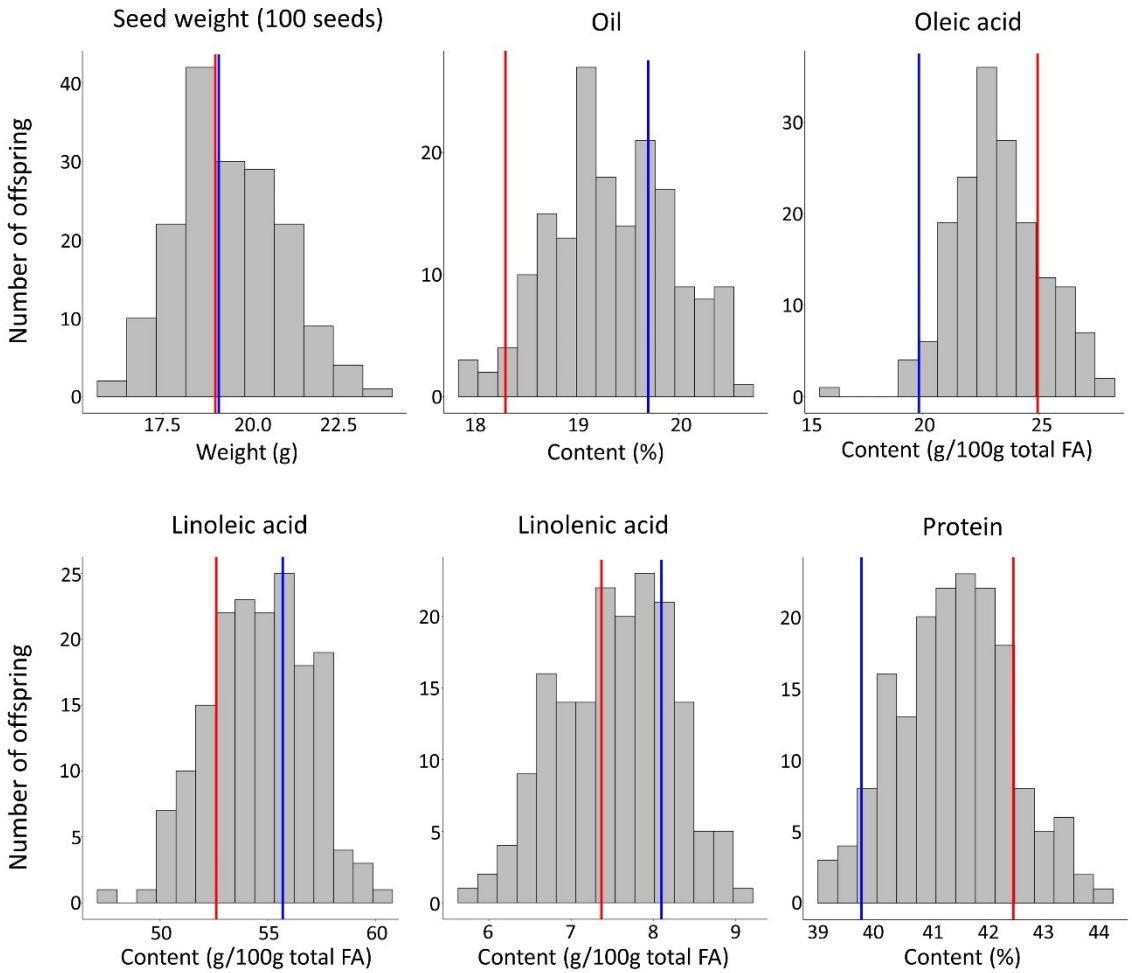

Figure 4.3 Phenotypic trait data distribution for the QS15524$_{F2:F3}$ population. Parental lines are indicated with vertical-coloured lines. Red lines, 'OAC Vision'; Blue lines, 'Maple Arrow'.

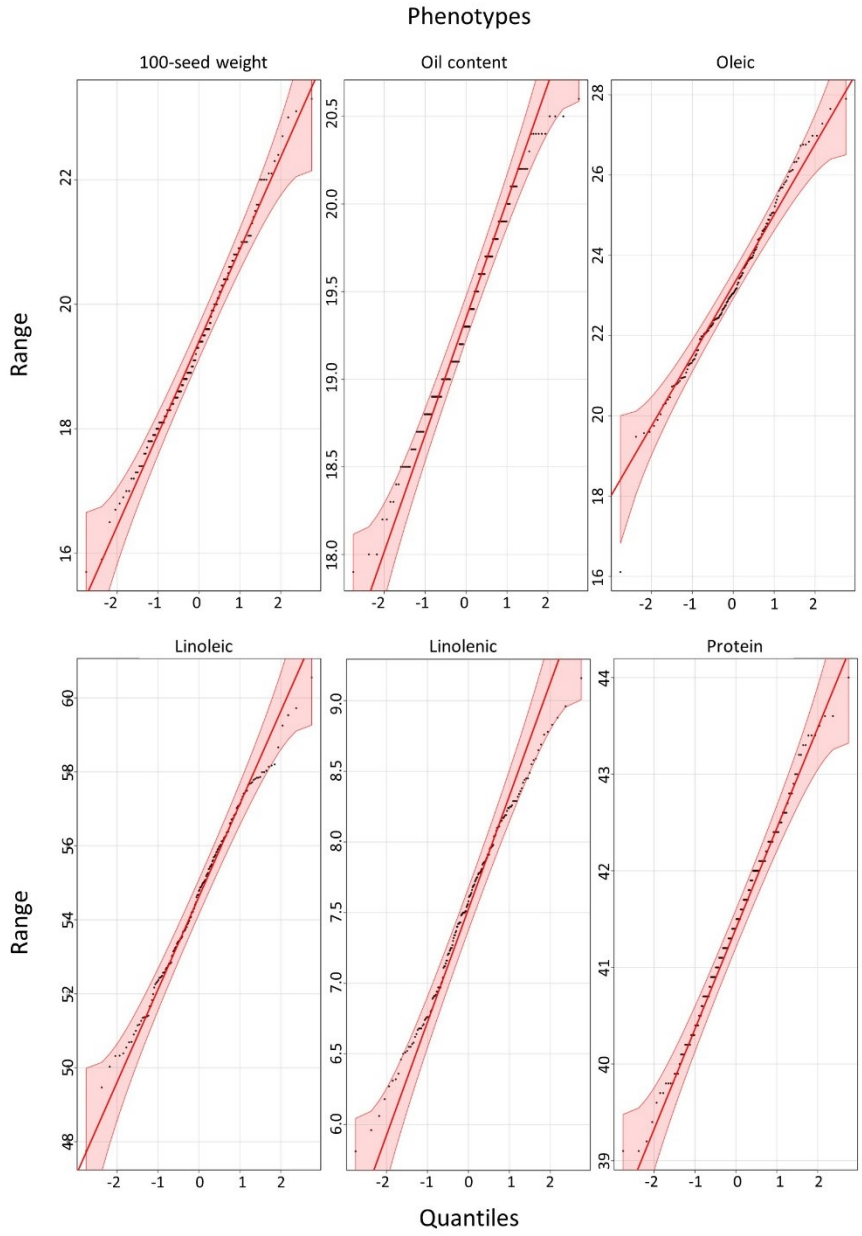

Figure 4.4 Quantile-quantile (Q-Q) plot demonstrating normal distribution for the seed quality phenotypes in the QS15524$_{F2:F3}$ population.

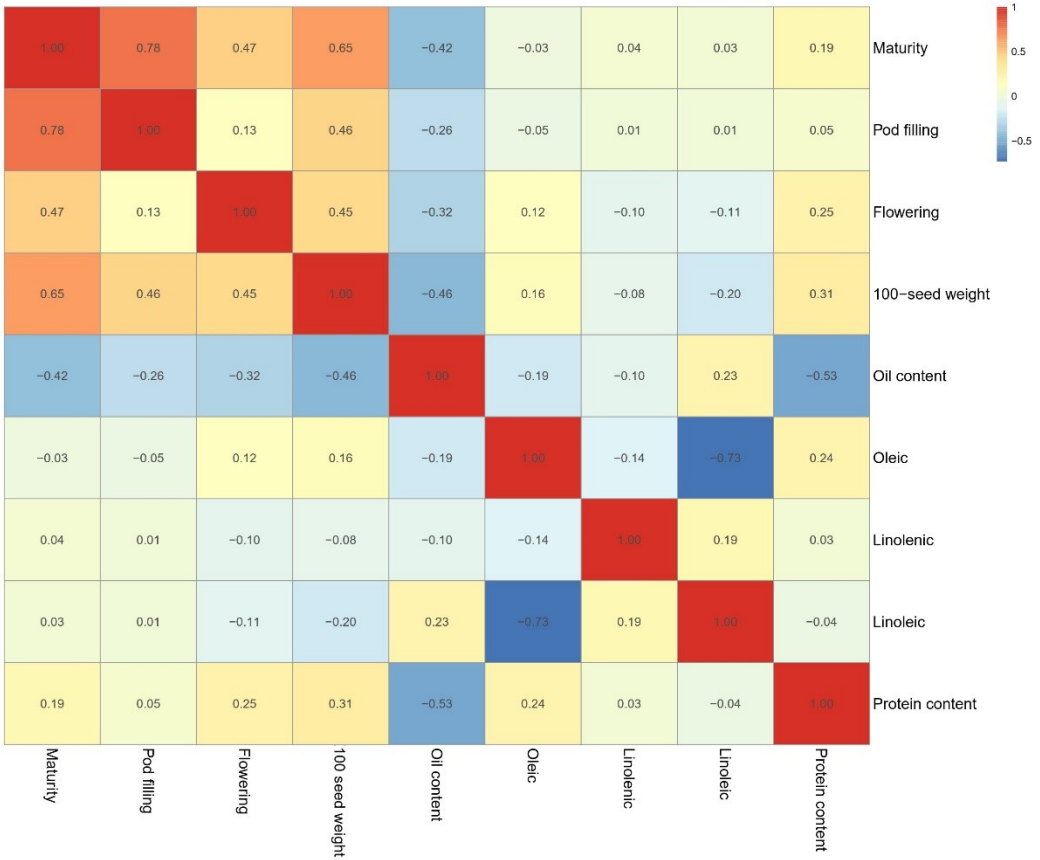

Figure 4.5 Pearson correlation coefficient matrix for the QS15544$_{RIL}$ population. Pearson correlations between six seed quality traits and three reproductive traits as available in Gélinas Bélanger *et al.* (2024).

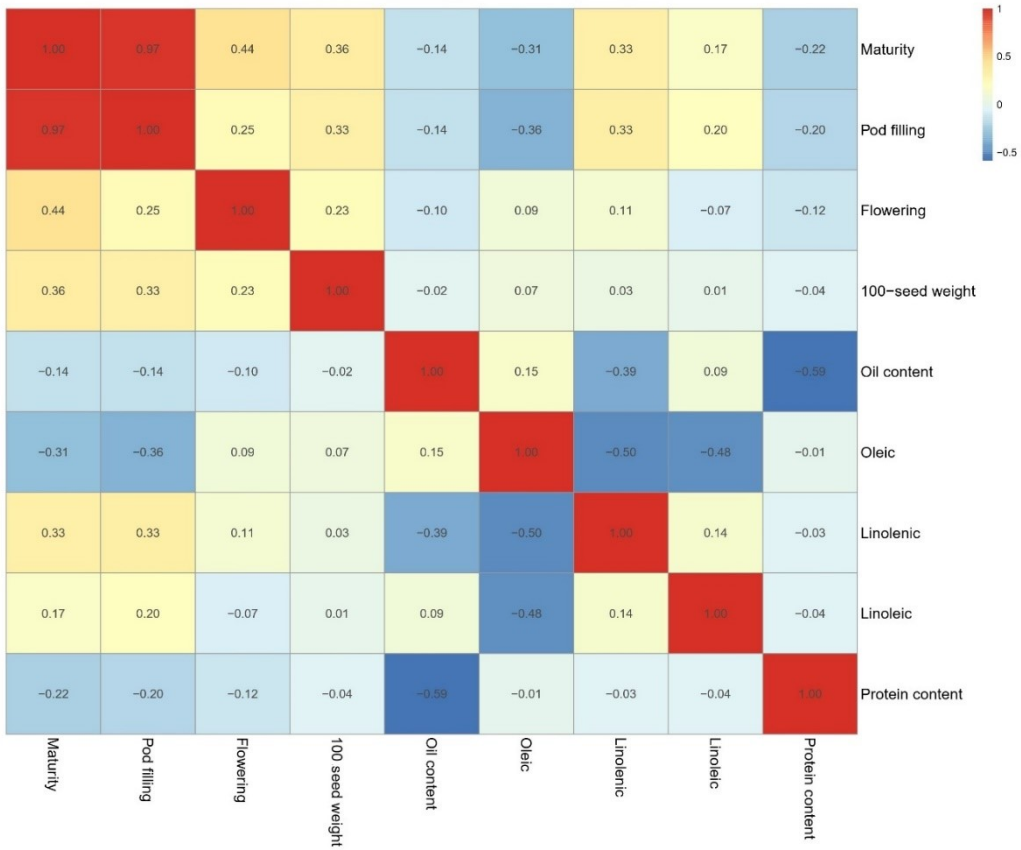

Figure 4.6 Pearson correlation coefficient matrix for the QS15524$_{F2:F3}$ population. Pearson correlations between six seed quality traits and three reproductive traits as available in Gélinas Bélanger *et al.* (2024).

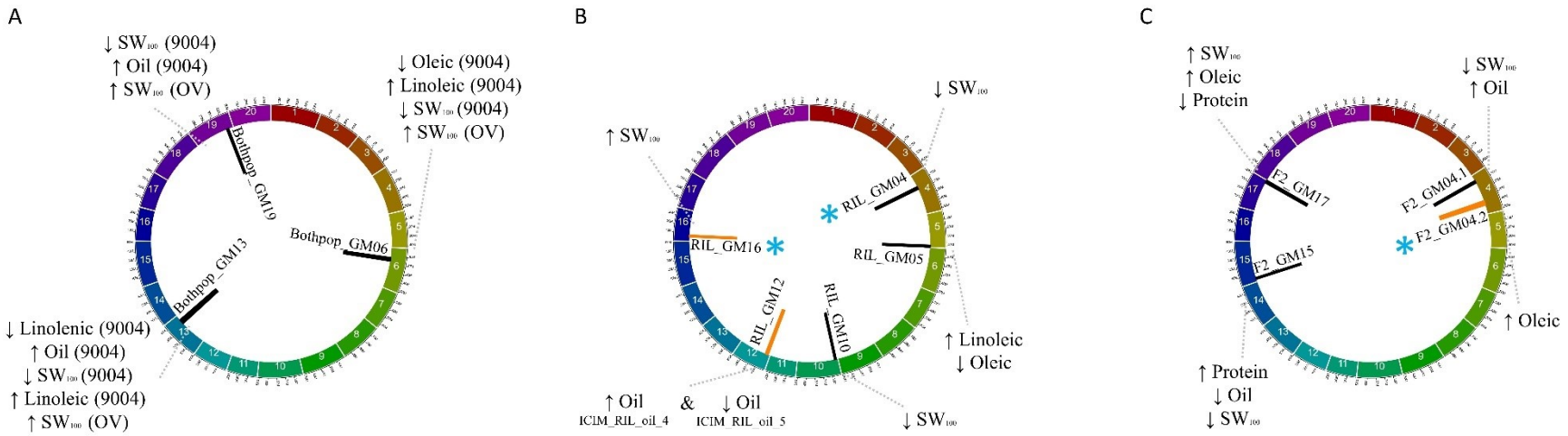

Figure 4.7 Identification of the QTL associated with different seed quality traits in the QS15544$_{RIL}$ and QS15524$_{F2:F3}$ populations. Circos plots illustrating the locations of the major QTL regions and the impacts of the alleles from the early-maturing parental lines on the studied traits. **(A)** QTL regions associated with both populations and effects associated with '9004' or 'OAC Vision'. **(B)** QTL regions associated with the QS15544$_{RIL}$ population and effects of '9004'. **(C)** QTL regions associated with the QS15524$_{F2:F3}$ population and effects of 'OAC Vision'. In Panel A, the traits associated with 'OAC Vision' are annotated as OV, whereas all the other traits found in the panel are associated with '9004' and annotated as such. Blue asterisk indicates the locations of three major regions, RIL_GM04, RIL_GM16, and F2_GM04.2, that respectively overlap the *E8-r1*, GM16:5,680,173-5,730,237, and *E8-r2* early-maturity and/or shorter pod-filling loci identified in Gélinas Bélanger *et al.* (2024). Orange and black bars respectively indicate the locations of novel loci and loci previously reported in Soybase.

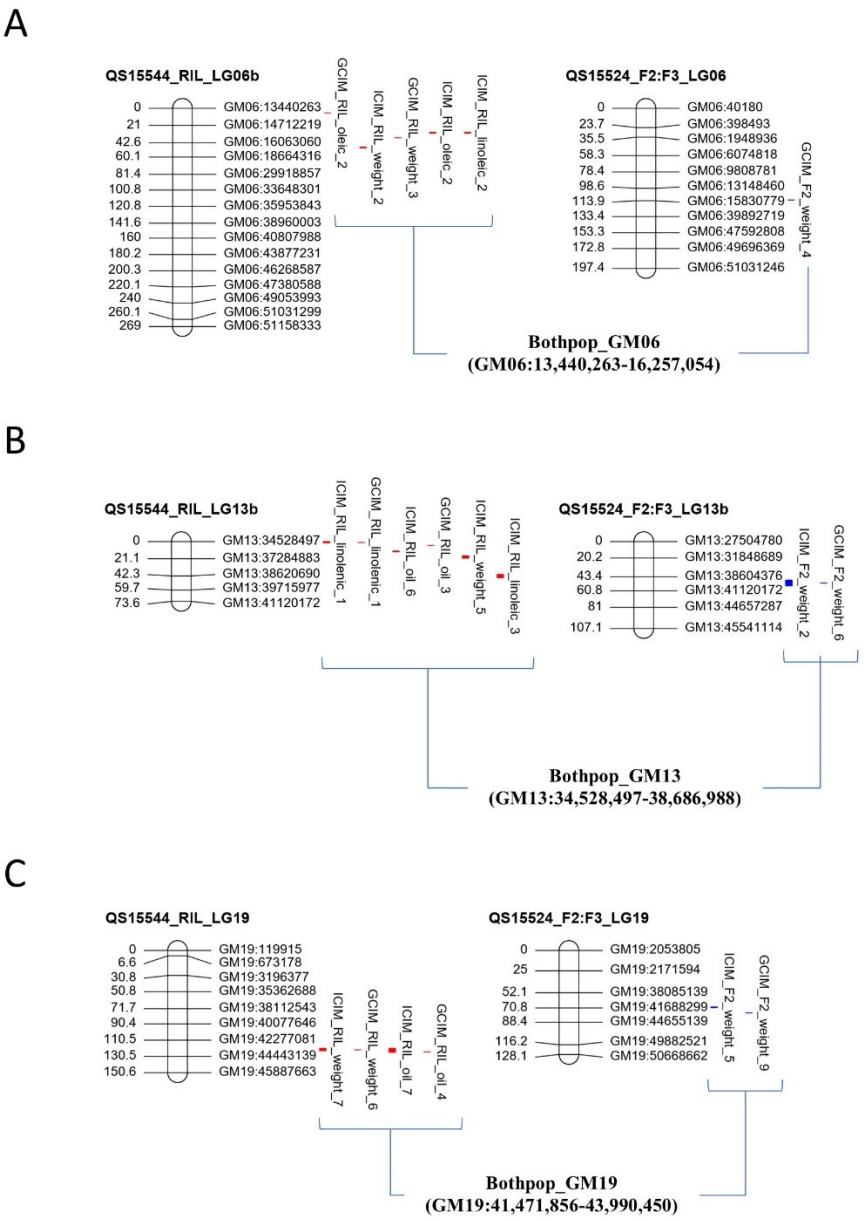

Figure 4.8 Overlapping quantitative trait loci identified in both populations. Overlapping QTL identified for the Bothpop_GM06 **(A)**, Bothpop_GM13 **(B)**, and Bothpop_GM19 **(C)**. Red-marked traits indicate the seed quality QTL identified in the $QS15544_{RIL}$ population, whereas the blue-marked traits indicate overlapping QTL identified in the $QS15524_{F2:F3}$ population. Blue parentheses show the names of the regions supporting the merged QTL. The number of markers has been decreased to facilitate visualization.

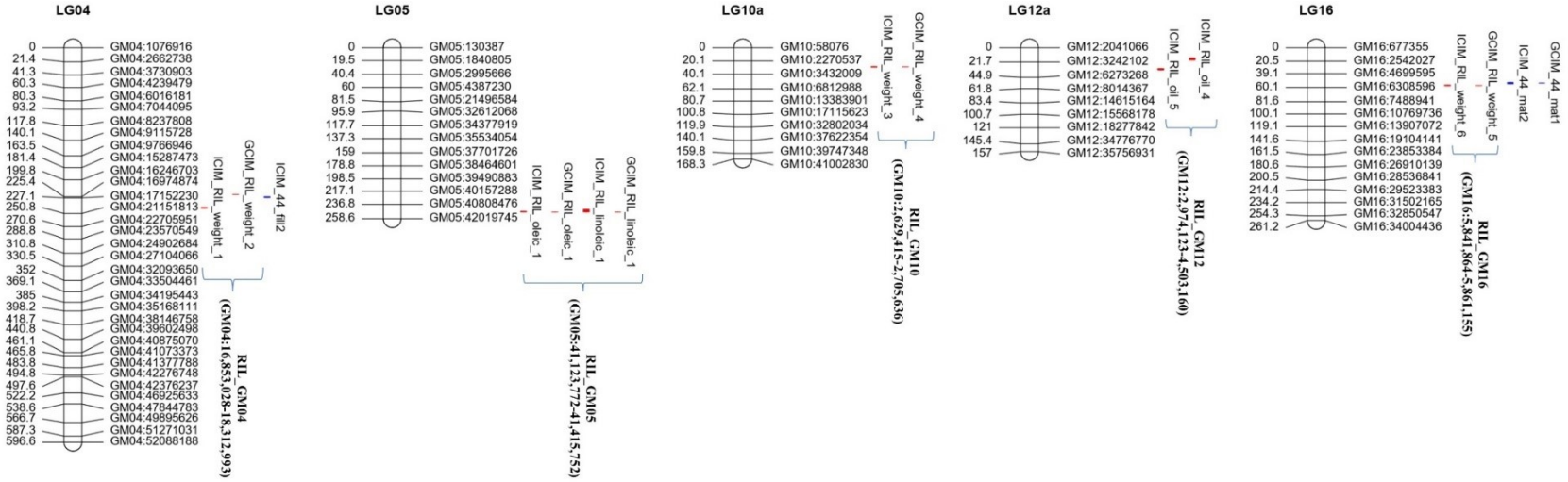

Figure 4.9 Quantitative trait loci specific to the QS15544$_{RIL}$ population. Red-marked traits indicate the seed quality QTL identified in this study, whereas the blue-marked traits indicate overlapping QTL associated with the reproductive traits detected in Gélinas Bélanger *et al.* (2024). Blue parentheses show the names of the regions supporting the merged QTL. The number of markers has been decreased to facilitate visualization.

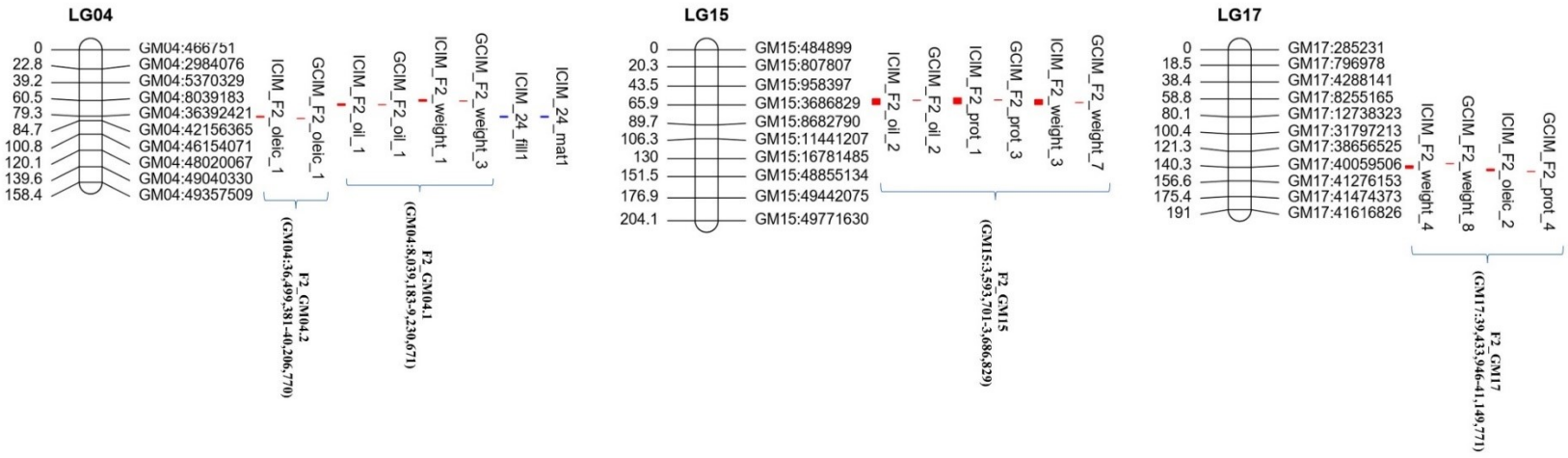

Figure 4.10 Quantitative trait loci specific to the QS15524$_{F2:F3}$ population. Red-marked traits indicate the seed quality QTL identified in this study, whereas the blue-marked traits indicate overlapping QTL associated with the reproductive traits detected in Gélinas Bélanger *et al.* (2024). Blue parentheses show the names of the regions supporting the merged QTL. The number of markers has been decreased to facilitate visualization.

5.9 Connecting text

Chapter 5 demonstrated that a few key regions regulate several critical seed quality traits and that three of these regions are in close linkage with major reproductive loci identified in Chapter 3. Amongst the QTL found in this study, three were novel and three were also found to be overlapping GWA signals from the MadMaturity$_{86}$ population for the seed weight trait. As such, this chapter proposes a catalog of SNPs for six quality traits regarding two early-maturing soybean populations that can be used as such by Canadian breeders desiring to expand their allele catalog. The next chapter discusses the impacts of the findings of Chapters 3, 4, and 5 on the crop improvement of early-maturing accessions and the low prevailing diversity in the germplasm of cultivars adapted to MG00 and MG000. In complement, Chapter 6 addresses potential future avenues for the QTL identified in the three aforementioned chapters, including their validation using genome editing and *in planta* transformation procedures.

6. From Prediction to Validation: A Roadmap to Breed Early-Maturing Soybean Cultivars Adapted to MG00, MG000, and Beyond.

Jérôme Gélinas Bélanger[1,2]

[1]CÉROM, Centre de recherche sur les grains Inc., St-Mathieu-de-Beloeil, Québec, Canada

[2]Department of Plant Science, McGill University, Montréal, Québec, Canada

Segments of this manuscript were reproduced from Plant Methods

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., Charron, J.-B., and O'Donoughue, L. (2024). A comprehensive review of in planta stable transformation strategies. ***Plant Methods*** 20, 79. doi: 10.1186/s13007-024-01200-8

6.1 Abstract

Soybean is a short-day crop that has been adapted to northern agricultural conditions through selective breeding and domestication. Multiple scientific studies have demonstrated that adaptation to high latitudinal conditions requires the stacking of multiple recessive alleles for the major *E1*, *E2*, *E3*, and *E4* loci, leaving breeders with a limited set of options. The mapping of novel regions regulating reproductive traits (Chapter 3), gene expression (Chapter 4), and seed quality traits (Chapter 5) opens new possibilities to develop early-maturing soybean cultivars with good seed-quality traits. In this chapter, I discuss the importance of these findings in regard to the lack of prevailing diversity in the early-maturing soybean germplasm and propose new avenues for breeders to integrate these alleles in their SNP catalogs, such as the implementation of genomic prediction and *in planta* transformation procedures in combination with RNA-guided endonucleases-based genome editing technologies (e.g., Cas9, Mad7, or other enzymes). Altogether, I propose a streamlined approach to validate the candidate genes discovered in Chapters 3-5 and propel the development of high-quality and high-yielding cultivars belonging to MG00, MG000, and beyond.

6.2 Introduction

Soybean is an interesting crop for farmers because of its high nutritional value, high yield, and nitrogen fixation capacities. Despite these benefits, the history of soybean cultivation in Canada is very recent. The earliest records of soybean cultivation in the country date back to 1893, but commercial production only began in the mid-1970s in Ontario with the development of early-maturing and cold-tolerant cultivars (Canadian Food Inspection Agency, 2021). In the mid-1990s, these varieties were more broadly adopted when Quebec and Manitoba started the cultivation of the crop (Canadian Food Inspection Agency, 2021). Consequently, Canada's soybean production now spans areas with maturity groups ranging from MG000 to MGIII, meaning that significant portions of Ontario, Quebec, and Manitoba can integrate soybeans into their crop rotations (Bagg *et al.*, 2002). Still, large portions of agricultural lands, located mainly in Saskatchewan, Alberta, and the northern parts of Manitoba, Quebec, and Ontario, remain largely unsuitable for soybean cultivation. Expanding a crop beyond its natural agroecological limits is a challenging process, not only because of the intricate complexity of such an endeavor from a biological standpoint but also because of the inherent requirements of developing genetic material that is commercially

acceptable. In the northernmost growing areas of Canada, the short summers limit the choice of crops available to farmers because of the negative correlation between yield and maturity (Jean *et al.*, 2021). Longer day lengths during the summers also limit the latitudinal expansion of short-day crops as the reproductive phenology of genetically unadapted cultivars does not allow for an optimal match between the plant's life cycle and the surrounding environmental conditions, meaning that a crop is often not mature enough to be harvested before the first frost kill. Soybean is no exception to both of these rules (i.e., negative correlation between yield and unadapted reproductive phenology) as a plethora of studies have confirmed the challenges of reaching commercially sustainable yields in remote northern regions. As it currently is, research using genomic prediction demonstrates that it is almost impossible to predict and generate soybean crosses with a near-zero or opposite correlation between yield and maturity (Jean *et al.*, 2021).

Moreover, soybean's expansion in northern areas is also hampered by the lack of high-yielding elite cultivars with good seed quality traits (e.g., protein content, oil content, and fatty acids) that also harbor critical disease and pest resistance genes. This paucity in the early-maturing germplasm ultimately leads to reduced selection options for breeders seeking to push the development of the crop in areas where the supply chains for other high-yielding field crops (e.g., wheat) are well-established. Ultimately, this smaller allelic variant catalog decreases the capacity of breeders to develop cultivars for specific market classes. In addition to these factors, multiple biological and genetic intricacies, such as linkage drags, potential pleiotropic effects, and/or epistatic interactions between regions of interest, limit the options for breeders. As a result, the genetic structures and mechanisms guiding the response to the surrounding environmental conditions (e.g., short summers with long days) are intertwined and hard to decipher, thus complexifying the work of plant scientists and breeders.

The initial hypothesis of this thesis stated that "a limited number of key genes and molecular mechanisms regulate the reproductive and seed quality traits in early-maturing Canadian soybean accessions". With the identification of only a few key loci regulating early reproductive traits in Chapter 3, but several small-effects regions with lower LOD and PVE, we demonstrated that this hypothesis is partially true. This observation was further confirmed with the identification of a handful of key regions regulating three candidate genes for the *E8-r3* region in Chapter 4. Furthermore, Chapter 5 demonstrated that a limited number of key regions regulate seed quality traits in each of the studied populations, with a few of them possibly subject to linkage drag or

pleiotropic events between loci regulating both seed quality (e.g., seed weight) and reproductive traits. In a similar fashion to Chapters 3 and 4, several additional minor regions for quality traits were identified in each of the populations, thus reinforcing the idea that my initial hypothesis was partially true. As a whole, these three chapters demonstrate the need to properly understand the limited sources of variation in the available germplasm for optimal breeding.

6.3 Lower Diversity, Better Understanding

The genetic diversity in the Canadian-bred soybean germplasm is known to be much lower than exotic gene pools due to a broad variety of historical (e.g., late introduction of soybean into the Canadian cropping systems and early immigration mainly from European descent), cultural (e.g., traditional Canadian diet does not integrate soybean), and agro-environmental (e.g., short summer seasons and long-day photoperiodism) constraints (Fu *et al.*, 2007; Iquira *et al.*, 2010). To compensate, breeders have incorporated new accessions into their programs to increase allelic diversity and introduce new traits (Bruce *et al.*, 2019). The highest pool of soybean genetic diversity is found in the Huanghe region of China (i.e., a region spanning soybean accessions classified as MG I to VI), an area suggested to be the origin of soybean domestication (Dong *et al.*, 2004; Li *et al.*, 2010). Researchers investigating the genetic structure of soybean landraces in China have demonstrated that allelic diversity for the major early maturity genes (*E1*, *E2*, *E3*, and *E4*) is dominated by a large presence of recessive loss-of-function variants in the peripheral MGs 00, 000 and 0000 regions (Liu *et al.*, 2020). The numbers speak for themselves as all of the *E1* alleles found for the Chinese landraces belonging to the MGs 000 and 0000 were recessive mutants (Liu *et al.*, 2020). In addition, only a small proportion (approximately 10 %) of the MGs 000 and 0000 landraces displayed the dominant form of the *E3* gene (Liu *et al.*, 2020). For *E4*, most of the super early landraces harbored either the *e4-kes* or *e4-SORE* photoperiod insensitivity-promoting alleles (Liu *et al.*, 2020). As a comparison, similar genotypic patterns with mostly recessive alleles were identified in the QS15524$_{F2:F3}$ (*e1-nl*/*e2-ns*/*E3Ha*/*e4-SORE-1*) and QS15544$_{RIL}$ (*e1-as*/*e2-ns*/*e3-tr*/*e4p.T832QfsX21*) populations used in Chapters 3, 4 and 5. The omnipresence of recessive alleles at the four major *E* genes reduces the flexibility and options for breeders as these genes can explain more than 60 % of the variation in the observed flowering time according to Liu *et al.* (2008) and Xia (2013). Interestingly, Jia *et al.* (2014) recently demonstrated that the 'Sunset' soybean cultivar (i.e., a variety belonging to the putative super early-maturity group MG0000) harbors the

e1/*E2*/*E3*/*E4* allelic combination, thus suggesting that other novel loci might be important regulators of the maturity process.

Due to the large influence of *E1* to *E4* on the observed phenotype, the populations that were used in Chapters 3, 4, and 5 were designed in such a way that both had fixed alleles for *E1-E4.* This approach was chosen to limit the background noise caused by these important genes and identify novel regions involved in reproduction (Chapter 3), the regulation of gene expression for candidates of the *E8-r3* locus (Chapter 4), and seed quality (Chapter 5), that could be harnessed for the breeding of a new soybean ideotype that could be cultivated beyond MG000. In Chapter 3, a total of 17 regions (i.e., 7 major and 10 minor) regulating three reproductive phenotypes (i.e., days to pod-filling, days to maturity in the field, and days to maturity in the greenhouse) were identified. Three of these major regions were identified in both populations (*E8-r1*, 1.81 days; *E8-r2,* 1.27-3.85 days; and *E8-r3*, 2.07-3.73 days), whereas the four others (GM04:16,974,874-17,152,230, 1.81 days; GM07:5,256,305-5,4049,71, 1.15-1.30 days; GM16:5,680,173-5,730,237, 1.51-1.55 days; and GM16:22,756,017-23,154,638, 1.12-1.35 days) were found only in $QS15544_{RIL}$. Several candidate genes were identified using a five-step variant analysis pipeline with *Glyma.04G124300*/*PROTEIN FAR-RED ELONGATED HYPOCOTYL 3* (*E8-r1*), *Glyma.04G156400*/*E1-like-a* (*E8-r2*), *Glyma.04G167900*/*LIGHT-HARVESTING CHLOROPHYLL-PROTEIN COMPLEX I SUBUNIT A4* (*E8-r3*), *Glyma.04G166300*/*PSEUDO–RESPONSE REGULATOR 1a* (*E8-r3*), *Glyma.04G159300*/*MADS-BOX DOWNREGULATED BY E1 04* (*E8-r3*) and *Glyma.04G168300*/*CYCLING DOF FACTOR 3* (*E8-r3*) being the best candidates for the *E8-r1*, *E8-r2*, and *E8-r3* overlapping regions.

Recent discoveries have identified and confirmed the role of *Glyma.04G156400* in the regulation of flowering time and maturity for the *Tof4* (*Time of flowering 4*) locus found in a RIL population derived from a cross between 'Dongnong50' and 'Williams 82' (Dong *et al.*, 2023). In their paper, Dong *et al.* (2023) identified an A→G single nucleotide polymorphism leading to a detrimental lysine-to-glutamate amino acid change in *Glyma.04G156400*. In Chapter 3, we screened the different variants in *Glyma.04G156400* using the GmHapMap dataset but only identified a mutation in the 3'UTR region, thus meaning that two different alleles could be currently available for this gene for breeding. The *Tof4* locus encodes the E1-like protein, E1La, which binds to the promoters of *FT2a*, *FT5a*, and *Tof5* (i.e., *Glyma.05G018800*/*MDE05*, a gene investigated in Chapter 4 also known as *GmFUL*) and inhibits their transcription under long days (Dong *et al.*,

2023). Ten additional minor regions were identified in either $QS15524_{F2:F3}$ (7 regions) or $QS15544_{RIL}$ (3 regions), with LOD scores and PVE ranging from 3.54 to 12.69 and 0.75 to 5.39%, respectively. Although it is impossible to currently confirm this affirmation, we think that it might be possible to harness these large and small-effect QTL by stacking them in such a way that the generated offspring will be earlier-maturing than their parents. Results obtained with various varieties belonging to the putative super early-maturity group MG0000 suggest it is possible to push early-maturity further (Jia *et al.*, 2014; Jiang *et al.*, 2019; Gupta *et al.*, 2022).

6.4 Unravelling the Transcriptional Landscape Associated With Physiological Maturity

To support our understanding of the intricate mechanisms regulating maturity (e.g., pleiotropism and TF co-regulation) in these two soybean populations, we generated an atlas of eQTL traits in Chapter 4 using leaf tissue that was collected from the V4 leaflet 4 hours after sunrise. Gene expression is regulated in *cis* (i.e., locally) or in *trans* (i.e., distantly) by TFs that bind with their DNA-binding domain to specific *cis*-regulatory elements (i.e., specific motifs) located in upstream, intron, or downstream regions of their targets to either activate or repress transcription (Vinson *et al.*, 2011; Mitsis *et al.*, 2020). Once bound, TFs recruit RNA polymerase II and interact with coactivators (i.e., other bound TFs) (Näär *et al.*, 2001). Understanding transcriptional dynamics can lead to the identification of master regulators triggering major developmental processes (e.g., flowering and maturity) that play critical roles in the architecture of agronomic features (e.g., yield and number of days to maturity) (Kaufmann and Airoldi, 2018). Although unraveling these complex transcriptional dynamics is crucial from a fundamental standpoint, the building of a compendium of TF interactions is a challenging endeavor as transcriptional interactions happen in the hundreds and thousands, with a large number of key players concomitantly acting on the same targets.

Experimentally, several biological factors represent additional challenges to the mapping of these interactions, the most prominent probably being the size of the studied genome. The soybean genome has a high structural redundancy due to past duplication events that happened approximately 13 and 59 million years ago which led to an increase in the number of homologous genes with redundant biological functions (Shultz *et al.*, 2006; Swarbreck *et al.*, 2008; Schmutz *et al.*, 2010). As detailed in Chapter 2, we predicted more genes coding for TFs involved in flowering

and/or maturity in soybean (i.e., 269) than *Arabidopsis* (i.e., 70), thus suggesting that mapping the downstream interactions in the former might be more challenging than the latter. To identify the putative TFs involved in the regulation transcriptional landscape of our two plant populations, we developed a custom pipeline divided into four main phases: (i) identifying eQTL interactions at a transcriptome-wide scale using three mapping algorithms; (ii) detecting hotspots associated with FRSPD function; (iii) finding gene pairs, including putative TFs, that were strongly co-expressed together using positive and negative CENs; and (iv) predicting putative TFs and their associated variants that were located within or near the hotspots.

Using this pipeline, we mapped a total of 2,218 trans (2,061 genes) / 7 cis (7 genes) in $QS15524_{F2:F3}$ and 4,073 trans (2,842 genes) / 3,083 cis (2,418 genes) interactions in $QS15544_{RIL}$. Following the mapping of the *cis* and *trans* interactions, we identified four hotspots (RIL_GM04:10,812,813-10,985,437, F2_GM06:39,892,719-43,437,125, F2_GM17:5,431,473-7,260,313, and F2_GM18:1,434,182-1,935,386) associated with FRSPD functions. In addition to this, we found that the F2_GM18:1,434,182-1,935,386 hotspot was regulating *Glyma.04G167900* (*GmLHCA4*) and five additional *LHCA* homologs that exhibit a role in the response to light. In Liu *et al.* (2021), a 1.8-day difference in the number of days to flowering was observed between two *GmLHCA4* haplotypes, thus suggesting that this gene might be the regulator found in this region. In both populations, several interactions with multiple *GmPRR* and *GmMDE* homologs were also successfully mapped, including one with the F2_GM15:49,385,092-49,442,237 hotspot. Using a custom variant analysis pipeline, we identified that *ALTERED PHLOEM DEVELOPMENT* and *DOMAIN-CONTAINING PROTEIN 2* as the best candidate TFs for the F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386 hotspots. To the best of my knowledge, the mapping of these interactions has not been previously reported and thus might serve as an interesting tool to better understand the regulation schemes for the candidate genes of *E8-r3*.

At present, few studies have used eQTL mapping to unravel the various interactions underlying reproductive traits. Chien *et al.* (2023) uncovered a hotspot in *Arabidopsis* regulating the flowering time trait and which covered the gene body of AT4G00650 (*FRIGIDA*). In the same research, it has been demonstrated that *FRIGIDA* harbors multiple haplotypes that differentially affect the expression of downstream genes involved in the regulation of flowering (e.g., *SUPPRESSOR OF OVEREXPRESSION OF CO 1* and *FLOWERING LOCUS C*). In a similar fashion, expression QTL mapping in a F_2 population of *Brassica napus* identified *FLOWERING*

LOCUS C as a candidate for regulating flowering time and their results suggest that this gene might be a master regulator controlling the expression of several downstream genes (Li *et al.*, 2018). In soybean, only a few eQTL studies were performed either in biparental (Bolon *et al.*, 2014) or association populations (Li *et al.*, 2021a, 2023a; Qin *et al.*, 2023b; Yuan *et al.*, 2024). To the best of our knowledge, this chapter is unique in that it is the only study currently available that tries to identify eQTL regions associated with the maturity trait in soybean. However, this novelty aspect limits our capacity to cross-compare our results with other datasets.

6.5 Breeding High-Quality Cultivars on the Frontier

Soybean breeding is partly driven by the development of cultivars bred for specific end-uses, such as the industrial, feed-grade, or food-grade (e.g., tofu, miso, natto, soy sauce, and soy beverage) markets (Soy Canada, 2022). Each market requires soybean lines that have specific profiles regarding their quality traits, whether it be for protein content, lipid quality, or grain size (Soy Canada, 2022). As such, the lack of prevailing diversity in the early-maturing germplasm limits the selection options for breeders to create cultivars with specific quality features. In Chapter 5, we investigated six different quality traits (100-seed weight, protein content, oil content, oleic acid content, linoleic acid content, and linolenic acid content) and identified a total of 12 major regions in $QS15524_{F2:F3}$ (four loci), $QS15544_{RIL}$ (five loci), or both (three loci) populations. During our analysis, we found that three of these regions (RIL_GM12, RIL_GM16, and F2_GM04.2) (out of 12) were not previously identified in the literature. Altogether, these loci might serve as novel sources of regulation for the oil content, seed weight, and oleic acid content, respectively. In a similar fashion to Chapters 3 and 4, we detected several small-effect QTL in $QS15524_{F2:F3}$ (9 regions) and $QS15544_{RIL}$ (14 regions) associated with the regulation of at least one of the aforementioned quality traits. Three other QTL regions (F2_GM04.1, Merg_GM06, and Merg_GM19) found in Chapter 5 were also previously identified by our group in a genome-wide association panel comprising 86 early-maturing soybean accessions (Copley *et al.*, 2018), thus reinforcing the quality of our results for both studies and demonstrating that these allelic variants are available for breeders.

To further improve the breeding of cultivars with specific seed quality features, two complementary approaches can be used by breeders. In the first approach, breeders can profile the offspring generated from specific crosses using marker-assisted selection for quality traits. In the

second approach, genomic prediction can be used to predict optimal crosses using a model based on ‘‘a large panel’’ (given that germplasm diversity in early-maturing lines is limited) of lines. Genomic prediction aims to predict the phenotypic performance of offspring using a training set that has been both phenotyped and genotyped with a set of genome-wide markers (Keller *et al.*, 2020; Jean *et al.*, 2021). Using genomic prediction, breeders can predict the performance of progeny lines from biparental crosses based on the *in silico* genomic estimated breeding values for the traits of interest (Miller *et al.*, 2023). From a practical standpoint, this means that an almost infinite number of crosses can be estimated to obtain an optimal allelic combination for early-maturity and yield-regulating genes – granted the use of an appropriate training set. Current research demonstrates that genomic prediction can be used to generate efficient predictions in soybean in general (Beche *et al.*, 2021; Miller *et al.*, 2023), but also more specifically for high-yielding early-maturing soybean cultivars (Jean *et al.*, 2021). For example, Jean *et al.* (2021) demonstrated that it is possible to identify crosses demonstrating high commercial potential (aka cultivars with high yield potential based on their respective MG) in early-maturing MGs as most of their predicted crosses (99 out of 101) were retained by breeders for advanced trials or commercialization. Furthermore, most of the crosses (i.e., 96.2 %) with a predicted low yield were eliminated during selection by breeders, thus demonstrating that the genomic prediction approach can successfully identify the best crosses in a breeding program (Jean *et al.*, 2021). Still, the researchers underlined that some traits, such as maturity, could be underestimated in early-maturing progeny, mainly because *E4* is a major maturity gene in Canadian lines but has no SNP marker tightly associated with it (Tardivel *et al.*, 2019).

For seed quality traits, Stewart-Brown *et al.* (2019) have demonstrated that predictive abilities of 0.81, 0.71, and 0.26 for protein content, oil content, and yield could be reached in genomic prediction trials using 483 elite soybean breeding lines, thus suggesting that these high-performing results for protein and oil contents could be emulated in an early-maturity breeding program. Similarly, prediction accuracies ranging between 0.40 and 0.83 were obtained for five traits (i.e., seed size, protein content, yield, test weight, and ergosterol content) using a panel comprising 309 advanced barley breeding lines (Nielsen *et al.*, 2016). Although it might be almost impossible to fully disentangle the negative correlation between yield and maturity, these papers suggest that genomic prediction can support good decision-making in breeding as a means to compensate for these undesirable correlations. Additionally, these articles show that the genomic

prediction of seed quality traits is highly accurate - given a large training set - and can be used in early-maturity breeding.

6.6 Pushing the Boundaries with Genome Editing

Before the advent of genetic engineering technologies, conventional breeding methods were the sole avenue to undergo agricultural plant improvement. To do so, breeders needed to cross and select the best-performing plants to achieve a relatively slow, but steady and reliable progress (Raina *et al.*, 2017). With the first development of plant genetic engineering in 1983 (Bevan *et al.*, 1983), scientists are now able to insert specific genetic sequences into plants' genomes either from the same species (a cisgenic modification) or from unprecedented sources such as other species, genera, or beyond (a transgenic modification) (Holme *et al.*, 2013). This breakthrough allows scientists to potentially shortcut the breeding process by directly inserting genetic material leading to improved field phenotypes (e.g., pest resistance) and novel traits (e.g., herbicide resistance) unseen before in cultivated plants. Genome editing with CRISPR endonucleases enzymes is the newest improvement in the realm of plant breeding and should become a mainstay because of its high-efficiency rate, low cost, and high customization due to modular cloning plant parts kits based on Golden Gate technology (Engler *et al.*, 2014; Hahn *et al.*, 2019; Bao *et al.*, 2020; Xu *et al.*, 2020). The first reports of successful genome editing in soybean were published in 2015 (Jacobs *et al.*, 2015) and a plethora of soybean genes controlling a wide variety of phenotypes, including flowering and maturity (Table 5.1), have been modified using different CRISPR-Cas systems since then. In addition to the reports in soybean, CRISPR-Cas systems have been used to induce mutagenesis in key orthologous genes involved in flowering in various monocots (e.g., rice) and dicots (e.g., *Arabidopsis* and tomato) species (Table 5.2).

Most often, genome editing experiments target the coding sequence of a gene to generate a knockout line and subsequently validate its specific biological function. For breeding, it might be valuable to target other types of sequences (e.g., promoter, 3'UTR, or 5'UTR) to generate offspring that exhibit a gradient of phenotypes. Using a multiplexed CRISPR-Cas9 promoter targeting approach, Rodríguez-Leal *et al.* (2017) demonstrated that it was possible to generate T_0 tomato transformants that can be used to create an F_1 population exhibiting a range of phenotypic variation for the number of inflorescence trait. Another interesting approach for breeders is to generate a multiplex mutagenesis population via pooled CRISPR-Cas9 as demonstrated by Bai *et al.* (2020).

In this article, the researchers constructed 70 CRISPR-Cas9 vectors targeting 102 candidate genes and generated a population exhibiting a large phenotypic variation (Bai *et al.*, 2020). Based on these observations, we think that these approaches could be combined (i.e., the creation of populations exhibiting mutations located in their promoter regions using a multiplex approach) to create a large range of novel alleles for the candidate genes identified in Chapter 3, 4 and 5 (e.g., *GmLCH4A*, *GmPRR1a*, and *GmCDF3*). Subsequently, this population could be incorporated into a conventional breeding program or used for functional validation. Ideally, such a strategy would require a strong transformation system to propel the creation of this novel germplasm on an extensive scale.

6.7 A Step-Forward with *In Planta* Transformation

Efficient stable plant transformation systems fulfill three basic requirements: i) a source of totipotent cells serving as recipients for the delivered DNA; ii) a means to deliver the DNA into the targeted cells; and (iii) a selection system to identify the transformed cells (Somers *et al.*, 2003). Soybean transformation, because of its low regeneration rates from calli, is notoriously challenging to perform and is considered one of the major bottlenecks that limit the understanding of fundamental biological processes and plant improvement in this crop. Over the years, several *in vitro* and *in vivo* methods have been developed to perform soybean transformation; however, the vast majority of the articles about soybean transformation use the indirect regeneration pathway under *in vitro* conditions. This strategy aims at producing shoots from adventitious clumps or embryogenic calli infected with *Agrobacterium tumefaciens* (Lee, 2013). At present, multiple iterations of this method are currently available in the literature, with a large proportion of the publications about this approach proposing new growth medium recipes to obtain higher regeneration rates (Lee, 2013). Still, the efficiency of this technique remains low as transformation rates ranging from 2.5 to 10 % are the norm (Song *et al.*, 2013). Furthermore, this strategy is especially slow to perform, taking approximately 9 months from the beginning (i.e., the transformation event) to the harvest of the T1 seeds (Luth *et al.*, 2015). In addition to the *Agrobacterium* technique, several techniques based on the *in vitro* indirect regeneration pathway have also been demonstrated to be feasible, albeit challenging, and include: (i) direct delivery of DNA in callus tissues using biolistics (Christou *et al.*, 1989; Simmonds and Canada, 2003; Rech *et al.*, 2008); (ii) ribonucleoprotein delivery in callus-derived protoplasts (Kim *et al.*, 2017; Kim

and Choi, 2021); and (iii) plasmid electroporation into protoplasts (Christou *et al.*, 1987; Dhir *et al.*, 1992). On the whole, the importance of these different *in vitro* techniques based on the indirect regeneration pathway remains modest. For most laboratories, the *in vitro* indirect regeneration route is still a major impediment to large-scale transformation experiments as the most successful labs often generate only a handful of mutants per construct using the classic techniques.

During the past few years, several laboratories have developed *in vitro* direct regeneration and *in planta* strategies to overcome the issues associated with *in vitro* indirect (i.e., callus-based) regeneration. *In planta* stable transformation, also called *in situ* transformation, techniques form a heterogeneous group of methods all aiming at performing the direct and stable integration of foreign T-DNA into a plant's genome (Gélinas Bélanger *et al.*, 2024). By definition, *in planta* transformation techniques must have none or minimal tissue culture steps (Gélinas Bélanger *et al.*, 2024). To be considered minimal, the tissue culture steps should meet these four following pivotal criteria: (i) use simple technical methods that are simple to master and reproduce; (ii) be genotype-independent or highly compatible with a high number of genotypes; (iii) regeneration is performed using a differentiated explant that does not undergo a callus development stage and thus relies on direct regeneration; and (iv) exhibit a short regeneration period with a limited number of medium transfers (Gélinas Bélanger *et al.*, 2024). Highly diversified, *in planta* techniques target a variety of organs, such as germ or vegetative cells. Recently, our lab performed a large-scale literature review with more than 300 references encompassing all species on the topic of *in planta* transformation (Gélinas Bélanger *et al.*, 2024). Based on the findings of this literature review, I argue that this route needs to be considered to increase the scale and speed of soybean transformation. In this review, we identified 75 references using 12 different types of *in planta* transformation methods for *Glycine max* and 11 other annual leguminous species that are highly similar to soybean at the morphological level (Fig. 5.1; Appendix 4.1). Amongst these references, we found a large number of highly promising protocols for soybean (Chee *et al.*, 1989; Gao *et al.*, 2007; Liu *et al.*, 2009; Zia *et al.*, 2011; Mily *et al.*, 2020), but also chickpea (Reddy *et al.*, 2007; Sreevathsa *et al.*, 2008; Asharani, B. M, 2011; Ganguly *et al.*, 2020a; Shriti *et al.*, 2023), pigeon pea (Rao *et al.*, 2008; Ramu *et al.*, 2012; Kaur *et al.*, 2016; Ganguly *et al.*, 2018, 2020b; Singh *et al.*, 2021), peanut (Rohini and Rao, 2000; Rohini and Sankara Rao, 2000; Zhai *et al.*, 2022; Zhou *et al.*, 2023), and common pea (Švábová *et al.*, 2005; Svabova *et al.*, 2007), thus suggesting that *in planta* transformation is feasible in these recalcitrant crops. On the whole, we identified four *in*

planta transformation techniques that we think could be more easily implemented in a large number of settings (i.e., small vs. large; private companies vs. academic institutions vs. non-profit research organizations): (i) agro-infection of imbibed embryos (Chee *et al.*, 1989; De Ronde *et al.*, 2001; Švábová *et al.*, 2005; Svabova *et al.*, 2007; Liu *et al.*, 2013); (ii) shoot apical meristem injury under *in vivo* conditions (Rohini and Rao, 2000; Entoori *et al.*, 2008; Sreevathsa *et al.*, 2008; Sundaresha *et al.*, 2010; Karthik *et al.*, 2018; Singh *et al.*, 2018); (iii) imbibition of dehydrated embryo (Arias *et al.*, 2003); and (iv) the direct regeneration of embryonic axis under *in vitro* conditions (Liu *et al.*, 2004; Rech *et al.*, 2008; Zhang *et al.*, 2014, 2016a; Paes de Melo *et al.*, 2020; Cho *et al.*, 2022). Altogether, these *in planta* transformation methods represent promising alternatives to circumvent the roadblocks associated with the conventional approach using callus-based *in vitro* regeneration and to validate the candidate genes identified in Chapters 3 to 5.

6.8 Conclusion

Historically speaking, soybean is a newcomer in Canada's agricultural production, but this crop is now firmly established in Ontario, Quebec and Manitoba's crop rotation. As a consequence of its high popularity, soybean production is deemed to increase northward into Canada's MG00 and MG000 cultivation zones, and possibly beyond. To propel this expansion in a timely manner, knowledge is key. Breeders need to have access to a compendium of genetic regions influencing both reproductive and seed quality traits to carefully design and plan their selection schemes. To improve the efficiency of their selection programs, breeders need to consider novel technological avenues (e.g., marker-assisted selection, genomic prediction, and stable *in planta* genome editing) to facilitate the introgression of these alleles. In this thesis, I identified several candidates involved in the regulation of a plethora of functions, including reproductive (Chapter 3), transcriptional regulation of candidates for the *E8-r3* locus (Chapter 4), and seed quality (Chapter 5) traits, in early-maturity backgrounds. In my opinion, the results generated in these three studies demonstrate that the parental lines used in these studies display the genetic variation required for breeders to select early-maturing cultivars with good seed quality. To fully harness the potential of these alleles, the next logical step is to perform the functional validation of each of these candidates through the generation of knockout mutant lines. To achieve this, I presented in this chapter several novel research articles disentangling the recalcitrance paradigm associated with soybean transformation and which are based on the *in planta* concept. Altogether, I believe that several *in*

planta methods need to be considered as a means to validate the candidate genes discovered in Chapters 3-5 and propel the development of high-quality and high-yielding cultivars adapted for MG00, MG000, and beyond.

6.9 Supplemental data

Appendix 4.1 References for the *in planta* transformation approaches for legumes.

Cultivar*	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
Jack	*E1*	B3 domain TF/regulation of photoperiodic flowering through *FT2a/FT5a*	11 and 40-bp deletions resulted in a frameshift mutation that produced premature translation	Early flowering under LD	Han *et al.* (2019)
Jack	*GmFT2a*	Homolog of *AtFT*/ florigen, flowering induction	InDels located in two regions resulting in frameshift mutations	Late flowering both under SD and LD	Cai *et al.* (2018)
Jack	*GmFT2a* *GmFT5a*	Homolog of *AtFT*/ florigen, flowering induction	*ft2a* has 1-bp insertion *ft5a* has 2-bp insertion	*ft2aft5a* double mutants showed late flowering (≈ 31.3 days) under SD conditions	Cai *et al.* (2020b)
Jack	*GmFT3b*	Homolog of *AtFT*/ florigen, flowering induction	*ft3b* has a 72-bp deletion	No significant differences in flowering time between the wild-type, the *FT3b* overexpressors, and the *ft3b* knockouts	Su *et al.* (2022)
Jack	*GmFT5b*	Homolog of *AtFT*/ florigen, flowering induction	Short deletions resulting in a frameshift mutation predicted to generate premature translation	Delayed flowering only under LD conditions	Cai *et al.*, 2020b)

Cultivar*	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
Jack	*GmFT2a* *GmFT4*	Homolog of *AtFT*/ florigen, flowering induction	C → G base substitutions at pos7	Late flowering (≈ 34.3 days) under SD conditions	Cai *et al.* (2020a)
HX3	*GmAP1a/b/c/d*	Homolog of *AtAP1* and is MADS-box TF / Floral meristem identity genes	5-bp deletion and 2-bp deletion in *ap1a* 3-bp deletion and 1-bp insertion in *ap1b* 8-bp deletion in *ap1c* 39-bp deletion in *ap1d*	*ap1* quadruple mutant exhibited delayed flowering and changes in flower morphology	Chen *et al.* (2020)
Harosoy	*GmLHY1a/*b *GmLHY2a/*b	Homologs of *AtLHY/AtCCA1* / Key components of the central oscillator encoding MYB TFs	2-bp deletion in *lhy2b/2a/1b* 1-bp deletion in *lhy1a* (Cheng *et al.*, 2019) 1-bp and 813-bp deletions *lhy2b* 2-bp and 1-bp insertions *lhy1b* 1-bp and 4-bp deletions in *lhy1a*	*lhy* quadruple mutants exhibited a reduced plant height, shortened internodes, and significant reduction in endogenous gibberellic acid (GA3) (Cheng *et al.*, 2019) *lhy* quadruple mutants exhibited delayed flowering under LD (Lu *et al.*, 2020)	Cheng *et al.* (2019); Lu *et al.* (2020)

Cultivar*	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
			1-bp and 10-bp deletions in *lhy2a*		
Wm82	*GmLNK2a/b/c/d*	Coactivators of dawn-phased MYB-like TFs, such as *AtRVE4* and *AtRVE8* / Regulate the expression of *GmTOC1*, *GmPRR5*, *GmE1* and *GmFT2a*	35-bp deletion at *lnk2a* 10-bp deletion at *lnk2b* 5- bp deletion at *lnk2c* 1-bp deletion at both *lnk2d-sg2* / *lnk2d-sg3*	*lnk2* quadruple mutants flower significantly earlier under SD	Li *et al.* (2021b)
Wm82 or Harosoy	*GmLCLa1/a2/b1/b2*	Myb-domain protein required for the maintenance of the circadian clock rhythm	1-bp and 813-bp deletions at *lcla1* 2-bp deletion and C → A mutation at *lcla2* 1-bp and 12-bp deletions at *lclb1* 1-bp and 4-bp deletions at *lclb2*	Quadruple mutants have an extremely short-period circadian rhythm and late flowering phenotype	Wang *et al.* (2020c)

Cultivar*	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
ZGDD and Jack	*GmPRR37*	Encode homeologous PRRs containing a pseudo-receiver domain	1-bp deletion in *prr37* (ZGDD) 1-bp deletion in *prr37* (Jack)	Early flowering (≈15.8 days) under LD and no effect in SD (ZGDD) Same as WT under LD and SD	Wang *et al.* (2020a)
Wm82	*GmPRR3b*	Encode homeologous PRRs containing a pseudo-receiver domain/	Small deletions causing a frameshift resulting in a premature termination	*prr3b* mutants had delayed growth and floral transition	Li *et al.* (2020)

*Maturity groups for the selected cultivars are Harosoy, MGII; Jack, MGII; Wm82, MGIII; and ZGDD, MGVII

Table 5.1 Summary of selected publications using CRISPR-based systems to induce mutagenesis in soybean flowering genes.

Species	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
A. thaliana	*AtFT*	PEPB protein/Florigen, flowering induction	Small InDels	Late flowering when compared to WT	Hyun *et al.* (2014)
A. thaliana	*AtAP1*	MADS-box TF / Floral meristem identity genes	Multiple mutants with small InDels	Abnormal floral meristem development abnormally with an increased number of petals, a degenerated number of sepals and petals, and an increase in plant branching	Liu *et al.* (2019)
A. thaliana	*AtSVP*	MADS-domain TF / Flowering repressor that determines inflorescence architecture	Multiple mutants with small InDels	Early floral meristem formation during the vegetative phase, but the inflorescence was aborted	Liu *et al.* (2019)
A. thaliana	*AtTFL1*	Repressor of floral initiation regulates indeterminate conversion to a determinate architecture and	Multiple mutants with small InDels	Partial transformation of the stem meristem into a floral meristem The transition from an indeterminate type of flowering to a determinate	Liu *et al.* (2019)

Species	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
		controls inflorescence development			
A. thaliana	*AtAP1*. *AtSVP* and *AtTFL1*	See above	See above	*ap1svptfl1* triple mutants had a continuous production of inflorescence meristem in place of flower production	Liu *et al.* (2019)
A. thaliana	*AtFT*	See above	Inversion mutation	-	Zhang *et al.* (2017a)
A. thaliana	*AtTFL1*	See above	Inversion mutation	-	Zhang *et al.* (2017a)
A. thaliana	*AtFDP*	bZIP TF that is a paralog of FD	Deletions varying from 1-bp to 58-bp in mutants	Flowering is slightly earlier than WT	Romera-Branchat *et al.* (2020)
O. sativa	*OsHBF1/2*	bZIP TF binding to *OsHd3a* and *OsRFT1* / repress flowering	Small InDels in both loci	*hbf1hbf2* double mutants are early-flowering	Brambilla *et al.* (2017)
O. sativa	*OsPHL3*	G2-like TF containing a Myb-CC domain	Multiple InDels in mutants with different genetic backgrounds	Mutants display an early flowering under LD and SD	Zeng *et al.* (2018)

Species	Target genes	Gene product/function	Type of editing events	Phenotypes	Reference
				regardless of their genetic background	
O. sativa	*OsEhd1*	B-type two-component response regulator / Regulates floral transition and regulates the expression of *OsHd3a* and *OsRFT1*	Short deletions generating in-frame and frameshift edits	Delayed vegetative growth and delayed flowering	Wu *et al.* (2020)
O. sativa	*OsFTL1/4/5/6/9/10/13/*	Homologs of *AtFT*/florigen, flowering induction	Frameshift mutations in their ORFs	Premature leaf senescence	Ma *et al.* (2015)
Solanum lycopersicum	*LsSP5G*	Homolog of *AtFT*/florigen, flowering induction, shape photoperiod adaptation	Deletions in targeted sequences	Loss of day-length-sensitive flowering Rapid flowering and enhancement of determinate growth habit	Soyk *et al.* (2017)

Table 5.2 Summary of the selected literature using CRISPR-based systems to induce mutagenesis in the flowering genes of *Arabidopsis*, rice, and tomato.

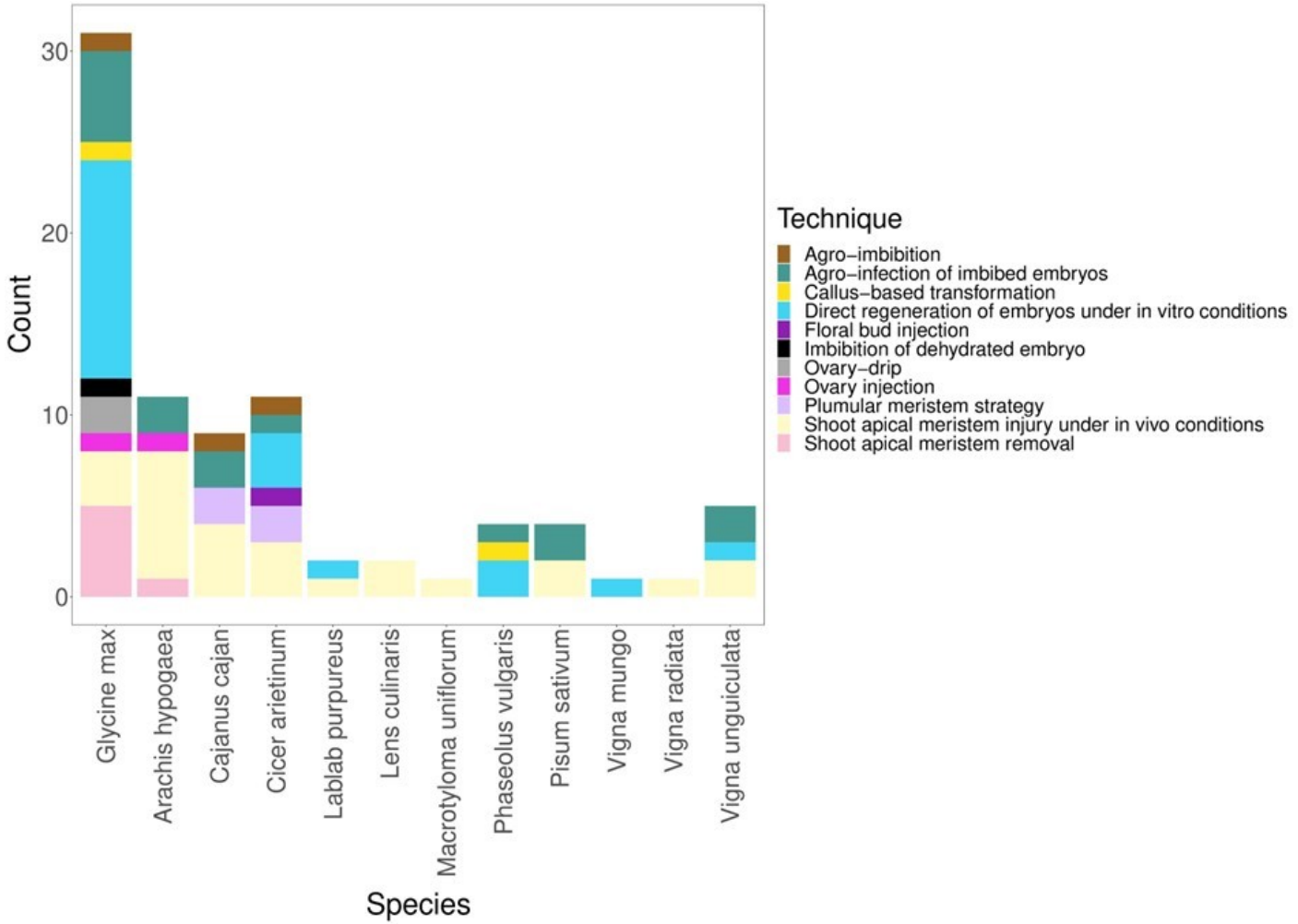

Figure 5.1 Distribution of *in planta* techniques across soybean and 11 annual leguminous species. Distribution of the publications associated with each type of explant and transformation technique. The count represents the number of unique publications using the technique.

7. Conclusion and Future Directions

7.1 Summary

The work presented in this thesis highlights the underlying genomic locations and gene regulatory networks regulating early reproductive traits in two early-maturing soybean populations adapted for MGs 00 and 000. The study presented in Chapter 3 demonstrated that only a few major regions regulate the pod-filling and maturity traits in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations. Importantly, this chapter dissected the *E8* region into three smaller regions (*E8-r1*, *E8-r2,* and *E8-r3*) encompassing major flowering genes. In addition to these findings, Chapter 3 identified the interactions between 13 genes, including the *E6* locus, and the *E8-r3* locus. Using a five-step variant analysis pipeline, several SNPs located in genes with predicted flowering and transcription factor functions were proposed as candidates for the identified loci.

The work in Chapter 4 proposed novel pipelines to identify *cis* and *trans* eQTL interactions and locate *trans* interactions-regulating hotspots. Using these pipelines, several *cis* and *trans* interactions for multiple genes with predicted FRSPD functions and four major FRSPD hotspots were identified in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations. In this study, deeper investigations using variant and co-expression network analyses unraveled two specific networks (F2_GM15:49,385,092-49,442,237 and F2_GM18:1,434,182-1,935,386) regulating three candidate genes (*GmLCH4a*, *GmPRR1a*, and *GmMDE04*) identified in Chapter 3, thus reinforcing their respective status as candidates.

The study presented in Chapter 5 identified 12 major loci regulating seed quality traits in $QS15524_{F2:F3}$ and $QS15544_{RIL}$. Three of these regions (Merg_GM06, Merg_GM13, and Merg_G19) were found to be regulating seed weight and identified in both mapping populations. Additionally, three other regions were identified as regulating the 100-seed weight or oleic acid traits and found to be overlapping or close to three other reproductive loci identified in Chapter 3.

Chapter 6 discussed the importance of the loci identified in Chapters 3, 4, and 5 for the improvement of diversity within the early-maturing soybean germplasm. Furthermore, this chapter investigated the toolkit available to breeders (e.g., genomic prediction and *in planta* transformation) to expand soybean's cultivation areas. This section also highlighted the work currently in preparation from our lab regarding *in planta* transformation to disentangle the bottlenecks associated with soybean transformation and promote the use of this conceptual framework.

7.2 Future Directions

As a whole, this thesis aimed at providing tools for an accelerated and more accurate breeding of early-maturing soybeans. Yet, the work presented here is partial and much still remains to be discovered. For instance, close to none of the candidates presented in Chapters 3-5 have been validated thoroughly using reverse genetics. Similarly, none of the suggested *in planta* transformation techniques referenced in Chapter 6 have been validated in our lab and their respective efficiencies thus remain hypothetical. In my humble opinion, I think that the work concerning *in planta* transformation should be of primary concern for Canadian research groups. Many of us, including our lab, cannot validate their candidates due to an incapacity of generating transformants. As a consequence, Canadian laboratories often outsource this task to other institutions which ultimately strengthen their transformation capacities while ours remain stagnant. Overall, major collaborative research efforts [e.g., SoyaGen project from Belzile *et al.* (2022)] have been displayed across Canada to develop genotyping platforms with the objective of implementing large-scale genomic projects [e.g., development of a genomic-assisted breeding toolkit from Genome Canada (2020)]. However, fewer engineering projects in soybean have been initiated over the years, meaning that this aspect is currently missing for Canadian breeders. The reasons supporting this choice are simple and are all mainly associated with the technical hurdles of soybean transformation.

In the longer term, this lack of interest from the Canadian public sector for soybean transformation will eventually impact our ability to be commercially competitive against entities investing larger amounts of financial and human resources in this field. Considering this, *in planta* strategies can be seen as an affordable, low-risk gateway to genome editing in soybean and many of the strategies described in Chapter 6 offer tremendous potential for a broad range of species. To overcome these challenges, a greater understanding of the mechanisms underlying the *in planta* concept is a must and will only be possible through collaborative efforts with other labs. As a starting option, I would suggest research groups interested in this approach to contact those who developed these technologies to build partnerships that will result in the implementation of these techniques in our laboratories. Moreover, I would suggest Canadian laboratories to boost their understanding of the concept by building a research platform grouping several labs interested in the genetic modification of soybean, but also other morphologically similar crops (e.g., common bean).

Along with *in planta* transformation, this research platform should also integrate a component regarding genome editing systems such as CRISPR-Cas9 or Mad7 endonucleases. As of now, our laboratory has succeeded the cloning phase of guide RNAs into the pGES201 (Bai *et al.*, 2020) and pHEE401E (Zheng *et al.*, 2020) CRISPR-Cas9 vectors to generate knockout lines for several candidate genes associated with maturity in soybean and their orthologs in *Arabidopsis*. The use of genome editing will lead to the generation of innovative phenotypes and genotypic combinations unseen before. Therefore, this option needs to be seriously considered to achieve the sustainable intensification of our cropping systems and considerable efforts thus need to be deployed to find high-throughput transformation techniques that are accessible to all. On a final note, latitudinal frontiers are deeply intertwined with scientific frontiers. The ongoing scientific development will reshape this world and push the frontiers farther than ever, with the possibility of extending soybean's cultivation range in locations that were previously out of range.

8. Appendix

8.1 Appendix 1 – Supplementary Information for Chapter 3

Appendix 1.1 Phenotypic data associated with each of the lines of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.2 Descriptive statistics associated with the four phenotypes for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.3 Pearson correlations associated with each of the phenotypes of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.4 Statistical analyses associated with each of the field phenotypes of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.5 Genotypes associated with each marker of the QS15524$_{F2:F3}$ population.

Appendix 1.6 Genotypes associated with each marker of the QS15544$_{RIL}$ population.

Appendix 1.7 Linkage maps for both populations.

Appendix 1.8 QTL analyses for each of the studied year for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 1.9 Minor QTL regions identified in each of the populations.

Appendix 1.10 Expression QTL regions for the *E8-r3* locus.

8.2 Appendix 2 – Supplementary Information for Chapter 4

Appendix 2.1 Gene expression dataset for the QS15524$_{F2:F3}$ parental lines.

Appendix 2.2 Gene expression dataset for the QS15544$_{RIL}$ parental lines.

Appendix 2.3 FRSPD gene ontology annotations from Soybase.

Appendix 2.4 Single nucleotide polymorphisms found in the E8-r3 region.

Appendix 2.5 Transcription factors from PlantTFDB and Soybase.

Appendix 2.6 Linkage maps for the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ populations.

Appendix 2.7 Genotypes for the 1,613 markers of the QS15524$_{F2:F3}$ population.

Appendix 2.8 Genotypes for the 2,746 markers of the QS15544$_{RIL}$ population.

Appendix 2.9 Differentially expressed genes for the QS15524$_{F2:F3}$ parental lines.

Appendix 2.10 Differentially expressed genes for the QS15544$_{RIL}$ parental lines.

Appendix 2.11 Gene ontology annotations for the differentially expressed genes of the QS15524$_{F2:F3}$ and QS15544$_{RIL}$ parental lines.

Appendix 2.12 Mapping of the eQTL interactions in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations using three algorithms (ICIM, IM, and GCIM).

Appendix 2.13 General statistical relative to the cis and trans interactions mapping in both populations.

Appendix 2.14 Number of interactions per gene in both populations before merging.

Appendix 2.15 Number of interactions per gene in both populations after merging.

Appendix 2.16 General statistics for the number of eQTL interactions before and after merging.

Appendix 2.17 Expression QTL interactions after the merging in the $QS15524_{F2:F3}$ and $QS15544_{RIL}$ populations.

Appendix 2.18 Statistics relevant to the identification of eQTL hotspots.

Appendix 2.19 Gene ontology annotations associated with the eQTL hotspots.

Appendix 2.20 Transcription-wide co-expression network statistics.

Appendix 2.21 Gene ontology annotations associated with the transcriptome-wide co-expression networks.

Appendix 2.22 Minor regions and candidate transcription factors associated with the regulation of the E8-r3 genes.

8.3 Appendix 3 – Supplementary Information for Chapter 5

Appendix 3.1 Threshold values for the logarithm of the odds obtained for each trait in both populations.

Appendix 3.2 Major and minor QTL identified in each of the populations.

Appendix 3.3 Lists of gene ontology annotations and candidate genes associated with PFASW functions.

Appendix 3.4 Linkage map for the $QS15544_{RIL}$ population.

Appendix 3.5 Linkage map for the $QS15524_{F2:F3}$ population.

Appendix 3.6 Descriptive statistics for each of the studied traits.

Appendix 3.7 Phenotypic data for each of the studied traits.

8.4 Appendix 4 – Supplementary Information for Chapter 6

Appendix 4.1 References for the *in planta* transformation approaches for legumes.

9. References

Arias, D., Mckersie, B., and Taylor, J. (2003). In planta transformation by embryo imbibition of Agrobacterium (Patent US 31478001). Available at: https://lens.org/144-522-898-736-153

Asharani, B. M (2011). Transformation of chickpea lines with Cry1X using in planta transformation and characterization of putative transformants T1 lines for molecular and biochemical characters. *J. Plant Breed. Crop Sci.* 3, 413–423. doi: 10.5897/jpbcs11.074

Bagg, J., Banks, S., Baute, T., Bohner, H., Brown, C., Griffiths, H., *et al.* (2002). *Agronomy guide for field crops*., 3rd editio, ed. C. Brown. Toronto: Ministry of Agriculture, Food and Rural Affairs. Available at: http://www.omafra.gov.on.ca/english/crops/pub811/pub811.pdf

Bahadur, K. K. C., Green, A. G., Wassmansdorf, D., Gandhi, V., Nadeem, K., and Fraser, E. D. G. (2021). Opportunities and trade-offs for expanding agriculture in Canada's North: an ecosystem service perspective. *Facets* 6, 1728–1752. doi: 10.1139/facets-2020-0097

Bai, M., Yuan, J., Kuang, H., Gong, P., Li, S., Zhang, Z., *et al.* (2020). Generation of a multiplex mutagenesis population via pooled CRISPR-Cas9 in soya bean. *Plant Biotechnol. J.* 18, 721–731. doi: 10.1111/pbi.13239

Bakels, C. (2014). The first farmers of the Northwest European plain: Some remarks on their crops, crop cultivation and impact on the environment. *J. Archaeol. Sci.* 51, 94–97. doi: 10.1016/j.jas.2012.08.046

Bao, A., Zhang, C., Huang, Y., Chen, H., Zhou, X., and Cao, D. (2020). Genome editing technology and application in soybean improvement. *Oil Crop Sci.* 5, 31–40. doi: 10.1016/j.ocsci.2020.03.001

Beche, E., Gillman, J. D., Song, Q., Nelson, R., Beissinger, T., Decker, J., *et al.* (2021). Genomic prediction using training population design in interspecific soybean populations. *Mol. Breed.* 41, 15. doi: 10.1007/s11032-021-01203-6

Belzile, F., Jean, M., Torkamaneh, D., Tardivel, A., Lemay, M. A., Boudhrioua, C., *et al.* (2022). The SoyaGen Project: Putting Genomics to Work for Soybean Breeders. *Front. Plant Sci.* 13, 1–10. doi: 10.3389/fpls.2022.887553

Benton, T., Bieg, C., Harwatt, H., Pudassaini, R., and Wellesley, L. (2021). Food system impacts on biodiversity loss: Three levers for food system transformation in support of nature. *Food Agric. Organ. United Nations*, 1–71.

Bevan, M. W., Flavell, R. B., and Chilton, M.-D. (1983). A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. *Nature* 304, 184–187. doi: 10.1038/304184a0

Bian, S., Jin, D., Li, R., Xie, X., Gao, G., Sun, W., *et al.* (2017). Genome-wide analysis of CCA1-like proteins in soybean and functional characterization of GmMYB138a. *Int. J. Mol. Sci.* 18, 2040. doi: 10.3390/ijms18102040

Bolon, Y., Hyten, D. L., Orf, J. H., Vance, C. P., and Muehlbauer, G. J. (2014). eQTL Networks Reveal Complex Genetic Architecture in the Immature Soybean Seed. *Plant Genome* 7, 1–14. doi: 10.3835/plantgenome2013.08.0027

Bonato, E. R., and Vello, N. A. (1999). E6, a dominant gene conditioning early flowering and maturity in soybeans. *Genet. Mol. Biol.* 22, 229–232. doi: 10.1590/S1415-47571999000200016

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2016). FLOR-ID: An interactive database of flowering-time gene networks in *Arabidopsis* thaliana. *Nucleic Acids Res.* 44, D1167–D1171. doi: 10.1093/nar/gkv1054

Brambilla, V., Martignago, D., Goretti, D., Cerise, M., Somssich, M., De Rosa, M., *et al.* (2017). Antagonistic transcription factor complexes modulate the floral transition in rice. *Plant Cell* 29, 2801–2816. doi: 10.1105/tpc.17.00645

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science (80-.).* 296, 752–755. doi: 10.1126/science.1069516

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., and Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. *Theor. Appl. Genet.* 132, 3089–3100. doi: 10.1007/s00122-019-03408-y

Buzzell, R. I. (1971). Inheritance of a Soybean Flowering Response To Fluorescent-Daylength Conditions. *Can. J. Genet. Cytol.* 13, 703–707. doi: 10.1139/g71-100

Buzzell, R. I., and Voldeng, H. D. (1980). Inheritance of Insensitivity to Long Daylength. *Soybean Genet. Newsl.* 7, 26–29.

Bylino, O. V., Ibragimov, A. N., and Shidlovskii, Y. V. (2020). Evolution of regulated transcription. *Cells* 9, 1–38. doi: 10.3390/cells9071675

Cai, Y., Chen, L., Liu, X., Guo, C., Sun, S., Wu, C., *et al.* (2018). CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. *Plant Biotechnol. J.* 16, 176–185. doi: 10.1111/pbi.12758

Cai, Y., Chen, L., Zhang, Y., Yuan, S., Su, Q., Sun, S., *et al.* (2020a). Target base editing in soybean using a modified CRISPR/Cas9 system. *Plant Biotechnol. J.* 18, 1996–1998. doi: 10.1111/pbi.13386

Cai, Y., Wang, L., Chen, L., Wu, T., Liu, L., Sun, S., *et al.* (2020b). Mutagenesis of GmFT2a and GmFT5a mediated by CRISPR/Cas9 contributes for expanding the regional adaptability of soybean. *Plant Biotechnol. J.* 18, 298–309. doi: 10.1111/pbi.13199

Canadian Food Inspection Agency (2021). The Biology of Glycine max (L.) Merr. (Soybean). *Biol. Doc. BIO2021-01 A companion Doc. to Dir. 94-08 (Dir94-08), Assess. Criteria Determ. Environ. Saf. Plant with Nov. Trait.* Available at: https://inspection.canada.ca/plant-varieties/plants-with-novel-traits/applicants/directive-94-08/biology-documents/glycine-max-l-merr-/eng/1330975306785/1330975382668 (Accessed September 23, 2023).

Challinor, A. J., and Wheeler, T. R. (2008). Crop yield reduction in the tropics under climate change: Processes and uncertainties. *Agric. For. Meteorol.* 148, 343–356. doi: 10.1016/j.agrformet.2007.09.015

Chee, P. P., Fober, K. A., and Slightom, J. L. (1989). Transformation of Soybean (Glycine max) by Infecting Germinating Seeds with Agrobacterium tumefaciens. *Plant Physiol.* 91, 1212–

1218. doi: 10.1104/pp.91.3.1212

Chen, L., Nan, H., Kong, L., Yue, L., Yang, H., Zhao, Q., *et al.* (2020). Soybean AP1 homologs control flowering time and plant height. *J. Integr. Plant Biol.* 62, 1868–1879. doi: 10.1111/jipb.12988

Cheng, J. Z., Zhou, Y. P., Lv, T. X., Xie, C. P., and Tian, C. E. (2017). Research progress on the autonomous flowering time pathway in *Arabidopsis*. *Physiol. Mol. Biol. Plants* 23, 477–485. doi: 10.1007/s12298-017-0458-3

Cheng, L., Wang, Y., Zhang, C., Wu, C., Xu, J., Zhu, H., *et al.* (2011). Genetic analysis and QTL detection of reproductive period and post-flowering photoperiod responses in soybean. *Theor. Appl. Genet.* 123, 421–429. doi: 10.1007/s00122-011-1594-8

Cheng, Q., Dong, L., Su, T., Li, T., Gan, Z., Nan, H., *et al.* (2019). CRISPR/Cas9-mediated targeted mutagenesis of GmLHY genes alters plant height and internode length in soybean. *BMC Plant Biol.* 19, 1–11. doi: 10.1186/s12870-019-2145-8

Chien, P. S., Chen, P. H., Lee, C. R., and Chiou, T. J. (2023). Transcriptome-wide association study coupled with eQTL analysis reveals the genetic connection between gene expression and flowering time in *Arabidopsis*. *J. Exp. Bot.* 74, 5653–5666. doi: 10.1093/jxb/erad262

Cho, H. J., Moy, Y., Rudnick, N. A., Klein, T. M., Yin, J., Bolar, J., *et al.* (2022). Development of an efficient marker-free soybean transformation method using the novel bacterium Ochrobactrum haywardense H1. *Plant Biotechnol. J.* 20, 977–990. doi: 10.1111/pbi.13777

Christou, P., Murphy, J. E., and Swain, W. F. (1987). Stable transformation of soybean by electroporation and root formation from transformed callus. *Proc. Natl. Acad. Sci.* 84, 3962–3966. doi: 10.1073/pnas.84.12.3962

Christou, P., Swain, W. F., Yang, N.-S., and McCabe, D. E. (1989). Inheritance and expression of foreign genes in transgenic soybean plants. *Proc. Natl. Acad. Sci.* 86, 7500–7504. doi: 10.1073/pnas.86.19.7500

Cinner, J. E., Caldwell, I. R., Thiault, L., Ben, J., Blanchard, J. L., Coll, M., *et al.* (2022). Potential impacts of climate change on agriculture and fisheries production in 72 tropical coastal communities. *Nat. Commun.* 13, 3530. doi: 10.1038/s41467-022-30991-4

Cober, E. R. (2011). Long juvenile soybean flowering responses under very short photoperiods. *Crop Sci.* 51, 140–145. doi: 10.2135/cropsci2010.05.0262

Cober, E. R., Molnar, S. J., Charette, M., and Voldeng, H. D. (2010). A new locus for early maturity in soybean. *Crop Sci.* 50, 524–527. doi: 10.2135/cropsci2009.04.0174

Cober, E. R., Tanner, J. W., and Voldeng, H. D. (1996). Soybean photoperiod-sensitivity loci respond differentially to light quality. *Crop Sci.* 36, 606–610. doi: 10.2135/cropsci1996.0011183X003600030014x

Cober, E. R., and Voldeng, H. D. (2001). Low R:FR light quality delays flowering of E7E7 soybean lines. *Crop Sci.* 41, 1823–1826. doi: 10.2135/cropsci2001.1823

Copley, T. R., Duceppe, M. O., and O'Donoughue, L. S. (2018). Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines.

BMC Genomics 19, 1–12. doi: 10.1186/s12864-018-4558-4

De Ronde, J. A., Cress, W. A., and Van der Mescht, A. (2001). Agrobacterium-mediated transformation of soybean (Glycine max) seed with the β-glucuronidase marker gene. *S. Afr. J. Sci.* 97, 421–424.

Dhir, S. K., Dhir, S., Savka, M. A., Belanger, F., Kriz, A. L., Farrand, S. K., *et al.* (1992). Regeneration of Transgenic Soybean (Glycine max) Plants from Electroporated Protoplasts. *Plant Physiol.* 99, 81–88. doi: 10.1104/pp.99.1.81

Dietz, N., Combs-Giroir, R., Cooper, G., Stacey, M., Miranda, C., and Bilyeu, K. (2021). Geographic distribution of the E1 family of genes and their effects on reproductive timing in soybean. *BMC Plant Biol.* 21, 1–13. doi: 10.1186/s12870-021-03197-x

Dijk, M., Morley, T., Rau, M. L., and Saghai, Y. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* 2. doi: 10.1038/s43016-021-00322-9

Dissanayaka, A., Rodriguez, T. O., Di, S., Yan, F., Githiri, S. M., Rodas, F. R., *et al.* (2016). Quantitative trait locus mapping of soybean maturity gene E5. *Breed. Sci.* 66, 407–415. doi: 10.1270/jsbbs.15160

Dong, L., Cheng, Q., Fang, C., Kong, L., Yang, H., Hou, Z., *et al.* (2022). Parallel selection of distinct Tof5 alleles drove the adaptation of cultivated and wild soybean to high latitudes. *Mol. Plant* 15, 308–321. doi: 10.1016/j.molp.2021.10.004

Dong, L., Fang, C., Cheng, Q., Su, T., Kou, K., Kong, L., *et al.* (2021). Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nat. Commun.* 12, 1–11. doi: 10.1038/s41467-021-25800-3

Dong, L., Li, S., Wang, L., Su, T., Zhang, C., Bi, Y., *et al.* (2023). The genetic basis of high-latitude adaptation in wild soybean. *Curr. Biol.* 33, 252-262.e4. doi: 10.1016/j.cub.2022.11.061

Dong, Y. S., Zhao, L. M., Liu, B., Wang, Z. W., Jin, Z. Q., and Sun, H. (2004). The genetic diversity of cultivated soybean grown in China. *Theor. Appl. Genet.* 108, 931–936. doi: 10.1007/s00122-003-1503-x

Druka, A., Potokina, E., Luo, Z., Jiang, N., Chen, X., Kearsey, M., *et al.* (2010). Expression quantitative trait loci analysis in plants. *Plant Biotechnol. J.* 8, 10–27. doi: 10.1111/j.1467-7652.2009.00460.x

Eiji, E., Alemayehu, T., Siljander, M., and Aragão, L. E. O. C. (2021). Large-scale commodity agriculture exacerbates the climatic impacts of Amazonian deforestation. 118. doi: 10.1073/pnas.2023787118

Engler, C., Youles, M., Gruetzner, R., Ehnert, T. M., Werner, S., Jones, J. D. G., *et al.* (2014). A Golden Gate modular cloning toolbox for plants. *ACS Synth. Biol.* 3, 839–843. doi: 10.1021/sb4001504

Entoori, K., Sreevathsa, R., Arthikala, M. K., Kumar, A., Raja, A., Kumar, V., *et al.* (2008). A chimeric cry1X gene imparts resistance to Spodoptera litura and Helicoverpa armigera in the transgenic groundnut. *EurAsia J. Biosci.* 65, 53–65.

Ericksen, P., Thornton, P., Notenbaert, A., Cramer, L., Jones, P., Herrero, M., *et al.* (2011). Mapping hotspots of climate change and food insecurity in the global tropics. *Change*, 29. Available at: http://ccafs.cgiar.org/sites/default/files/assets/docs/ccafsreport5-climate_hotspots_final.pdf

Fan, C., Hu, R., Zhang, X., Wang, X., Zhang, W., Zhang, Q., *et al.* (2014). Conserved CO-FT regulons contribute to the photoperiod flowering control in soybean. *BMC Plant Biol.* 14, 9. doi: 10.1186/1471-2229-14-9

Fang, C., Liu, J., Zhang, T., Su, T., Li, S., Cheng, Q., *et al.* (2021a). A recent retrotransposon insertion of J caused E6 locus facilitating soybean adaptation into low latitude. *J. Integr. Plant Biol.* 63, 995–1003. doi: 10.1111/jipb.13034

Fang, Y., Wang, L., Sapey, E., Fu, S., Wu, T., Zeng, H., *et al.* (2021b). Speed-Breeding System in Soybean: Integrating Off-Site Generation Advancement, Fresh Seeding, and Marker-Assisted Selection. *Front. Plant Sci.* 12, 1–9. doi: 10.3389/fpls.2021.717077

Fornara, F., de Montaigu, A., and Coupland, G. (2010). SnapShot: Control of flowering in *Arabidopsis*. *Cell* 141, 3–5. doi: 10.1016/j.cell.2010.04.024

Fu, Y. B., Peterson, G. W., and Morrison, M. J. (2007). Genetic diversity of Canadian soybean cultivars and exotic germplasm revealed by simple sequence repeat markers. *Crop Sci.* 47, 1947–1954. doi: 10.2135/cropsci2006.12.0843

Ganguly, S., Ghosh, G., Ghosh, S., Purohit, A., Chaudhuri, R. K., Das, S., *et al.* (2020a). Plumular meristem transformation system for chickpea: an efficient method to overcome recalcitrant tissue culture responses. *Plant Cell. Tissue Organ Cult.* 142, 493–504. doi: 10.1007/s11240-020-01873-8

Ganguly, S., Ghosh, G., Purohit, A., Kundu Chaudhuri, R., and Chakraborti, D. (2018). Development of transgenic pigeonpea using high throughput plumular meristem transformation method. *Plant Cell. Tissue Organ Cult.* 135, 73–83. doi: 10.1007/s11240-018-1444-3

Ganguly, S., Purohit, A., Chaudhuri, R. K., Das, S., and Chakraborti, D. (2020b). "Embryonic explant and plumular meristem transformation methods for development of transgenic pigeon pea," in *Methods in Molecular Biology*, (Springer), 317–333. doi: 10.1007/978-1-0716-0235-5_17

Gao, X. R., Wang, G. K., Su, Q., Wang, Y., and An, L. J. (2007). Phytase expression in transgenic soybeans: Stable transformation with a vector-less construct. *Biotechnol. Lett.* 29, 1781–1787. doi: 10.1007/s10529-007-9439-x

Gélinas Bélanger, J., Copley, T. R., Hoyos-Villegas, V., Charron, J.-B., and O'Donoughue, L. (2024). A comprehensive review of in planta stable transformation strategies. *Plant Methods* 20, 79. doi: 10.1186/s13007-024-01200-8

Genome Canada (2020). Development and implementation of a Toolkit for Genomics-Assisted Breeding in Soybean. Available at: https://genomecanada.ca/project/development-and-implementation-toolkit-genomics-assisted-breeding-soybean/ (Accessed September 25, 2023).

Gonzalez, D. H. (2016). "Introduction to Transcription Factor Structure and Function," in *Plant Transcription Factors: Evolutionary, Structural and Functional Aspects*, ed. D. H. B. T.-P. T. F. Gonzalez (Boston: Academic Press), 3–11. doi: 10.1016/B978-0-12-800854-6.00001-4

Government of Canada (2020). Climate change impacts on agriculture - agriculture. Available at: https://agriculture.canada.ca/en/environment/climate-scenarios-agriculture (Accessed July 5, 2024).

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, 843–846. doi: 10.1093/nar/gkp798

Greenham, K., and McClung, C. R. (2015). Integrating circadian dynamics with physiological processes in plants. *Nat. Rev. Genet.* 16, 598–610. doi: 10.1038/nrg3976

Guo, G., Xu, K., Zhang, X., Zhu, J., Lu, M., Chen, F., *et al.* (2015). Extensive Analysis of GmFTL and GmCOL Expression in Northern Soybean Cultivars in Field Conditions. *PLoS One* 10, e0136601. doi: 10.1371/journal.pone.0136601

Guo, J., Wang, Y., Song, C., Zhou, J., Qiu, L., Huang, H., *et al.* (2010). A single origin and moderate bottleneck during domestication of soybean (Glycine max): Implications from microsatellites and nucleotide sequences. *Ann. Bot.* 106, 505–514. doi: 10.1093/aob/mcq125

Gupta, S., Kumawat, G., Agrawal, N., Tripathi, R., Rajesh, V., Nataraj, V., *et al.* (2022). Photoperiod trait: Insight in molecular mechanism for growth and maturity adaptation of soybean (Glycine max) to different latitudes. *Plant Breed.* 141, 483–500. doi: 10.1111/pbr.13041

Haas, M., Schreiber, M., and Mascher, M. (2019). Domestication and crop evolution of wheat and barley: Genes, genomics, and future directions. *J. Integr. Plant Biol.* 61, 204–225. doi: 10.1111/jipb.12737

Hahn, F., Korolev, A., Loures, L. S., and Nekrasov, V. (2019). A modular cloning toolkit for genome editing in plants. *bioRxiv*, 1–10. doi: 10.1101/738021

Han, J., Guo, B., Guo, Y., Zhang, B., Wang, X., and Qiu, L. J. (2019). Creation of Early Flowering Germplasm of Soybean by CRISPR/Cas9 Technology. *Front. Plant Sci.* 10, 1–10. doi: 10.3389/fpls.2019.01446

Hannah, L., Roehrdanz, P. R., Krishna Bahadur, K. C., Fraser, E. D. G., Donatti, C. I., Saenz, L., *et al.* (2020). The environmental consequences of climate-driven agricultural frontiers. *PLoS One* 15, e0228305. doi: 10.1371/journal.pone.0228305

Holme, I. B., Wendt, T., and Holm, P. B. (2013). Intragenesis and cisgenesis as alternatives to transgenic crop development. *Plant Biotechnol. J.* 11, 395–407. doi: 10.1111/pbi.12055

Hou, Z., Liu, B., and Kong, F. (2022). "Chapter Two - Regulation of flowering and maturation in soybean," in *Soybean Physiology and Genetics*, eds. H.-M. Lam and M.-W. B. T.-A. in B. R. Li (Academic Press), 43–75. doi: https://doi.org/10.1016/bs.abr.2022.02.007

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y. Y., Nelson, R. L., Costa, J. M., *et al.* (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci.* 103, 16666–16671.

doi: 10.1073/pnas.0604379103

Hyun, Y., Kim, J., Cho, S. W., Choi, Y., Kim, J. S., and Coupland, G. (2014). Site-directed mutagenesis in *Arabidopsis* thaliana using dividing tissue-targeted RGEN of the CRISPR/Cas system to generate heritable null alleles. *Planta* 241, 271–284. doi: 10.1007/s00425-014-2180-5

Iquira, E., Gagnon, É., and Belzile, F. (2010). Comparison of genetic diversity between canadian adapted genotypes and exotic germplasm of soybean. *Genome* 53, 337–345. doi: 10.1139/G10-009

Jacobs, T. B., LaFayette, P. R., Schmitz, R. J., and Parrott, W. A. (2015). Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnol.* 15, 1–10. doi: 10.1186/s12896-015-0131-2

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., *et al.* (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nat. Food* 2, 873–885. doi: 10.1038/s43016-021-00400-y

Jähne, F., Hahn, V., Würschum, T., and Leiser, W. L. (2020). Speed breeding short-day crops by LED-controlled light schemes. *Theor. Appl. Genet.* 133, 2335–2342. doi: 10.1007/s00122-020-03601-4

Jean, M., Cober, E., O'Donoughue, L., Rajcan, I., and Belzile, F. (2021). Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. *Crop Sci.* 61, 3908–3918. doi: 10.1002/csc2.20583

Jia, H., Jiang, B., Wu, C., Lu, W., Hou, W., Sun, S., *et al.* (2014). Maturity group classification and maturity locus genotyping of early-maturing soybean varieties from high-latitude cold regions. *PLoS One* 9. doi: 10.1371/journal.pone.0094139

Jiang, B., Nan, H., Gao, Y., Tang, L., Yue, Y., Lu, S., *et al.* (2014). Allelic combinations of soybean maturity loci E1, E2, E3 and E4 result in diversity of maturity and adaptation to different latitudes. *PLoS One* 9. doi: 10.1371/journal.pone.0106042

Jiang, B., Zhang, S., Song, W., Khan, M. A. A., Sun, S., Zhang, C., *et al.* (2019). Natural variations of FT family genes in soybean varieties covering a wide range of maturity groups. *BMC Genomics* 20, 1–16. doi: 10.1186/s12864-019-5577-5

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., *et al.* (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Jones, M. K., and Lister, D. L. (2022). The Domestication of the Seasons: The Exploitation of Variations in Crop Seasonality Responses by Later Prehistoric Farmers. *Front. Ecol. Evol.* 10, 607. doi: 10.3389/fevo.2022.907536

Jung, C. H., Wong, C. E., Singh, M. B., and Bhalla, P. L. (2012). Comparative genomic analysis of soybean flowering genes. *PLoS One* 7. doi: 10.1371/journal.pone.0038250

Kanazawa, A., Liu, B., Kong, F., Arase, S., and Abe, J. (2009). Adaptive evolution involving gene duplication and insertion of a novel Ty1/copia-like retrotransposon in soybean. *J. Mol. Evol.*

69, 164–175. doi: 10.1007/s00239-009-9262-1

Karthik, S., Pavan, G., Sathish, S., Siva, R., Kumar, P. S., and Manickavasagam, M. (2018). Genotype-independent and enhanced in planta Agrobacterium tumefaciens-mediated genetic transformation of peanut [Arachis hypogaea (L.)]. *3 Biotech* 8. doi: 10.1007/s13205-018-1231-1

Kaufmann, K., and Airoldi, C. A. (2018). Master regulatory transcription factors in plant development: A blooming perspective. *Methods Mol. Biol.* 1830, 3–22. doi: 10.1007/978-1-4939-8657-6_1

Kaur, A., Sharma, M., Sharma, C., Kaur, H., Kaur, N., Sharma, S., *et al.* (2016). Pod borer resistant transgenic pigeon pea (Cajanus cajan L.) expressing cry1Ac transgene generated through simplified Agrobacterium transformation of pricked embryo axes. *Plant Cell. Tissue Organ Cult.* 127, 717–727. doi: 10.1007/s11240-016-1055-9

Keller, B., Ariza-Suarez, D., de la Hoz, J., Aparicio, J. S., Portilla-Benavides, A. E., Buendia, H. F., *et al.* (2020). Genomic Prediction of Agronomic Traits in Common Bean (Phaseolus vulgaris L.) Under Environmental Stress. *Front. Plant Sci.* 11, 1–15. doi: 10.3389/fpls.2020.01001

Kim, D. H. (2020). Current understanding of flowering pathways in plants: focusing on the vernalization pathway in *Arabidopsis* and several vegetable crop plants. *Hortic. Environ. Biotechnol.* 61, 209–227. doi: 10.1007/s13580-019-00218-5

Kim, H., and Choi, J. (2021). A robust and practical CRISPR/crRNA screening system for soybean cultivar editing using LbCpf1 ribonucleoproteins. *Plant Cell Rep.* 40, 1059–1070. doi: 10.1007/s00299-020-02597-x

Kim, H., Kim, S. T., Ryu, J., Kang, B. C., Kim, J. S., and Kim, S. G. (2017). CRISPR/Cpf1-mediated DNA-free plant genome editing. *Nat. Commun.* 8, 1–7. doi: 10.1038/ncomms14406

Kim, M. Y., Shin, J. H., Kang, Y. J., Shim, S. R., and Lee, S. H. (2012). Divergence of flowering genes in soybean. *J. Biosci.* 37, 857–870. doi: 10.1007/s12038-012-9252-0

Kong, F., Liu, B., Xia, Z., Sato, S., Kim, B. M., Watanabe, S., *et al.* (2010). Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in Soybean. *Plant Physiol.* 154, 1220–1231. doi: 10.1104/pp.110.160796

Kong, F., Nan, H., Cao, D., Li, Y., Wu, F., Wang, J., *et al.* (2014). A new dominant gene E9 conditions early flowering and maturity in soybean. *Crop Sci.* 54, 2529–2535. doi: 10.2135/cropsci2014.03.0228

Kong, L., Lu, S., Wang, Y., Fang, C., Wang, F., Nan, H., *et al.* (2018). Quantitative trait locus mapping of flowering time and maturity in soybean using next-generation sequencing-based analysis. *Front. Plant Sci.* 9, 1–20. doi: 10.3389/fpls.2018.00995

Kou, K., Yang, H., Li, H., Fang, C., Chen, L., Yue, L., *et al.* (2022). A functionally divergent SOC1 homolog improves soybean yield and latitudinal adaptation. *Curr. Biol.* 32, 1728–1742. doi: 10.1016/j.cub.2022.02.046

Lee, H. (2013). “An Overview of Genetic Transformation of Soybean,” ed. S.-Y. Park (Rijeka:

IntechOpen), Ch. 23. doi: 10.5772/51076

Li, C., Li, Y. Y. hui, Li, Y. Y. hui, Lu, H., Hong, H., Tian, Y., *et al.* (2020). A Domestication-Associated Gene GmPRR3b Regulates the Circadian Clock and Flowering Time in Soybean. *Mol. Plant* 13, 745–759. doi: 10.1016/j.molp.2020.01.014

Li, D., Liu, Q., and Schnable, P. S. (2021a). TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol.* 186, 1800–1811. doi: 10.1093/plphys/kiab161

Li, D., Wang, Q., Tian, Y., Lyu, X., Zhang, H., Sun, Y., *et al.* (2023a). Transcriptome brings variations of gene expression, alternative splicing, and structural variations into gene-scale trait dissection in soybean. *bioRxiv*, 2007–2023. doi: 10.1101/2023.07.03.545230

Li, F., Zhang, X., Hu, R., Wu, F., Ma, J., Meng, Y., *et al.* (2013). Identification and molecular characterization of FKF1 and GI homologous genes in soybean. *PLoS One* 8, 26–28. doi: 10.1371/journal.pone.0079036

Li, H., Du, H., Huang, Z., He, M., Kong, L., Fang, C., *et al.* (2023b). The AP2/ERF transcription factor TOE4b regulates photoperiodic flowering and grain yield per plant in soybean. *Plant Biotechnol. J.* 21, 1682–1694. doi: 10.1111/pbi.14069

Li, R., Jeong, K., Davis, J. T., Kim, S. S., Lee, S., Michelmore, R. W., *et al.* (2018). Integrated QTL and eQTL mapping provides insights and candidate genes for fatty acid composition, flowering time, and growth traits in a F2 population of a novel synthetic allopolyploid brassica napus. *Front. Plant Sci.* 871, 1–20. doi: 10.3389/fpls.2018.01632

Li, Y., Li, W., Zhang, C., Yang, L., Chang, R., Gaut, B. S., *et al.* (2010). Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytol.* 188, 242–253.

Li, Z., Cheng, Q., Gan, Z., Hou, Z., Zhang, Y., Li, Y., *et al.* (2021b). Multiplex CRISPR/Cas9-mediated knockout of soybean LNK2 advances flowering time. *Crop J.* 9, 767–776. doi: https://doi.org/10.1016/j.cj.2020.09.005

Lin, X., Dong, L., Tang, Y., Li, H., Cheng, Q., Li, H., *et al.* (2022). Novel and multifaceted regulations of photoperiodic flowering by phytochrome A in soybean. *Proc. Natl. Acad. Sci. U. S. A.* 119, 1–10. doi: 10.1073/pnas.2208708119

Lin, X., Fang, C., Liu, B., and Kong, F. (2021a). Natural variation and artificial selection of photoperiodic flowering genes and their applications in crop adaptation. *aBIOTECH* 2, 156–169. doi: 10.1007/s42994-021-00039-0

Lin, X., Liu, B., Weller, J. L., Abe, J., and Kong, F. (2021b). Molecular mechanisms for the photoperiodic regulation of flowering in soybean. *J. Integr. Plant Biol.* 63, 981–994. doi: 10.1111/jipb.13021

Linde, A. M., Eklund, D. M., Kubota, A., Pederson, E. R. A., Holm, K., Gyllenstrand, N., *et al.* (2017). Early evolution of the land plant circadian clock. *New Phytol.* 216, 576–590. doi: 10.1111/nph.14487

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., and Abe, J. (2008). Genetic

redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180, 995–1007. doi: 10.1534/genetics.108.092742

Liu, C., Chen, X., Wang, W., Hu, X., Han, W., He, Q., *et al.* (2021). Identifying wild versus cultivated gene-alleles conferring seed coat color and days to flowering in Soybean. *Int. J. Mol. Sci.* 22, 1–22. doi: 10.3390/ijms22041559

Liu, H. K., Yang, C., and Wei, Z. M. (2004). Efficient Agrobacterium tumefaciens-mediated transformation of soybeans using an embryonic tip regeneration system. *Planta* 219, 1042–1049. doi: 10.1007/s00425-004-1310-x

Liu, J., Su, Q., An, L., and Yang, A. (2009). Transfer of a minimal linear marker-free and vector-free smGFP cassette into soybean via ovary-drip transformation. *Biotechnol. Lett.* 31, 295–303. doi: 10.1007/s10529-008-9851-x

Liu, L. feng, Gao, L., Zhang, L. xin, Cai, Y. peng, Song, W. wen, Chen, L., *et al.* (2022a). Co-silencing E1 and its homologs in an extremely late-maturing soybean cultivar confers super-early maturity and adaptation to high-latitude short-season regions. *J. Integr. Agric.* 21, 326–335. doi: 10.1016/S2095-3119(20)63391-3

Liu, L., Song, W., Wang, L., Sun, X., Qi, Y., Wu, T., *et al.* (2020). Allele combinations of maturity genes E1-E4 affect adaptation of soybean to diverse geographic regions and farming systems in China. *PLoS One* 15, 1–15. doi: 10.1371/journal.pone.0235397

Liu, S., Won, H., Clarke, D., Matoba, N., Khullar, S., Mu, Y., *et al.* (2022b). Illuminating links between cis-regulators and trans-acting variants in the human prefrontal cortex. *Genome Med.* 14, 133. doi: 10.1186/s13073-022-01133-8

Liu, X., Fuller, D. Q., and Jones, M. (2015). "Early agriculture in China," in *The Cambridge World History*, eds. C. Goucher and G. Barker (Cambridge: Cambridge University Press), 310–334. doi: 10.1017/cbo9780511978807.013

Liu, Y., Gao, Y., Gao, Y., and Zhang, Q. (2019). Targeted deletion of floral development genes in *Arabidopsis* with CRISPR/Cas9 using the RNA endoribonuclease Csy4 processing system. *Hortic. Res.* 6. doi: 10.1038/s41438-019-0179-6

Liu, Z. M., Xiong, H. W., Xie, H., Qin, Y. T., Zhao, Y. R., and Guo, B. (2013). A technique for agrobacterium-mediated transformation via germinating seeds of soybean. *Adv. Mater. Res.* 749, 413–417. doi: 10.4028/www.scientific.net/AMR.749.413

Lu, S., Dong, L., Fang, C., Liu, S., Kong, L., Cheng, Q., *et al.* (2020). Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. *Nat. Genet.* 52, 428–436. doi: 10.1038/s41588-020-0604-7

Lu, S., Li, Y., Wang, J., Srinives, P., Nan, H., Cao, D., *et al.* (2015). QTL mapping for flowering time in different latitude in soybean: QTL mapping for flowering time. *Euphytica* 206, 725–736. doi: 10.1007/s10681-015-1501-5

Lu, S., Zhao, X., Hu, Y., Liu, S., Nan, H., Li, X., *et al.* (2017). Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. *Nat. Genet.* 49, 773–779. doi: 10.1038/ng.3819

Luth, D., Warnberg, K., and Wang, K. (2015). "Soybean [Glycine max (L.) Merr.] BT - Agrobacterium Protocols: Volume 1," ed. K. Wang (New York, NY: Springer New York), 275–284. doi: 10.1007/978-1-4939-1695-5_22

Lv, T., Wang, L., Zhang, C., Liu, S., Wang, J., Lu, S., *et al.* (2022). Identification of two quantitative genes controlling soybean flowering using bulked-segregant analysis and genetic mapping. *Front. Plant Sci.* 13, 987073. doi: 10.3389/fpls.2022.987073

Ma, X., Zhang, Q., Zhu, Q., Liu, W., Chen, Y., Qiu, R., *et al.* (2015). A Robust CRISPR/Cas9 System for Convenient, High-Efficiency Multiplex Genome Editing in Monocot and Dicot Plants. *Mol. Plant* 8, 1274–1284. doi: 10.1016/j.molp.2015.04.007

Masson-Delmotte, V., Pörtner, H. O., Skea, J., Slade, R., Ferrat, M., Neogi, S., *et al.* (2022). Climate Change and Land: An IPCC Special Report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. *Clim. Chang. L. An IPCC Spec. Rep. Clim. Chang. Desertif. L. Degrad. Sustain. L. Manag. food Secur. Greenh. gas fluxes Terr. Ecosyst.*, 1–896. doi: 10.1017/9781009157988

Matsumura, H., Liu, B., Abe, J., and Takahashi, R. (2008). AFLP mapping of soybean maturity gene E4. *J. Hered.* 99, 193–197. doi: 10.1093/jhered/esm114

Mcblain, B. A., and Bernard, R. L. (1987). A new gene affecting the time of flowering and maturity in soybeans. *J. Hered.* 78, 160–162. doi: 10.1093/oxfordjournals.jhered.a110349

Meyfroidt, P. (2021). Emerging agricultural expansion in northern regions: Insights from land-use research. *One Earth* 4, 1661–1664. doi: 10.1016/j.oneear.2021.11.019

Miller, M. J., Song, Q., Fallen, B., and Li, Z. (2023). Genomic prediction of optimal cross combinations to accelerate genetic improvement of soybean (Glycine max). *Front. Plant Sci.* 14, 1–12. doi: 10.3389/fpls.2023.1171135

Mily, R., Neelima, S., Anne, B., and Moran, F. (2020). Generation of Heritably Gene-Edited Plants Without Tissue Culture (Patent US 201862727431). 1. Available at: https://lens.org/011-524-499-432-378

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos P., G., and Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression. *World Acad Sci J* 2, 3–8. doi: 10.3892/wasj.2020.32

Molnar, S. J., Rai, S., Charette, M., and Cober, E. R. (2003). Simple sequence repeat (SSR) markers linked to E1, E3, E4, and E7 maturity genes in soybean. *Genome* 46, 1024–1036. doi: 10.1139/g03-079

Muluneh, M. G. (2021). Impact of climate change on biodiversity and food security : a global perspective — a review article. *Agric. Food Secur.*, 1–25. doi: 10.1186/s40066-021-00318-5

Murphy, D. J. (2007). People, Plants and Genes: The Story of Crops and Humanity. *People, Plants Genes Story Crop. Humanit.*, 1–436. doi: 10.1093/acprof:oso/9780199207145.001.0001

Näär, A. M., Lemon, B. D., and Tjian, R. (2001). Transcriptional coactivator complexes. *Annu. Rev. Biochem.* 70, 475–501. doi: 10.1146/annurev.biochem.70.1.475

Nakamichi, N. (2015). Adaptation to the local environment by modifications of the photoperiod response in crops. *Plant Cell Physiol.* 56, 594–604. doi: 10.1093/pcp/pcu181

Neto, E. C., Keller, M. P., Broman, A. F., Attie, A. D., Jansen, R. C., Broman, K. W., *et al.* (2012). Quantile-based permutation thresholds for quantitative trait loci hotspots. *Genetics* 191, 1355–1365. doi: 10.1534/genetics.112.139451

Nielsen, N. H., Jahoor, A., Jensen, J. D., Orabi, J., Cericola, F., Edriss, V., *et al.* (2016). Genomic prediction of seed quality traits using advanced barley breeding lines. *PLoS One* 11, e0164494. doi: 10.1371/journal.pone.0164494

Nissan, N., Cober, E. R., Sadowski, M., Charette, M., Golshani, A., and Samanfar, B. (2021). Identifying new variation at the J locus, previously identified as e6, in long juvenile 'Paranagoiana' soybean. *Theor. Appl. Genet.* 134, 1007–1014. doi: 10.1007/s00122-020-03746-2

Osnato, M. (2023). Evolution of flowering time genes in rice: From the paleolithic to the anthropocene. *Plant Cell Environ.* 46, 1046–1059. doi: 10.1111/pce.14495

Paes de Melo, B., Lourenço-Tessutti, I. T., Morgante, C. V., Santos, N. C., Pinheiro, L. B., de Jesus Lins, C. B., *et al.* (2020). Soybean Embryonic Axis Transformation: Combining Biolistic and Agrobacterium-Mediated Protocols to Overcome Typical Complications of In Vitro Plant Regeneration. *Front. Plant Sci.* 11, 1–14. doi: 10.3389/fpls.2020.01228

Pagano, M. C., and Miransari, M. (2016). "The importance of soybean production worldwide," in *Abiotic and Biotic Stresses in Soybean Production*, (Elsevier Inc.), 1–26. doi: 10.1016/B978-0-12-801536-0.00001-3

Pattang, A. (2022). A Functional Genomics Approach to Identifying Potential Candidates Underlying the E7 Maturity Locus in Soybean (Glycine max). Available at: https://curve.carleton.ca/0eda6bf9-5412-41ed-bdde-6d4787eeb169

Ping, J., Liu, Y., Sun, L., Zhao, M., Li, Y., She, M., *et al.* (2014). Dt2 Is a Gain-of-Function MADS-Domain Factor Gene That Specifies Semideterminacy in Soybean. *Plant Cell* 26, 2831–2842. doi: 10.1105/tpc.114.126938

Preece, C., Livarda, A., Christin, P. A., Wallace, M., Martin, G., Charles, M., *et al.* (2017). How did the domestication of Fertile Crescent grain crops increase their yields? *Funct. Ecol.* 31, 387–397. doi: 10.1111/1365-2435.12760

Qin, C., Li, H., Zhang, S., Lin, X., Jia, Z., Zhao, F., *et al.* (2023a). GmEID1 modulates light signaling through the Evening Complex to control flowering time and yield in soybean. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2212468120. doi: 10.1073/pnas.2212468120

Qin, C., Li, Y., Li, D., Zhang, X., Kong, L., Zhou, Y., *et al.* (2023b). PH13 improves soybean shade traits and enhances yield for high-density planting at high latitudes. *Nat. Commun.* 14, 6813. doi: 10.1038/s41467-023-42608-5

Raina, M., Pandotra, P., Salgotra, R. K., Ali, S., Mir, Z. A., Bhat, J. A., *et al.* (2017). "Genetic engineering and environmental risk," in *Modern Age Environmental Problems and their Remediation*, 69–82. doi: 10.1007/978-3-319-64501-8_4

Ramu, S. V., Rohini, S., Keshavareddy, G., Gowri Neelima, M., Shanmugam, N. B., Kumar, A. R. V., *et al.* (2012). Expression of a synthetic cry1AcF gene in transgenic Pigeon pea confers resistance to Helicoverpa armigera. *J. Appl. Entomol.* 136, 675–687. doi: 10.1111/j.1439-0418.2011.01703.x

Rao, K. S., Sreevathsa, R., Sharma, P. D., Keshamma, E., and Kumar, M. U. (2008). In planta transformation of pigeon pea: A method to overcome recalcitrancy of the crop to regeneration in vitro. *Physiol. Mol. Biol. Plants* 14, 321–328. doi: 10.1007/s12298-008-0030-2

Rasche, L., Habel, J. C., Stork, N., Schmid, E., and Schneider, U. A. (2022). Food versus wildlife : Will biodiversity hotspots benefit from healthier diets ? 1090–1103. doi: 10.1111/geb.13485

Rausenberger, J., Tscheuschler, A., Nordmeier, W., Wüst, F., Timmer, J., Schäfer, E., *et al.* (2011). Photoconversion and nuclear trafficking cycles determine phytochrome A's response profile to far-red light. *Cell* 146, 813–825. doi: 10.1016/j.cell.2011.07.023

Ray, J. D., Hinson, K., Mankono, J. E., and Malo, M. F. (1995). Genetic control of a long-juvenile trait in soybean. *Crop Sci.* 35, 1001–1006. doi: 10.2135/cropsci1995.0011183X003500040012x

Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Lv, Y., *et al.* (2019). Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. *Plants* 8, 34. doi: 10.3390/plants8020034

Rech, E. L., Vianna, G. R., and Aragão, F. J. L. (2008). High-efficiency transformation by biolistics of soybean, common bean and cotton transgenic plants. *Nat. Protoc.* 3, 410–418. doi: 10.1038/nprot.2008.9

Reddy, Y. S., Sridevi, O., Krishnaraj, P. U., Salimath, P. M., and Reddy, S. (2007). In planta strategy for gene transfer in chickpea (Cicer arietinum L .): embryo transformation. *Indian J. Crop Sci.* 2, 323–326.

Rockman, M. V., and Kruglyak, L. (2006). Genetics of global gene expression. *Nat. Rev. Genet.* 7, 862–872. doi: 10.1038/nrg1964

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., and Lippman, Z. B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* 171, 470-480.e8. doi: 10.1016/j.cell.2017.08.030

Rohini, V. K., and Rao, K. S. (2000). Transformation of peanut (Arachis hypogaea L.): A non-tissue culture based approach for generating transgenic plants. *Plant Sci.* 150, 41–49. doi: 10.1016/S0168-9452(99)00160-0

Rohini, V. K., and Sankara Rao, K. (2000). Embryo transformation, a practical approach for realizing transgenic plants of safflower (Carthamus tinctorius L.). *Ann. Bot.* 86, 1043–1049. doi: 10.1006/anbo.2000.1278

Romera-Branchat, M., Severing, E., Pocard, C., Ohr, H., Vincent, C., Née, G., *et al.* (2020). Functional Divergence of the *Arabidopsis* Florigen-Interacting bZIP Transcription Factors FD and FDP. *Cell Rep.* 31. doi: 10.1016/j.celrep.2020.107717

Saavedra, J. (2019). Early maturing soybeans in Canada and Potential market growing areas

[dissertation/master's thesis]. Iowa State University. Available at: https://lib.dr.iastate.edu/creativecomponents/240%0A

Sadowski, M. (2020). A functional genomics approach in identifying the underlying gene for the E8 maturity locus in soybean (Glycine max) [dissertation/master's thesis]. Department of Biology, Carleton University. Carleton University. Available at: https://curve.carleton.ca/5c96c758-a146-4cd6-aadd-d7d83398fd3e

Saindon, G., Voldeng, H. D., Beversdorf, W. D., and Buzzell, R. I. (1989). Genetic control of long daylength response in soybean. *Crop Sci.* 29, 1436–1439. doi: 10.2135/cropsci1989.0011183X002900060021x

Samanfar, B., Molnar, S. J., Charette, M., Schoenrock, A., Dehne, F., Golshani, A., *et al.* (2017). Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. *Theor. Appl. Genet.* 130, 377–390. doi: 10.1007/s00122-016-2819-7

Sawa, M., and Kay, S. A. (2011). GIGANTEA directly activates Flowering Locus T in *Arabidopsis* thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 108, 11698–11703. doi: 10.1073/pnas.1106771108

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., *et al.* (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Sedivy, E. J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.* 214, 539–553. doi: 10.1111/nph.14418

Shan, N., Wang, Z., and Hou, L. (2019). Identification of trans-eQTL using mediation analysis with multiple mediators. *BMC Bioinformatics* 20, 126. doi: 10.1186/s12859-019-2651-6

Shriti, S., Paul, S., and Das, S. (2023). Overexpression of CaMYB78 transcription factor enhances resistance response in chickpea against Fusarium oxysporum and negatively regulates anthocyanin biosynthetic pathway. *Protoplasma* 260, 589–605. doi: 10.1007/s00709-022-01797-4

Shultz, J. L., Kurunam, D., Shopinski, K., Iqbal, M. J., Kazi, S., Zobrist, K., *et al.* (2006). The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. *Nucleic Acids Res.* 34, 758–765. doi: 10.1093/nar/gkj050

Simmonds, D., and Canada, A. (2003). Genetic Transformation of Plants. *Genet. Transform. Plants*. doi: 10.1007/978-3-662-07424-4

Singh, M., Kaur, A., Singh, S., and Sandhu, J. S. (2021). in Planta Genetic Transformation in Pigeonpea: Occurrence and Analysis of Chimerism in Transformants. *Agric. Res. J.* 58, 989–997. doi: 10.5958/2395-146X.2021.00140.X

Singh, R. J., and Hymowitz, T. (1988). The genomic relationship between Glycine max (L.) Merr. and G. soja Sieb. and Zucc. as revealed by pachytene chromosome analysis. *Theor. Appl. Genet.* 76, 705–711. doi: 10.1007/BF00303516

Singh, S., Kumar, N. R., Maniraj, R., Lakshmikanth, R., Rao, K. Y. S., Muralimohan, N., *et al.* (2018). Expression of Cry2Aa, a Bacillus thuringiensis insecticidal protein in transgenic pigeon pea confers resistance to gram pod borer, Helicoverpa armigera. *Sci. Rep.* 8, 8820. doi:

10.1038/s41598-018-26358-9

Somers, D. A., Samac, D. A., and Olhoft, P. M. (2003). Recent advances in legume transformation. *Plant Physiol.* 131, 892–899. doi: 10.1104/pp.102.017681

Song, Z., Tian, J., Fu, W., Li, L., Lu, L., Zhou, L., *et al.* (2013). Screening Chinese soybean genotypes for Agrobacterium-mediated genetic transformation suitability. *J. Zhejiang Univ. Sci. B* 14, 289–298. doi: 10.1631/jzus.b1200278

Soy Canada (2022). Soy Canada: Canada's soybean. Available at: https://soycanada.ca/ (Accessed August 2, 2022).

Soyk, S., Müller, N. A., Park, S. J., Schmalenbach, I., Jiang, K., Hayama, R., *et al.* (2017). Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. *Nat. Genet.* 49, 162–168. doi: 10.1038/ng.3733

Sreevathsa, R. S. R., Neelima, M. G., Ramu, S. V, Rohini, S., Rani, B. M. A., Kumar, A. R. V, *et al.* (2008). In planta transformation strategy to generate transgenic plants in chickpea: proof of concept with a cry gene. *J Plant Biol* 35, 201–206.

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., and Li, Z. (2019). Genomic Selection for Yield and Seed Composition Traits Within an Applied Soybean Breeding Program. *G3 Genes|Genomes|Genetics* 9, 2253–2265. doi: 10.1534/g3.118.200917

Su, Q., Chen, L., Cai, Y., Chen, Y., Yuan, S., Li, M., *et al.* (2022). Functional Redundancy of FLOWERING LOCUS T 3b in Soybean Flowering Time Regulation. *Int. J. Mol. Sci.* 23. doi: 10.3390/ijms23052497

Sui, M., Jing, Y., Li, H., Zhan, Y., Luo, J., Teng, W., *et al.* (2020). Identification of Loci and Candidate Genes Analyses for Tocopherol Concentration of Soybean Seed. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.539460

Sun, Y., Shen, E., Hu, Y., Wu, D., Feng, Y., Lao, S., *et al.* (2022). Population genomic analysis reveals domestication of cultivated rye from weedy rye. *Mol. Plant* 15, 552–561. doi: 10.1016/j.molp.2021.12.015

Sundaresha, S., Kumar, A. M., Rohini, S., Math, S. A., Keshamma, E., Chandrashekar, S. C., *et al.* (2010). Enhanced protection against two major fungal pathogens of groundnut, Cercospora arachidicola and Aspergillus flavus in transgenic groundnut over-expressing a tobacco β 1-3 glucanase. *Eur. J. Plant Pathol.* 126, 497–508. doi: 10.1007/s10658-009-9556-6

Svabova, L., Smykal, P., and Griga, M. (2007). Agrobacterium-mediated transformation of pea (Pisum sativum L.): Transformant production in vitro and by non-tissue culture approach. *Food Legum. Nutr. Secur. Sustain. Agric. IS-GPB*, 208–220. Available at: https://www.researchgate.net/publication/235652958

Švábová, L., Smýkal, P., Griga, M., and Ondřej, V. (2005). Agrobacterium-mediated transformation of Pisum sativum in vitro and in vivo. *Biol. Plant.* 49, 361–370. doi: 10.1007/s10535-005-0009-6

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., *et al.* (2008). The *Arabidopsis* Information Resource (TAIR): Gene structure and function

annotation. *Nucleic Acids Res.* 36, 1009–1014. doi: 10.1093/nar/gkm965

Tan, Z., Peng, Y., Xiong, Y., Xiong, F., Zhang, Y., Guo, N., *et al.* (2022). Comprehensive transcriptional variability analysis reveals gene networks regulating seed oil content of Brassica napus. *Genome Biol.* 23, 1–25. doi: 10.1186/s13059-022-02801-z

Tardivel, A., Torkamaneh, D., Lemay, M.-A., Belzile, F., and O'Donoughue, L. S. (2019). A Systematic Gene-Centric Approach to Define Haplotypes and Identify Alleles on the Basis of Dense Single Nucleotide Polymorphism Datasets. *Plant Genome* 12, 180061. doi: 10.3835/plantgenome2018.08.0061

Thakare, D., Kumudini, S., and Dinkins, R. D. (2011). The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in *Arabidopsis*. *Planta* 234, 933–943. doi: 10.1007/s00425-011-1450-8

The American Soybean Association (2023). SoyStats - International: World Soybean Production (2021/2022 year). *Am. Soybean Assoc.* Available at: http://soystats.com/ (Accessed December 11, 2023).

Tigchelaar, M., Battisti, D. S., Naylor, R. L., and Ray, D. K. (2018). Future warming increases probability of globally synchronized maize production shocks. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6644–6649. doi: 10.1073/pnas.1718031115

Tsubokura, Y., Matsumura, H., Xu, M., Liu, B., Nakashima, H., Anai, T., *et al.* (2013). Genetic variation in soybean at the maturity locus e4 is involved in adaptation to long days at high latitudes. *Agronomy* 3, 117–134. doi: 10.3390/agronomy3010117

Tsubokura, Y., Watanabe, S., Xia, Z., Kanamori, H., Yamagata, H., Kaga, A., *et al.* (2014). Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean. *Ann. Bot.* 113, 429–441. doi: 10.1093/aob/mct269

Unc, A., Altdorff, D., Abakumov, E., Adl, S., Baldursson, S., Bechtold, M., *et al.* (2021). Expansion of Agriculture in Northern Cold-Climate Regions: A Cross-Sectoral Perspective on Opportunities and Challenges. *Front. Sustain. Food Syst.* 5, 663448. doi: 10.3389/fsufs.2021.663448

Van roekel, R. J., and Purcell, L. C. (2016). Understanding and increasing soybean yields., in *Proceedings of the Integrated Crop Management Conference*, ed. Iowa State University (Ames: Iowa State University, Digital Press), 1–6. doi: 10.31274/icm-180809-195

Vanhanen, S., Gustafsson, S., Ranheden, H., Björck, N., Kemell, M., and Heyd, V. (2019). Maritime Hunter-Gatherers Adopt Cultivation at the Farming Extreme of Northern Europe 5000 Years Ago. *Sci. Rep.* 9, 4756. doi: 10.1038/s41598-019-41293-z

Venters, B. J., and Pugh, B. F. (2009). How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.* 44, 117–141. doi: 10.1080/10409230902858785

Vinson, C., Chatterjee, R., and Fitzgerald, P. (2011). Transcription factor binding sites and other features in human and Drosophila proximal promoters. *Subcell. Biochem.* 52, 205–222. doi: 10.1007/978-90-481-9069-0_10

Vollmann, J., and Škrabišová, M. (2023). Going north: adaptation of soybean to long-day

environments. *J. Exp. Bot.* 74, 2933–2936. doi: 10.1093/jxb/erad105

Wan, Z., Liu, Y., Guo, D., Fan, R., Liu, Y., Xu, K., *et al.* (2022). CRISPR/Cas9-mediated targeted mutation of the E1 decreases photoperiod sensitivity, alters stem growth habits, and decreases branch number in soybean. *Front. Plant Sci.* 13, 1–12. doi: 10.3389/fpls.2022.1066820

Wang, F., Nan, H., Chen, L., Fang, C., Zhang, H., Su, T., *et al.* (2019). A new dominant locus, E11, controls early flowering time and maturity in soybean. *Mol. Breed.* 39, 70. doi: 10.1007/s11032-019-0978-3

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016a). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep.* 6, 1–10. doi: 10.1038/srep20728

Wang, J., Kong, L., Yu, K., Zhang, F., Shi, X., Wang, Y., *et al.* (2018). Development and validation of InDel markers for identification of QTL underlying flowering time in soybean. *Crop J.* 6, 126–135. doi: 10.1016/j.cj.2017.08.001

Wang, L., Li, H., He, M., Dong, L., Huang, Z., Chen, L., *et al.* (2023). GIGANTEA orthologs, E2 members, redundantly determine photoperiodic flowering and yield in soybean. *J. Integr. Plant Biol.* 65, 188–202. doi: 10.1111/jipb.13398

Wang, L., Sun, S., Wu, T., Liu, L., Sun, X., Cai, Y., *et al.* (2020a). Natural variation and CRISPR/Cas9-mediated mutation in GmPRR37 affect photoperiodic flowering and contribute to regional adaptation of soybean. *Plant Biotechnol. J.* 18, 1869–1881. doi: 10.1111/pbi.13346

Wang, W., Wang, Z., Hou, W., Chen, L., Jiang, B., Liu, W., *et al.* (2020b). GmNMHC5, a neoteric positive transcription factor of flowering and maturity in Soybean. *Plants* 9, 1–16. doi: 10.3390/plants9060792

Wang, Y., Gu, Y., Gao, H., Qiu, L., Chang, R., Chen, S., *et al.* (2016b). Molecular and geographic evolutionary support for the essential role of GIGANTEAa in soybean domestication of flowering time. *BMC Evol. Biol.* 16, 1–13. doi: 10.1186/s12862-016-0653-9

Wang, Y., Han, Y., Teng, W., Zhao, X., Li, Y., Wu, L., *et al.* (2014). Expression quantitative trait loci infer the regulation of isoflavone accumulation in soybean (Glycine max L. Merr.) seed. *BMC Genomics* 15, 680. doi: 10.1186/1471-2164-15-680

Wang, Y., Yuan, L., Su, T., Wang, Q., Gao, Y., Zhang, S., *et al.* (2020c). Light- and temperature-entrainable circadian clock in soybean development. *Plant Cell Environ.* 43, 637–648. doi: 10.1111/pce.13678

Wang, Z., Zhou, Z., Liu, Y., Liua, T., Li, Q., Ji, Y., *et al.* (2015). Functional evolution of phosphatidylethanolamine binding proteins in soybean and *Arabidopsis*. *Plant Cell* 27, 323–336. doi: 10.1105/tpc.114.135103

Watanabe, S., Hideshima, R., Zhengjun, X., Tsubokura, Y., Sato, S., Nakamoto, Y., *et al.* (2009). Map-based cloning of the gene associated with the soybean maturity locus E3. *Genetics* 182, 1251–1262. doi: 10.1534/genetics.108.098772

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., *et al.* (2011). A map-

based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering. *Genetics* 188, 395–407. doi: 10.1534/genetics.110.125062

Weller, J. L., Beauchamp, N., Kerckhoffs, L. H. J., Damien Platten, J., and Reid, J. B. (2001). Interaction of phytochromes A and B in the control of de-etiolation and flowering in pea. *Plant J.* 26, 283–294. doi: 10.1046/j.1365-313X.2001.01027.x

Wen, Y., Zhang, Y., Zhang, J., Feng, J., and Zhang, Y. (2020). The improved FASTmrEMMA and GCIM algorithms for genome-wide association and linkage studies in large mapping populations. *Crop J.* 8, 723–732. doi: 10.1016/j.cj.2020.04.008

Westra, H. J., and Franke, L. (2014). From genome to function by studying eQTL. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1896–1902. doi: 10.1016/j.bbadis.2014.04.024

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., *et al.* (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. *Plant Cell* 28, 2365–2384. doi: 10.1105/tpc.16.00158

Wiréhn, L. (2018). Nordic agriculture under climate change: A systematic review of challenges, opportunities and adaptation strategies for crop production. *Land use policy* 77, 63–74. doi: 10.1016/j.landusepol.2018.04.059

Wolfe, D. W., DeGaetano, A. T., Peck, G. M., Carey, M., Ziska, L. H., Lea-Cox, J., *et al.* (2018). Unique challenges and opportunities for northeastern US crop production in a changing climate. *Clim. Change* 146, 231–245. doi: 10.1007/s10584-017-2109-7

Wolfgang, G., and An, Y. qiang C. (2017). Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. *Mol. Breed.* 37. doi: 10.1007/s11032-016-0611-7

Wu, F., and Hanzawa, Y. (2018). A simple method for isolation of soybean protoplasts and application to transient gene expression analyses. *J. Vis. Exp.* 2018, 1–7. doi: 10.3791/57258

Wu, F., Kang, X., Wang, M., Haider, W., Price, W. B., Hajek, B., *et al.* (2019). Transcriptome-Enabled Network Inference Revealed the GmCOL1 Feed-Forward Loop and Its Roles in Photoperiodic Flowering of Soybean. *Front. Plant Sci.* 10, 1–14. doi: 10.3389/fpls.2019.01221

Wu, M., Liu, H., Lin, Y., Chen, J., Fu, Y., Luo, J., *et al.* (2020). In-Frame and Frame-Shift Editing of the Ehd1 Gene to Develop Japonica Rice With Prolonged Basic Vegetative Growth Periods. *Front. Plant Sci.* 11, 1–14. doi: 10.3389/fpls.2020.00307

Xia, Z.-J. (2013). Research Progresses on Photoperiodic Flowering and Maturity Genes in Soybean (Glycine max Merr.). *Acta Agron. Sin.* 39, 571. doi: 10.3724/sp.j.1006.2013.00571

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., *et al.* (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2155–E2164. doi: 10.1073/pnas.1117982109

Xu, H., Zhang, L., Zhang, K., and Ran, Y. (2020). Progresses, Challenges, and Prospects of Genome Editing in Soybean (Glycine max). *Front. Plant Sci.* 11, 1–19. doi: 10.3389/fpls.2020.571138

Xu, M., Xu, Z., Liu, B., Kong, F., Tsubokura, Y., Watanabe, S., *et al.* (2013). Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. *BMC Plant Biol.* 13, 91. doi: 10.1186/1471-2229-13-91

Xu, M., Yamagishi, N., Zhao, C., Takeshima, R., Kasai, M., Watanabe, S., *et al.* (2015). The soybean-specific maturity gene E1 family of floral repressors controls night-break responses through down-regulation of FLOWERING LOCUS T orthologs. *Plant Physiol.* 168, 1735–1746. doi: 10.1104/pp.15.00763

Yang, X., Li, X., Shan, J., Li, Y., Zhang, Y., Wang, Y., *et al.* (2021a). Overexpression of GmGAMYB Accelerates the Transition to Flowering and Increases Plant Height in Soybean. *Front. Plant Sci.* 12, 1–14. doi: 10.3389/fpls.2021.667242

Yang, X., Zhang, Y., Shan, J., Sun, J., Li, D., Zhang, X., *et al.* (2021b). GmIDD Is Induced by Short Days in Soybean and May Accelerate Flowering When Overexpressed in *Arabidopsis* via Inhibiting AGAMOUS-LIKE 18. *Front. Plant Sci.* 12, 1–12. doi: 10.3389/fpls.2021.629069

Yin, Z., Meng, F., Song, H., Wang, X., Xu, X., and Yu, D. (2010). Expression quantitative trait loci analysis of two genes encoding Rubisco activase in soybean. *Plant Physiol.* 152, 1625–1637. doi: 10.1104/pp.109.148312

Yuan, X., Jiang, X., Zhang, M., Wang, L., Jiao, W., Chen, H., *et al.* (2024). Integrative omics analysis elucidates the genetic basis underlying seed weight and oil content in soybean. *Plant Cell*, koae062. doi: 10.1093/plcell/koae062

Yue, L., Li, X., Fang, C., Chen, L., Yang, H., Yang, J., *et al.* (2021). FT5a interferes with the Dt1-AP1 feedback loop to control flowering time and shoot determinacy in soybean. *J. Integr. Plant Biol.* 63, 1004–1020. doi: 10.1111/jipb.13070

Yue, Y., Liu, N., Jiang, B., Li, M., Wang, H., Jiang, Z., *et al.* (2017). A Single Nucleotide Deletion in J Encoding GmELF3 Confers Long Juvenility and Is Associated with Adaption of Tropic Soybean. *Mol. Plant* 10, 656–658. doi: 10.1016/j.molp.2016.12.004

Zeng, L., Liu, X., Zhou, Z., Li, D., Zhao, X., Zhu, L., *et al.* (2018). Identification of a G2-like transcription factor, OsPHL3, functions as a negative regulator of flowering in rice by co-expression and reverse genetic analysis. *BMC Plant Biol.* 18, 1–12. doi: 10.1186/s12870-018-1382-6

Zhai, H., Lü, S., Liang, S., Wu, H., Zhang, X., Liu, B., *et al.* (2014). GmFT4, a homolog of FLOWERING LOCUS T, is positively regulated by E1 and functions as a flowering repressor in soybean. *PLoS One* 9. doi: 10.1371/journal.pone.0089030

Zhai, Q., Chen, R., Liang, X., Zeng, C., Hu, B., Li, L., *et al.* (2022). Establishment and Application of a Rapid Genetic Transformation Method for Peanut. *Chinese Bull. Bot.* 57, 327–339. doi: 10.11983/CBB21192

Zhang, C., Liu, C., Weng, J., Cheng, B., Liu, F., Li, X., *et al.* (2017a). Creation of targeted inversion

mutations in plants using an RNA-guided endonuclease. *Crop J.* 5, 83–88. doi: 10.1016/j.cj.2016.08.001

Zhang, C., Xu, X., Chen, F., Yuan, S., Wu, T., Jiang, B., *et al.* (2023). Establishment of a novel experimental system for studying the photoperiodic response of short-day dicots using soybean 'cotyledon-only plant' as material. *Front. Plant Sci.* 13, 1–11. doi: 10.3389/fpls.2022.1101715

Zhang, F., Chen, C., Ge, H., Liu, J., Luo, Y., Liu, K., *et al.* (2014). Efficient soybean regeneration and Agrobacterium-mediated transformation using a whole cotyledonary node as an explant. *Biotechnol. Appl. Biochem.* 61, 620–625. doi: 10.1002/bab.1207

Zhang, F., Ruan, X., Wang, X., Liu, Z., Hu, L., and Li, C. (2016a). Overexpression of a Chitinase Gene from Trichoderma asperellum Increases Disease Resistance in Transgenic Soybean. *Appl. Biochem. Biotechnol.* 180, 1542–1558. doi: 10.1007/s12010-016-2186-5

Zhang, L., Su, W., Tao, R., Zhang, W., Chen, J., Wu, P., *et al.* (2017b). RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nat. Commun.* 8, 1–12. doi: 10.1038/s41467-017-02445-9

Zhang, L., Yu, Y., Shi, T., Kou, M., Sun, J., Xu, T., *et al.* (2020a). Genome-wide analysis of expression quantitative trait loci (eQTL) reveals the regulatory architecture of gene expression variation in the storage roots of sweet potato. *Hortic. Res.* 7, 90. doi: 10.1038/s41438-020-0314-4

Zhang, S.-R. R., Wang, H., Wang, Z., Ren, Y., Niu, L., Liu, J., *et al.* (2017c). Photoperiodism dynamics during the domestication and improvement of soybean. *Sci. China Life Sci.* 60, 1416–1427. doi: 10.1007/s11427-016-9154-x

Zhang, S., Singh, M. B., and Bhalla, P. L. (2021a). Molecular characterization of a soybean FT homologue, GmFT7. *Sci. Rep.* 11, 1–11. doi: 10.1038/s41598-021-83305-x

Zhang, X., Wu, T., Wen, H., Song, W., Xu, C., Han, T., *et al.* (2021b). Allelic variation of soybean maturity genes e1–e4 in the huang-huai-hai river valley and the northwest china. *Agric.* 11, 1–9. doi: 10.3390/agriculture11060478

Zhang, X., Zhai, H., Wang, Y., Tian, X., Zhang, Y., Wu, H., *et al.* (2016b). Functional conservation and diversification of the soybean maturity gene E1 and its homologs in legumes. *Sci. Rep.* 6, 1–14. doi: 10.1038/srep29548

Zhang, Y., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2020b). QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked *QTL* for quantitative traits in bi-parental segregation populations. *Comput. Struct. Biotechnol. J.* 18, 59–65. doi: 10.1016/j.csbj.2019.11.005

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., *et al.* (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci.* 114, 9326–9331. doi: 10.1073/pnas.1701762114

Zhao, C., Takeshima, R., Zhu, J., Xu, M., Sato, M., Watanabe, S., *et al.* (2016). 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 Biol.* 16, 1–15. doi: 10.1186/s12870-016-

0704-9

Zheng, N., Li, T., Dittman, J. D., Su, J., Li, R., Gassmann, W., *et al.* (2020). CRISPR/Cas9-Based Gene Editing Using Egg Cell-Specific Promoters in *Arabidopsis* and Soybean. *Front. Plant Sci.* 11, 1–15. doi: 10.3389/fpls.2020.00800

Zhou, M., Luo, J., Xiao, D., Wang, A., He, L., and Zhan, J. (2023). An efficient method for the production of transgenic peanut plants by pollen tube transformation mediated by Agrobacterium tumefaciens. *Plant Cell. Tissue Organ Cult.* 152, 207–214. doi: 10.1007/s11240-022-02388-0

Zhu, J., Takeshima, R., Harigai, K., Xu, M., Kong, F., Liu, B., *et al.* (2019). Loss of function of the E1-like-B gene associates with early flowering under long-day conditions in Soybean. *Front. Plant Sci.* 9, 1–13. doi: 10.3389/fpls.2018.01867

Zhu, X., Leiser, W. L., Hahn, V., and Würschum, T. (2023). The genetic architecture of soybean photothermal adaptation to high latitudes. *J. Exp. Bot.* 74, 2987–3002. doi: 10.1093/jxb/erad064

Zia, M., Arshad, W., Bibi, Y., Nisa, S., and Chaudhary, M. F. (2011). Does Agro-injection to soybean pods transform embryos? *Plant Omics* 4, 384–390.

Zimmer, G., Miller, M. J., Steketee, C. J., Jackson, S. A., de Tunes, L. V. M., and Li, Z. (2021). Genetic control and allele variation among soybean maturity groups 000 through IX. *Plant Genome* 14, 1–25. doi: 10.1002/tpg2.20146

Zou, W., and Zeng, Z. B. (2009). Multiple interval mapping for gene expression QTL analysis. *Genetica* 137, 125–134. doi: 10.1007/s10709-009-9365-z