# Mapping of Genomic Regions Underlying the Early Flowering Trait in 'RE2', a Mutant Derived from Flax (*Linum usitatissimum* L.) Cultivar 'Royal'

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
In Partial Fulfillment of the Requirements
For the Degree of Master of Science
In the Department of Plant Sciences
University of Saskatchewan
Saskatoon

By
Akshaya Vasudevan

## Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a postgraduate degree from the University of Saskatchewan, I agree that the libraries of this University may make it freely available for inspection. I further agree that permission for copying this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

## Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Plant Sciences
51 Campus Drive,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada,
S7N 5A8

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan, Canada
S7N 5C9

# Abstract

Canada is a world leader in flax production, and the expansion of the crop into the northern region of the prairies requires early flowering, consequently early maturing cultivars to overcome the frost damage. New sources of variation for flowering time thus hold great interest. Flax genomics resources including chromosome level assembly are now sufficiently developed to examine traits with complex inheritance. An early flowering mutant 'RE2' was selected from cultivar 'Royal' after treatment with 5-Azacytidine (5-AzaC). The mutant line flowered nearly seven to 13 days earlier than the progenitor 'Royal'. A large recombinant inbred line (RIL) population encompassing 656 lines, derived from 'Royal' x 'RE2' was used to identify the potential genomic region underlying the trait. Firstly, the RIL population was phenotyped for early vigour, days to- start of flowering, full flowering, maturity and height in three field seasons (2015, 2016 and 2017) using a modified augmented design type 2, and once in the growth-cabinet. Secondly, the distributional extremes for flowering time identified from the RIL population were subjected to sequencing based bulked segregant analysis. Thirdly, the QTL-seq bioinformatics pipeline (Takagi et al. 2013) was used for the identification of SNP, which were annotated using SnpEff. QTL-seq pipeline identified a SNP upstream of the flax gene homologous to Arabidopsis *LUMINIDEPENDENS*. Later, the sequencing data were reanalysed with customized variant calling steps succeeded by statistical analysis using QTLseqr (Mansfeld and Grumet 2018), a recent improved pipeline. QTLseqr detected two genomic regions having significant association with early flowering trait on chromosomes 9 and 12. The variants in these regions were found to be associated with genes encoding LATE EMBRYOGENESIS ABUNDANT (LEA) HYDROXYPROLINE-RICH GLYCOPROTEIN FAMILY, MAINTENANCE OF MERISTEMS-LIKE, CYTOCHROME P450 87A3 and PHLOEM PROTEIN 2-A12, based on homology analysis. As 'RE2' was derived from the population resulting from the treatment of 'Royal' with the demethylating agent 5-AzaC, whole genome bisulfite sequencing data were generated to identify variation in methylation patterns and its association with early flowering. A total of 260,193 cytosines were transformed from methylated state in the late flowering bulk to the unmethylated state in the early flowering bulk, potentially owing to the hypomethylating action of 5-AzaC. Out of the 127 significant differentially methylated regions (DMRs) detected, 59 were overlapping with genes, and 35 DMRs and 33 DMRs were within the upstream- (5kb interval) and intergenic regions, respectively.

Interestingly, a cluster of significant DMRs were also present on chromosome 12. Three DMRs (on chromosomes 1, 6 and 7) were overlapping the genes whose homologues encode FASCICLIN-LIKE ARABINOGALACTAN group of proteins, and two DMRs (on chromosome 12) were present upstream to *SUPPRESSOR OF FRI 4* and *FRIGIDA-ESSENTIAL 1*. This study is first of its kind in flax, providing the basis for identifying novel epialleles underlying the early flowering phenotype.

# Acknowledgements

I would like to sincerely thank my co-supervisors – Dr. Helen Booker for this wonderful opportunity, guidance on field trials and phenotypic data analysis, thoughtful discussions and fellowship support; and Dr. Steve Robinson for his moral support, guidance in wet-lab and in bioinformatics analysis, and his excellent mentorship and intellectual discussions so that I receive training for the next phase of my academic career, through this research project.

I am thankful to my advisory committee members Dr. Tom Warkentin (chair), Dr. Karen Tanino and Dr. Tim Sharbel for sharing their reservoir of knowledge, critical inputs and for orienting me to think in an inter-disciplinary way. I specially thank Dr. Carlos Carvalho for serving as the external examiner of this thesis.

I sincerely thank Dr. Raja Ragupathy for his invaluable guidance and support throughout my programme.

I thank the members of the flax team – Dr. Gaofeng Jia, Dr. Megan House and Dr. Lester Young for their help, and I am very grateful to Mrs. Shannon Froese, Mr. Ken Jackle, Ms. Mackenzie Holland and Mrs. Barbara Boon for their wonderful support and technical assistance.

At Agriculture and Agri-Food Canada, I sincerely thank Mrs. Tricia Bender and Mrs. Kyla Horner for their sound technical guidance. I am very grateful to Mr. Miles Buchwaldt for his support in bioinformatic analysis. I specially thank Dr. Dwayne Hegedus and his lab members Mrs. Cathy Coutu, Mrs. Diana Bekkaoui, Mrs. Myrtle Harrington for the lab space and the wonderful working environment.

I am very grateful to Dr. Sylvie. J. Cloutier and Dr. Frank. M. You at the Agriculture and Agri-Food Canada, Ottawa Research and Development Centre for kindly providing the pseudomolecule-level reference sequence of flax before its publication.

The financial support from Natural Sciences and Engineering Research Council (NSERC), SeCan, SaskFlax Graduate Student Scholarship and Seed of the Year Scholarship are gratefully acknowledged.

I thank my wonderful parents K. Vasudevan and Leela Vasudevan, and my sister Aardhra Vasudevan for being with me in all the ups and downs and believing in me more than I did. A deep thanks to Meenamma, whom I wish was still around to see me completing the programme. My special thanks to my friends Abiraami, Aberame, Shruthi, Barath, Manpreet, Surendra, Amara and Dr. Sampath Perumal, for all their support.

**Dedicated to**
My Grandparents
My Parents
My Sister and
My Guru

## Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations

| | |
|---|---|
| 5-AzaC | 5-Azacytidine |
| AFLP | Amplified Fragment Length Polymorphism |
| BLAST | Basic Local Alignment Search Tool |
| BSA | Bulked Segregant Analysis |
| BWA | Burrows Wheeler Aligner |
| C | Cytosine |
| CV | Coefficient of Variation |
| DMRs | Differentially Methylated Regions |
| DNA | Deoxyribonucleic Acid |
| *epi*RIL | epigenetic Recombinant Inbred Line |
| GATK | Genome Analysis Tool Kit |
| GCV | Genotypic Coefficient of Variation |
| GFF | General Feature Format |
| gVCF | genomic Variant Call Format |
| LD | Linkage Disequilibrium |
| LINE | Long Interspersed Nuclear Element |
| NGS | Next Generation Sequencing |
| PCR | Polymerase Chain Reaction |
| qPCR | quantitative Polymerase Chain Reaction |
| QTL | Quantitative Trait Locus |
| RIL | Recombinant Inbred Line |
| RNA | Ribonucleic Acid |
| RSB | Resuspension Buffer |
| SAM | Sequence Alignment Map |
| siRNA | small interfering Ribonucleic Acid |
| SNP | Single Nucleotide Polymorphisms |
| SPB | Sample Purification Beads |
| T | Thymine |
| TAIR | The Arabidopsis Information Resource |
| WGBS | Whole Genome Bisulfite Sequencing |

# Chapter 1 Introduction

Canada is the world leader in flax production and exports (www.saskflax.ca) with the majority of flax grown in the Province of Saskatchewan (~75% based on Canada Grains and Oilseeds Outlook, by AAFC). However, the climate throughout the Canadian Prairie provinces is harsh, resulting in a short growing season with a duration of 110 frost-free days (FFD). This environment imposes limits on flax production in Saskatchewan (Figure 1.1). In order to expand the area of flax production beyond current growing areas typically restricted to areas of southern Saskatchewan and Manitoba with a greater number of FFD, into the more northern parts of the prairies, there is a need to develop early flowering and subsequently early maturing cultivars, able to escape the potential frost damage during the physiological seed maturity phase. In addition, with the changing climate, cropping areas are expected to expand into further northern latitudes in the 21st century (King et al. 2018), and hence, developing cultivars with adaptation to these regions is of prime interest. Further, early flowering genotypes possibly escape high temperature and drought stress, and breeding for earlier flowering has been carried out in other crops in western Canada including chickpea (Bueckert and Clarke 2013).

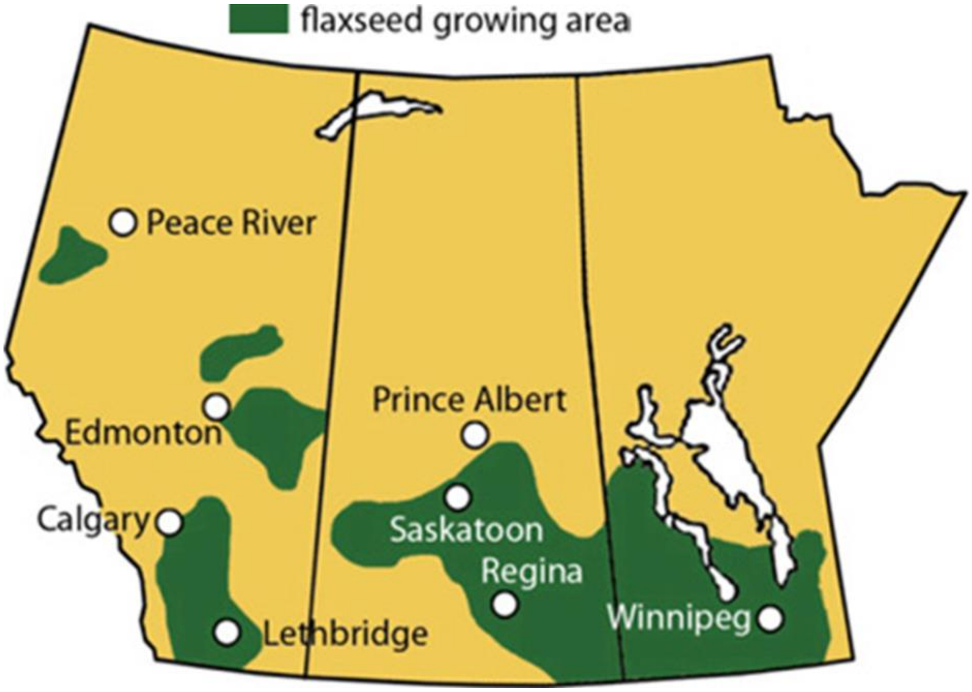

**Figure 1.1** Flax growing regions in Canada (Image source: Flax Council of Canada)

Flowering time is a complex trait governed by nearly 300 genes associated with eight different physiological pathways. These include responses to photoperiod, aging, vernalization, ambient temperature and regulations through the circadian clock, phytohormones, nutrition (particularly trehalose-6-phosphate; Wahl et al. 2013) and the autonomous flowering pathways. In Arabidopsis, the molecular mechanisms underlying flowering time have been elucidated in detail (Bouché et al. 2015). Key floral integrator genes including, *FLOWERING LOCUS T* (*FT*; Yoo et al. 2005) and *TWIN SISTER OF FT* (Yamaguchi et al. 2005), that govern the transition from vegetative to reproductive phase act as convergence points linking these diverse pathways (Andrés and Coupland 2012). In contrast to Arabidopsis, the genetic control of flowering time is not well understood in non-model plant species and direct extrapolation from Arabidopsis to flax might be misleading since they diverged ~106 million years ago (http://www.timetree.org/; Hedges et al. 2006).

Variation in flowering time occurs naturally and the cultivar ‘Royal’ is among the moderately early flowering flax genotypes (McGregor 1953). ‘Royal’ is an older flax cultivar, that has been superseded by other cultivars with improved agronomic performance including traits leading to higher yield and oil content, along with improved resistance to diseases. However, the alleles controlling flowering time remain valuable towards the development of earlier flowering flax varieties. Among the 29 flax cultivars that have been registered in Canada since 2000, ‘2126’, a low linoleic acid containing flax cultivar is the earliest, taking ~95 days to maturity (Dribnenki et al. 2005). Although this improvement is significant, it has been estimated that a cultivar with duration of ~90 days is desirable for northern part of the prairies (Duguid 2009). Efforts to further reduce the flowering time of ‘Royal’ include unconventional approaches including altering chromatin structure through epigenetic modification (Fieldes 1994; Fieldes and Amyot 1999). These approaches not only provide additional insight into the regulation of flowering time in flax, but also explore the potential prospects for exploiting both genetic and epigenetic variation towards crop improvement.

Methodologies to alter chromatin structure can be applied to plants using pharmacological inhibitors that restrict chromatin modifications. One such approach targets alterations to DNA methylation patterns using 5-Azacytidine (5-AzaC), a cytidine analog that inhibits methylation (Veselý et al. 1978). The utility of this approach was demonstrated in flax where three early flowering lines named ‘RE1’, ‘RE2’ and ‘RE3’ were selected from a population of the variety ‘Royal’ after mutagenesis using 5-AzaC (Fieldes 1994). The early flowering trait identified after 5-AzaC treatment was found to be heritable and was observed to be stably transmitted through meiosis for at least nine generations (Sun 2015, M. Sc., Thesis, University of Saskatchewan). The heritability of this trait enabled the generation of three recombinant inbred line (RIL) populations developed by crossing each of the three early flowering lines to the progenitor genotype ‘Royal’. The crosses and RIL populations were developed at the Crop Development Centre, University of Saskatchewan. These RIL populations offer a unique genetic resource to elucidate the underlying factors controlling the variation in these early flowering lines. These genetic resources were examined to characterize the factors underlying the observed variation in flowering time between ‘Royal’ and the early flowering line ‘RE2’.

The use of a DNA methylation inhibitor to induce new phenotypic variation has the potential to do so without altering the favourable allelic combinations selected by flax breeders in their vision for ‘Royal’ since, negligible primary sequence variation is induced by 5-AzaC (Xu et al. 2016). The use of 5-AzaC to induce new phenotypic variation suggests that the underlying variation might result from alterations in DNA methylation patterns between ‘RE2’ and its progenitor genotype ‘Royal’. Although, the possibility exists that exposure to 5-AzaC might act as a weak mutagen inducing a genetic mutation (Single Nucleotide Polymorphisms-SNP or chromosomal rearrangements) resulting in the observed phenotypic variation. The stable inheritance of early flowering time in ‘RE2’, and the moderate heritability of flowering time trait in ‘RC’ x ‘RE2’ RIL population, probably suggest that a genetic mutation might underlie this variation. However, the origin and the number of loci responsible for this variation is unknown and is the subject of this thesis.

## 1.1 Objectives and hypotheses tested

Since the genetic basis of variation for the phenotype of early flowering time in ‘RE2’ is unknown, there is a need to uncover the underlying cause. Firstly, the trait was identified from a population treated with 5-AzaC suggesting that changes in DNA methylation control the variant flowering time trait. However, the trait was found to be stable for multiple generations suggesting, a potential genetic control might have been induced and selected. The compound 5-AzaC is a known hypomethylating chemical suggesting a possible epigenetic basis for the origin of early flowering trait, though the mutagenic potential of 5-AzaC (i.e. its ability to cause genetic mutations) has not been thoroughly investigated and genetic variation remains a possibility. This thesis characterizes the basis (genetic or epigenetic) of the trait and was taken up for investigation with the following objectives:

1. To perform field-based phenotyping of recombinant inbred line (RIL) mapping populations (‘Royal’ x ‘RE2’ and ‘RE2’ x ‘Royal’) using modified augmented design type 2 (MAD2).

2. To identify individuals representing the distributional extremes of the phenotypic spectrum for flowering time (early- and late- flowering) corrected for potential soil heterogeneity.

3. To perform whole genome sequencing-based BSA using individuals representing extremes of the variation for flowering time, for identifying causative or linked SNP underlying the phenotype.

4. To perform bisulfite sequencing of individuals representing extremes of the variation for flowering time for identification of potential hypomethylated regions associated with early flowering phenotype.

**Null hypothesis:** ‘The phenotypic variation, induced by 5-AzaC, observed for flowering time in ‘RE2’ as early flowering trait, is governed by specific genomic regions (Quantitative trait loci - QTLs) harboring candidate genes underlying the trait’.

**Alternative hypothesis**: ‘The early flowering phenotype of ‘RE2’, induced by 5-AzaC, is conditioned by the epiallelic state of certain discrete loci, beyond DNA sequence variation’.

# Chapter 2 Literature Review

## 2.1 Flax

Flax (*Linum usitatissimum* L.) taxonomically placed under the family *Linaceae*, is a multifunctional crop. Flax, along with seven other crops belonging to Neolithic agriculture were unearthed in the fertile crescent zone (Lev-Yadun et al. 2000) indicating its early domestication. The crop was introduced into Canada ~400 years ago by French immigrants (Atton 1989) who later spread it across the continent. There was an increase in demand for flaxseed oil during the second world war, as well as a rise in world flax consumption in the mid-20th century increasing the scope of commercial flax production in the country (www.saskflax.ca). Today Canada has grown into the world's leading producer (591,000 tonnes, 2016-2017 data; source: Statistics Canada) and exporter (500,000 tonnes, 2016-2017 data; source: Statistics Canada) of the crop with most of the production since 1993/94 in the province of Saskatchewan (www.saskflax.ca).

## 2.2 Significance of early flowering lines

Flax production in Canada is currently restricted to the southern parts of Saskatchewan and Manitoba due to the risk of damage by early fall frost in the northern part of prairies at maturity phase. Frost during the maturity stage of the crop results in unfilled seeds, lower oil content and reduced ability for germination (Gubbels et al. 1994). However, enhanced seed quality due to lower ambient temperature and longer photoperiods lead to higher concentrations of unsaturated fatty acids in the linseed oil exhibiting the advantages of flax production in northern prairies (Dillman and Hopper 1943; Sosulki and Gore 1964). Hence, the development of early flowering cultivars would help to overcome the frost damage in northern prairies and make harvest easier by avoiding tangling of green stems in farm equipment in the current growing regions as well, without compromising on oil quality.

## 2.3 Flax breeding

Flax is a diploid (2n=2x=30) and self-pollinating annual requiring a growing period of 90-150 days to reach maturity (Diederichsen and Richards 2003). ‘Non-shattering’ was the earliest selected trait since domestication of the crop (Cullis 2007). In Canada, there has been flax breeding since early 1900s. Breeding objectives were increased yield and disease resistance, initially for wilt-resistance followed by rust-resistance (Cullis 2007). Currently, breeding objectives vary with requirements of different production areas (Duguid 2009) including yield, early maturity (early flowering), oil content and fatty acid profile, and resistance to evolving races of the pathogens. Flax being an autogamous crop involves the development of pure-line varieties through conventional breeding methods such as pedigree and single seed descent. The absence of male sterility in flax has made it difficult to exploit heterosis (Hall et al. 2016). The University of Saskatchewan’s Crop Development Centre’ (CDC) flax breeding program was established in 1974 to fulfil the needs of the provincial producers.

### 2.4 Molecular basis of flowering time

Flowers bear the reproductive organs which produce male and female gametes of the plants, further producing seeds which are of high economic significance. The switch controlling the transition from vegetative to reproductive phase is critical, where the shoot apical meristem is transformed to an inflorescence meristem. Hence, manipulation of flowering time remains one of the prime objectives in crop breeding.

Flowering time is controlled by both external and internal environmental cues such as photoperiod, vernalization and hormonal signals (Andres and Coupland 2012).

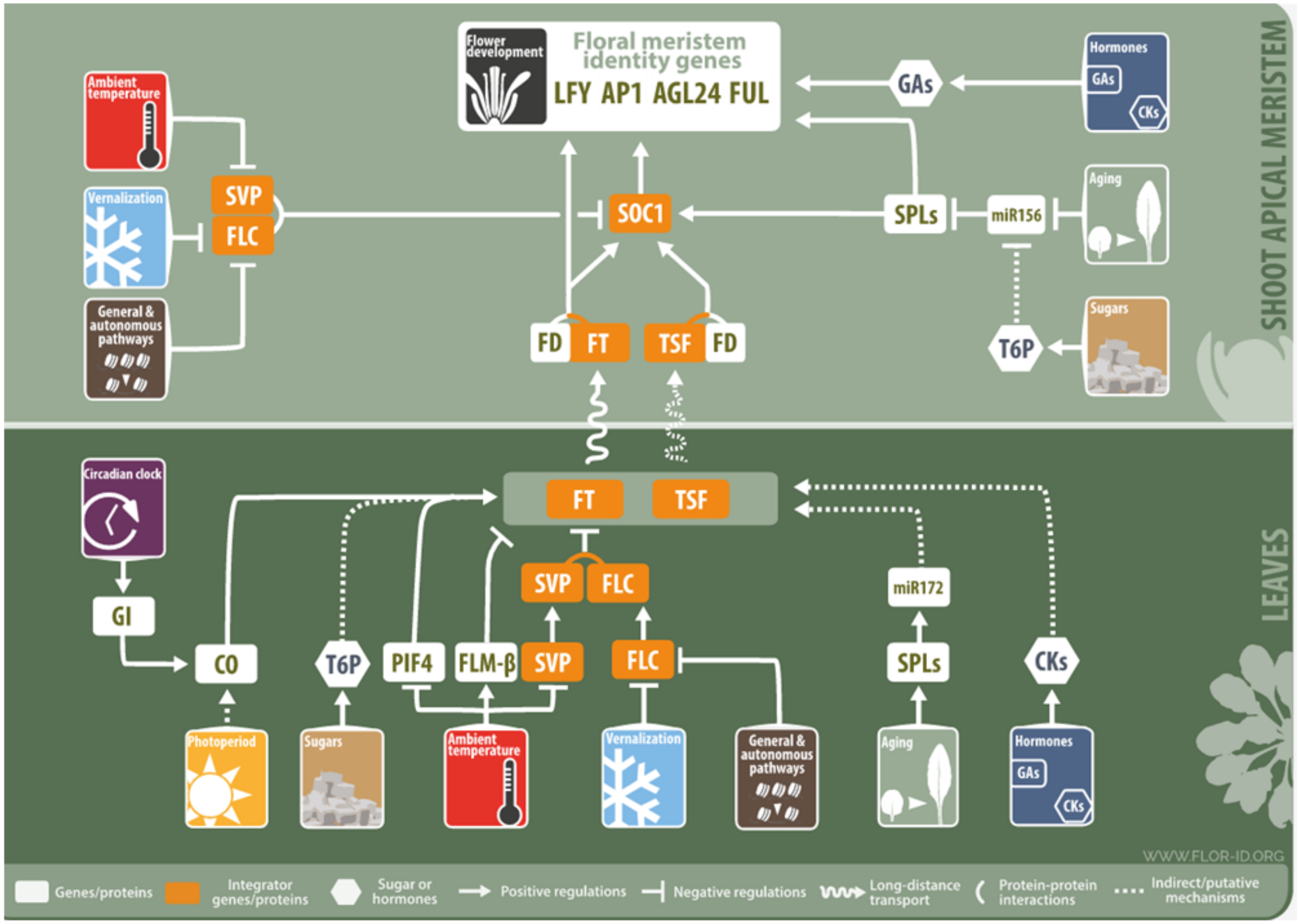

**Figure 2.1** Overview of pathways involved in controlling flowering time (Used with permission from Bouché et al. 2015).

<u>SHOOT APICAL MERISTEM</u>: *SOC - SUPPRESSOR OF OVEREXPRESSION OF CONSTANS; FD - FLOWERING LOCUS D;* <u>FLORAL MERISTEM IDENTITY GENES:</u> *LFY- LEAFY; AP1 - APETALA 1; AGL24 - AGAMOUS-LIKE 24; FUL - FRUITFULL;* <u>LEAVES:</u> *GI - GIGANTEA; CO - CONSTANS; PIF4 - PHYTOCHROME INTERACTING FACTOR 4; FLM - FLOWERING LOCUS M; FLC - FLOWERING LOCUS C; FLT - FLOWERING LOCUS T; TSF - TWIN SISTER OF FT; SPL - SQUMOSA PROMOTER BINDING PROTEIN LIKE; SVP - SHORT VEGETATIVE PHASE*

Eight distinct genetic pathways are involved in flowering time:

1. Photoperiod pathway
2. Autonomous pathway
3. Circadian clock
4. Vernalization pathway
5. Aging pathway
6. Thermosensory pathway
7. Gibberellin signalling
8. Endogenous sugars

The floral integrator genes such as *FLOWERING LOCUS T (FT)* act as the convergence point of these diverse pathways (Andres and Coupland 2012) (Figure 2.1).

### 2.4.1 Photoperiod and Circadian clock pathways

Early experiments have shown that day length has a major influence on flowering time, classifying plants into short day, long day and day neutral based on their responses (Garner and Allard 1920). The study of the model organism *Arabidopsis thaliana* has led to the dissection of the underlying pathway. *CONSTANS (CO)* and *FT* have a principal role in the response to day length (Koorneef et al. 1991).

Classical experiments indicate that the signals for flowering are produced in the leaves and transmitted to the shoot apex through the phloem which was named ‘florigen’. The FT protein is transmitted through the phloem in a similar pattern (Corbesier et al. 2007) and hence it has been hypothesized as the ‘florigen’.

In photoperiod-sensitive plants, circadian clock is responsible for day length measurement. The clock regulated genes by means of both internal and external coincidence control light dependent flowering. The mediation between circadian rhythm and the photoperiod pathway is done by CO which is regulated by the plant clock-controlled *FLAVIN-BINDING, KELCH REPEAT, F-BOX 1 (FKF1) and GIGANTEA (GI)* (Suarez-Lopez et al. 2001; Greenham and McClung 2015). *GI* expression is complexly linked with other genes of the circadian clock such as *CIRCADIAN CLOCK-ASSOCIATED 1 (CCA1), LATE-ELONGATED HYPOCOTYL (LHY)* and *EARLY FLOWERING 3 (ELF3)* (Fowler et al., 1999). *CRYPTOCHROME 2 (CRY2)* and *PHYTOCHROME A (PHYA)* photoreceptors are essential for enhanced *CO* expression (Yavnovsky and Kay 2002). The level of *CO* expression is found corresponding to the FKF1-GI complex formed in the presence of blue light, by degradation of *CYCLING DOF FACTORS (CDF)* (Sawa et al. 2007; Fornara et al. 2009) and enhancing its stability (Song et al. 2012). Also, *CRY2* mediates blue light mediated suppression of *CONSTITUTIVE PHOTOMORPHOGENIC 1* (*COP1)* and *SUPPRESSOR OF PHYTOCHROME A 1* (*SPA1)* complex which is responsible for proteolysis of CO (Zuo et al. 2011).

Under long day conditions high levels of *CO* mRNA is found only in the end of light period which cannot be achieved under short day conditions (Suarez-Lopez et al. 2001) considering the stabilization of CO under light, this facilitates early flowering under the former condition. *CO* binds to *FT* promoter region and brings about its transcriptional activation (Tiwari et al. 2010). *FD* transcribed bZIP domain transcription factor interacts with *FT* and has a role to play in its expression (Abe et al. 2005). *CO* through *FT* also activates *SOC1* (Yoo et al. 2005) which along with *FT* activates the floral meristem identity genes such as *LEAFY (LFY), APETALA (AP1), AGAMOUS-LIKE 24 (AGL24)*, *FRUITFULL (FUL).*

### 2.4.2 Vernalization pathway

FLOWERING LOCUS C (FLC), a MADS box domain protein is primarily a repressor of flowering initiation (Michaels and Amasino 1999). *FLC* is epigenetically silenced by prolonged exposure to cold, a process called vernalization. The duration of cold treatment is proportional to the extent to which flowering is accelerated. Two plant homeodomain proteins VERNALIZATION INSENSITIVE 3 (VIN3) and VERNALIZATION 5 (VRN5) associated with Polycomb Repressor Complex 2 (PRC2) accumulate in the region of *FLC*, containing its promoter, first exon and parts of the first intron, and modify lysine 27 of histone-3 (H3) with trimethylation (Qüesta et al. 2016). From a mechanistic perspective, during vernalization, a three-dimensional chromatin loop involving the promoter, first exon and intron of the *FLC* gene and the downstream promoter of the gene encoding a long non-coding RNA (COOLAIR) is disrupted leading to repression of *FLC* (Zhu et al. 2015).

### 2.4.3 Aging and thermosensory pathways

The *SQUAMOSA PROMOTER BINDING LIKE (SPL)* genes are found to be immediate activators of the floral meristem identity genes but are repressed by the microRNA 156 (miR156) during the juvenile stages of the plant. With aging, there is decrease in the levels of the miR156 facilitating *SPL* expression (Fornara and Coupland 2009) leading to transition to flowering phase.

While low temperature treatment (vernalization) is required for flowering in few crops, including a few species of flax, control of flowering under low temperature is brought about by SHORT VEGETATIVE PHASE-FLOWERING LOCUS M-β (SVP-FLMβ) complex belonging to the MADS box transcription factor family by repressing the *FT* gene (Pose et al. 2013). Under warmer conditions, the transcription factor PHYTOCHROME INTERACTING FACTOR 4 (PIF4) binds to *FT*, the floral integrator and activates its transcription (Kumar et al. 2012).

### 2.4.4 Floral organ development

During the process of flower development, the transition to floral meristem is by the activation of *LEAFY (LFY), APETALA 1 (AP1), PISTILLATA (PI),* the floral meristem identity genes, by an integrator of varied pathways responding to environmental cues, such as the FT. In the floral meristem, *AP1* and *AGAMOUS LIKE-24 (AGL24),* the determinants of the inflorescence meristem fate, are repressed (Yu et al. 2004). These homeotic genes which determine the floral organ identity are grouped into three classes namely, A, B and C, influencing the formation of different whorls of a flower, though their functions are overlapping (Bowman et al., 1991). LFY along with *UNUSUAL FLORAL ORGANS (UFO)* and *AP1* activates *APETALA 3 (AP3)* and along with *WUSCHEL (WUS)* activates *AGAMOUS (AG)* (Ng and Yanovsky 2001; Lenhard et al., 2001).

The proper development of petals, stamens and carpels requires *SEPALATA 1 (SEP1), SEPALATA 2 (SEP2)* and *SEPALATA 3 (SEP3),* called the class E genes, the triple mutants of which result in the development of sepals in all the whorls (Pelaz et al., 2000). The class D group of genes confer ovule identity (Colombo et al., 1995). In *Arabidopsis thaliana*, *SEED STICK (STK)* and *SHATTERPROOF 1 (SHP1)* and *SHATTERPROOF 2* (*SHP2)* along with *AG* influence the ovule identity (Table 2.1). The A, B, C and E class genes, except *AP2* encode MADS domain transcription factors. They form ‘protein quartets’ (complexes of four proteins) which control the genes of floral organ development.

**Table 2.1** Classes of genes involved in flower development

| Class | Genes | Function |
|---|---|---|
| A | *AP1 and AP2* | Sepal identity in whorl-1 and petal identity in whorl-2 |
| B | *AP3 and PI* | Petal identity in whorl-2 along with class A and stamen identity in whorl-3 along with class C |
| C | *AG* | Stamen identity in whorl-3 and carpel identity in whorl-4 |
| D | *STK, SHP1* | Ovule identity |
| E | *SEP1, SEP2, SEP3* | Petals, stamen and carpel development |

The cadastral genes restrict the boundary for the activity of other genes. *LEUNIG (LUG)* and *SEUSS (SEU)* repress the activity of the gene *AG* in the whorls 1 and 2 of the flower (Sridhar et al., 2004), while *SUPERMAN (SUP)* maintains the boundary of *AP3*, since its mutant resulted in the development of stamens in the innermost whorl, thus its function being specifying boundaries for male and female floral organs and also the termination of floral meristem (Breuil-Broyer et al., 2016).

During the juvenile stages of the crop, all these floral organ identity genes are repressed by genes such as *EMBRYONIC FLOWER 1 (EMF1)*, *EMF2* and *FERTILIZATION-INDEPENDENT ENDOSPERM (FIE)* (Chanvivattana et al., 2004).

### 2.5 Mapping of genomic regions underlying flowering time

Since the first study of tagging of a quantitative trait (seed size) using a qualitative trait (seed-color; Sax 1923), quantitative trai locus (QTL) mapping strategies have been widely used to identify genomic regions harboring candidate genes underlying various phenotypes in several crops. Advent of DNA markers helped accurate positioning of genomic regions controlling target phenotypes. Conventional QTL mapping studies have identified genomic regions governing flowering time in crops such as *Brassica napus*, grapevine, watermelon and tomato (Table 2.2).

**Table 2.2** Studies of QTL mapping for flowering time in crop plants

| Crop | Number of identified QTLs for flowering time | Reference |
|---|---|---|
| Rapeseed (*Brassica napus*) | 4 | Luo et al. 2014 |
| Rapeseed (*Brassica napus*) | 4 | Liu et al. 2016 |
| Grapevine | 6 | Duchêne et al. 2012 |
| Grapevine | 8 | Fechter et al. 2014 |
| Tomato | 2 | Nakano et al. 2016 |
| Watermelon | 1 | McGregor et al. 2014 |

### 2.6 Effect of 5-Azacytidine on plant genome

The eukaryotic genomes carry epigenetic marks on DNA and histones (Suzuki and Bird 2008). DNA methylation may occur on cytosine in varied sequence contexts such as at symmetric sites CG, CHG and also asymmetric sites CHH (Law and Jacobsen 2010).

5-Azacytidine (5-AzaC), a DNA methylation inhibitor is a cytidine analog with a nitrogen atom in place of carbon at the fifth position of the ring (Figure 2.2). During cytosine methylation reaction, ‘C’ at the sixth position of the cytosine ring forms a covalent bond with DNA methyltransferases which transfers the methyl group from S-adenosyl methionine of metabolite pool to the ‘C’ at the fifth position of the ring. Methyltransferases will be released from its covalent linkage after the methyl-transfer reaction (Zhang et al. 2013). However, the non-methylable nature of 5-AzaC due to the presence of ‘N’ in place of ‘C’ in the fifth position of the ring, prevents the methyl-transfer reaction needed for the release of methyltransferases. This leads to genome-wide demethylation due to reduction in the number of available DNA methyltransferases, in addition to specific hypomethylation at sites where 5-AzaC is incorporated (Pecinka and Liu 2014).

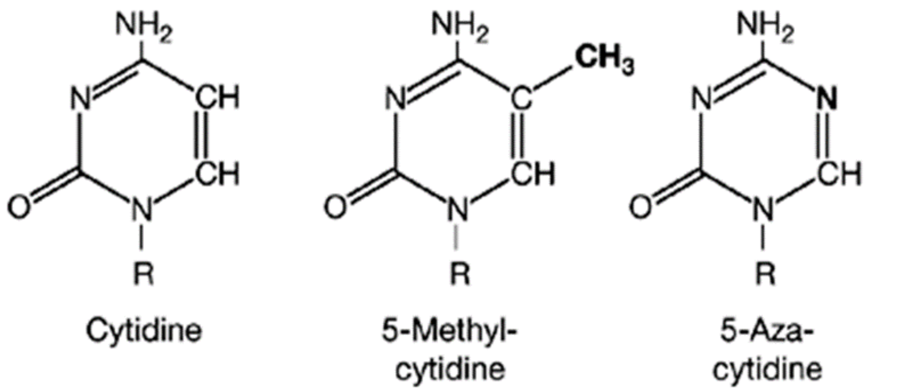

**Figure 2.2** Chemical structure of cytidine and 5-Azacytidine (Used with permission from Claus and Lübbert 2003)

Currently, bisulfite sequencing remains the most efficient technology to investigate the methylation status at single nucleotide level (Figure 2.3). During this process, the bisulfite anion gets added at the C5-C6 double bond of unmethylated cytosine generating a cytosine sulfonate which in turn undergoes hydrolysis and deamination to form uracilsulfonate. Under alkaline condition, the sulfonate group is released giving rise to uracil. This event when followed by PCR amplification and sequencing, 5-methylcytosine is read as cytosine while the unmethylated cytosine as thymine (Peng et al. 2016). 5-Azacytidine has been used to study the effects of differential methylation of DNA on phenotypes of various crop and non-crop species over years. Some of the studies are listed below (Table 2.3).

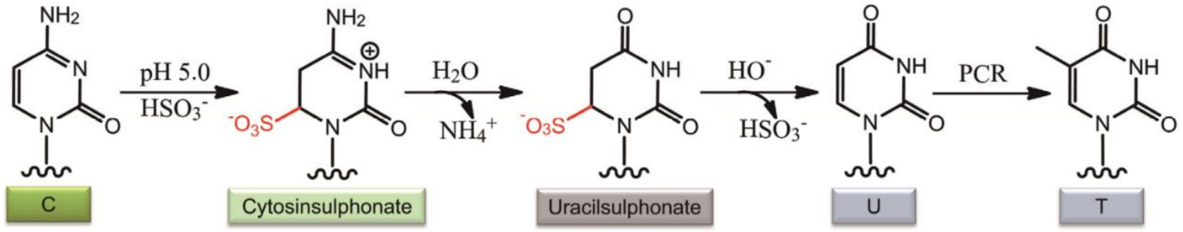

**Figure 2.3** Principle of bisulfite sequencing (Used with permission from Peng et al. 2016)

**Table 2.3** Effect of 5-Azacytidine on plant species

| Crop | Genotypic/phenotypic observation | Reference |
|---|---|---|
| A. Crop plants | | |
| Rice | Reduced plant height | Sano et al. 1990 |
| Triticale | Observation of inherited hypomethylation | Heslop-Harrison 1990 |
| Tobacco | Hypomethylated repetitive DNA sequences which are inherited; dwarf phenotypes | Vystok et al. 1995 |
| Triticale | The expression of rye rDNA in treated plants is higher | Amado et al. 1997 |
| Wheat | Reduces vernalization requirement in winter wheat | Horvath et al. 2003 |
| Wheat | Increase in callus formation rate helping in developing doubled haploid population | Belchev et al. 2004 |
| Rice | Inheritance of dwarf phenotype and resistance to *Xanthomonas oryzae pv oryzae* as the result of hypomethylation | Akimoto et al. 2007 |
| *Brassica rapa* | Lines with decreased flowering time and epialleles showing prospects of hypomethylated population in crop breeding | Amoah et al. 2012 |
| Wild potato | Early flowering phenotypes | Marfil et al. 2012 |
| Sugarcane | Lines showing smut tolerance and herbicide (Imazapyr) tolerance | Munsamy et al. 2013 |
| Spinach | Lines showing reduction in days to flowering and monoecy | Li et al. 2015 |
| B. Non-crop plants | | |
| *Melandrium album* | Expression of andromonoecy in male plants | Janousek et al. 1996 |
| *Perilla frutescens* var. *crispa* | Flowering under non-inductive photoperiods and dwarf phenotype | Kondo et al. 2010 |
| *Silene armeria* | Flowering under non-inductive photoperiod | Kondo et al. 2010 |

| | | |
|---|---|---|
| Trifoliate orange | Increased expression of citrus floral integrator and floral identity genes | Zhang et al. 2014 |
| Strawberry | Early- and late- flowering phenotypes, reduced rosette diameter | Xu et al. 2016 |

From the above studies (Table 2.3), it is evident that 5-AzaC is known to induce genome-wide DNA demethylation with potential impact on epigenetic regulation of gene expression. However, the true mutagenic potential of 5-AzaC has not been investigated in detail, so far. Preliminary results from comparison of resequencing data from ‘Royal’ and its three early flowering derivatives (‘RE1’, ‘RE2’ and ‘RE3’) had uncovered around 400 high quality SNP. (Dr. Stephen. J. Robinson, personal communication). The SNP were filtered with a quality score cut-off of 5000, which is a phred-like probability score, scaled-up to measure the quality of variant call for a given nucleotide position in the reference (De Pristo et al. 2011). However, the origin of these SNP need to be investigated further since there is a lack of evidence in literature in support of the potential of 5-AzaC to induce point mutations. Moreover, since ‘RE1’, ‘RE2’, ‘RE3’ used for generating RILs are nine generations past 5-AzaC treatment (Sun 2015, M. Sc., Thesis, University of Saskatchewan), the SNP may represent the accumulated point mutations over multiple generations, and thus still can be used as tags for Differentially Methylated Regions (DMRs) between normal and early flowering lines.

A large body of evidence suggested that hypomethylated genomic condition would lead to activation of transposable elements because of the resetting of the epigenetic mechanisms repressing the transcription and transposition of these elements (Grandbastien 2015). Transposition of jumping genes would create genetic variation including insertions-deletions (InDels). Changes in DNA methylation levels may lead to significant heritable phenotypic variation in eukaryotes (Johannes et al. 2008). The hypomethylation of transposable elements lead to their transposition into genes resulting in phenotypic variation as well (Miura et al. 2001).

### 2.7 Early flowering studies in flax

Treatment of flax seeds with 5-AzaC led to modified phenotypes such as decreased plant height, fewer leaves and reduction in time to flower (Fieldes 1994). The study of successive progeny of these lines showed stable inheritance of the modified phenotypes (Fieldes and Amyot 1999). The earliness in flowering was suggested to be due to hastening of the late vegetative phase (Fieldes and Harvey 2004). Studies on response to photoperiod of the three epimutant lines ‘RE1’, ‘RE2’ and ‘RE3’ along with five adapted Canadian cultivars namely ‘CDC Sorrel’, ‘CDC Bethune’, ‘Flanders’, ‘Prairie thunder’ and ‘Royal’ were carried out by Jia Sun (2015) in the Universty of Saskatchewan by transfering plants between short day and long day conditions at different time points. The results indicate that the five Canadian cultivars are highly sensitive to photoperiod and the epimutant line ‘RE2’, earliest flowering among all the eight lines studied, was the least photoperiod sensitive line.

### 2.8 Principle of bulked segregant analysis (BSA)

Most traits of agricultural importance are quantitative in nature involving many genes with minor effects, environmental effects and their interaction (Holland 2007). Identification of these quantitative trait loci has a significant impact from a breeding perspective. Bulked segregant analysis was proposed by Michelmore et al. (1991) as a quick method to identify markers to genomic regions controlling distinct phenotypes. Two bulks, each with extreme values for the trait of interest are generated from the segregating population of a single cross. The individuals within each pool/bulk are uniform for the particular trait but not for other loci and hence, the marker that is polymorphic between the two pools is linked to the genomic region responsible for the trait of interest. The application of bulk segregant analysis was extended to identify the markers linked with QTL in later studies (Mansur et al. 1993).

The segregating population with phenotypic extremes for BSA include $F_2$, $BC_1$, Doubled Haploids (DHs) and Recombinant Inbred Lines (RILs) of which DHs and RILs are most preferred because of their homozygosity which is maintained over generations, making them suitable for evaluation under different environments over years (Zou et al. 2016). The BSA combined with NGS also requires precise phenotyping (Sun et al. 2010) since the power of this strategy is primarily dependent on the accuracy to group the individuals based on the phenotypic value (Zou et al. 2016).

### 2.9 Principle of quantitative trait locus -sequencing (QTL-seq)

The recent QTL-seq methodology is a novel strategy which combines the advantages of both BSA and whole genome resequencing, enabling the identification of genomic regions accountable for the extreme trait values between the two bulks and also among the two parents (Figure 2.4). As QTL-seq involves neither marker development nor individual marker-based genotyping, it is cost effective and less time consuming than conventional QTL analysis.

Two parents with contrasting phenotypes for the trait of interest are crossed to generate a mapping population segregating for the trait. When more number of loci are involved, the frequency of the trait of interest will show a normal distribution in the $F_2$. The phenotype is scored in the progeny of the mapping population based on which two bulks with highest and lowest values are generated. The DNA of the bulks are resequenced and aligned to the reference genome. The bulked DNA is expected to contain genomic regions from both parents for most of the part except those segments harboring the QTL for the trait of interest.

A statistical parameter called SNP index (ratio between the count of alternate SNP to total number of reads aligned to the reference assembly corresponding to the SNP position) generated by genome-wide scan (chromosome-wise) will help to identify regions underlying the mutant phenotype. The observed SNP-index is ‘0’ when all the short reads from resequencing are the same as the reference genome sequence and it is 1 while all the short reads are as that of the other parent. A SNP-index of 0.5 would indicate the equal representation of both parental genomes in the bulk. Candidate genes underlying a few important agronomic traits have already been successfully mapped using the QTL-seq method (Table 2.4).

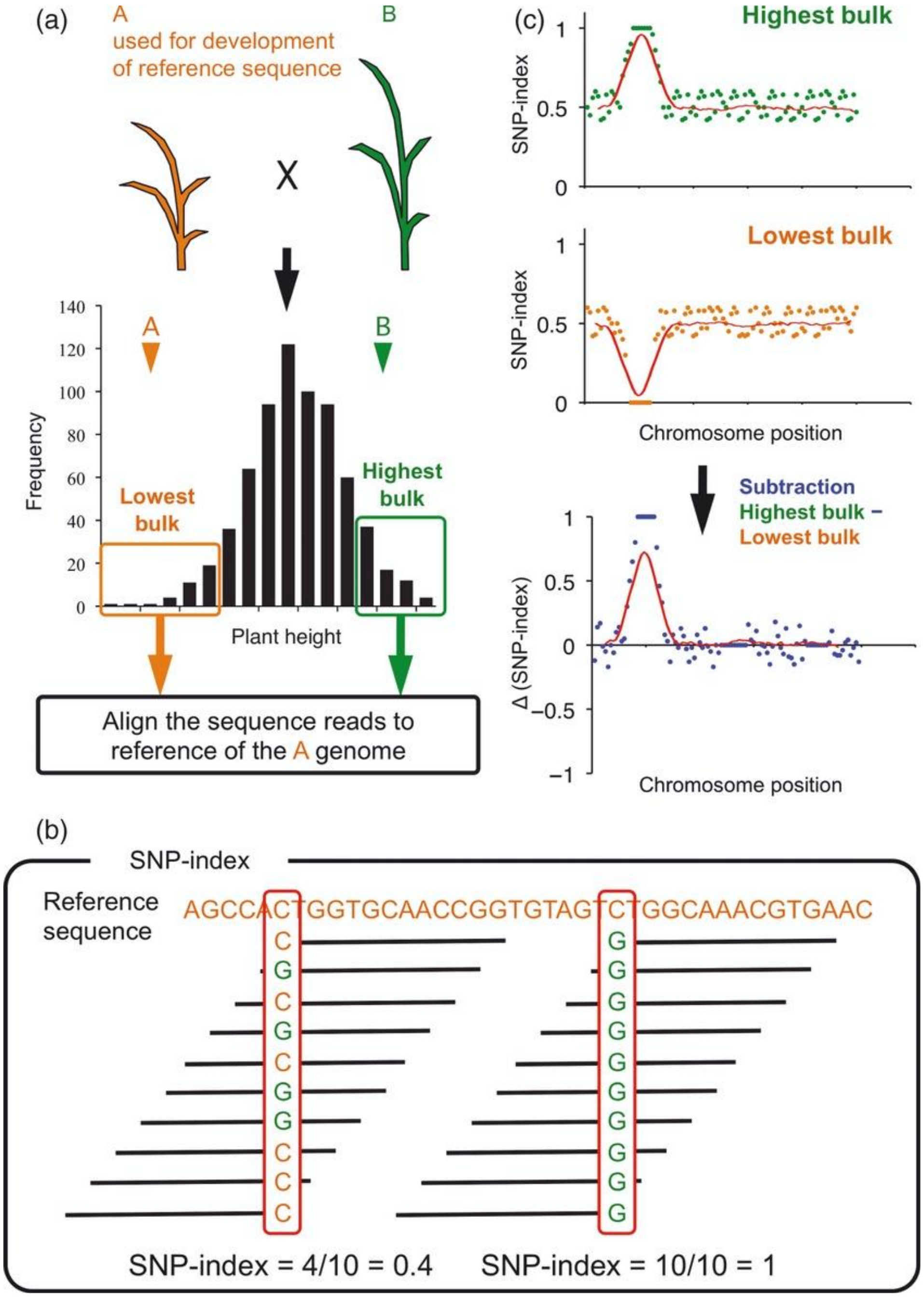

**Figure 2.4** Principle of QTL-seq strategy and estimation of SNP-index (Used with permission from Takagi et al. 2013)

**Table 2.4** List of traits mapped by adopting the QTL-seq method

| **Crop** | **Phenotype** | **Number of identified QTLs** | **Reference** |
|---|---|---|---|
| Rice | Resistance to blast | 1 | Takagi et al. 2013 |
| | Seedling vigor | 1 | |
| | Seedling vigor under low temperature | 3 | |
| | Cold tolerance | 6 | Yang et al. 2013b |
| Cucumber | Early flowering | 1 | Lu et al. 2014 |
| Tomato | Fruit weight | 3 | Illa-Berenguer et al. 2015 |
| | Locule number | 3 | |
| Chickpea | 100 seed weight | 1 | Das et al. 2015 |
| Sorghum | Stem moisture | 1 | Han et al. 2015 |
| Pigeon pea | Fusarium wilt resistance | 4 | Singh et al. 2015 |
| | Sterility mosaic resistance | 3 | |
| Chickpea | Pod number | 2 | Das et al. 2016 |
| Cucumber | Fruit length | 8 | Wei et al. 2016 |
| Rice | Salinity tolerance | 21 | Tiwari et al. 2016 |
| Chickpea | Plant height | 6 | Kujur et al. 2016 |
| Foxtail millet | Panicle branching | 1 | Masumoto et al. 2016 |
| Cucumber | Subgynoecy | 4 | Bu et al. 2016 |
| *Brassica napus* | Branch angle | 1 | Wang et al. 2016 |
| Cucumber | Downy mildew resistance | 5 | Win et al. 2016 |
| Chickpea | 100 seed weight | 2 | Singh et al. 2016 |
| | Root/total plant dry weight | 1 | |
| Soybean | Phytophthora root rot resistance | 1 | Zhong et al. 2018 |
| Broccoli x Cabbage | Flowering time | 1 | Shu et al. 2018 |
| Rice | Dwarfness | 1 | Kadambari et al.2018 |

# Chapter 3 Phenotypic Characterization of RIL Mapping Population Derived from a Cross of ‘Royal’ Flax and ‘RE2’ an Early Flowering Derivative Line

## 3.1 Abstract

‘Royal’, a heritage flax cultivar, and three early flowering lines ‘RE1’, ‘RE2’ and ‘RE3’ derived from treatment with an epimutagen 5-Azacytidine (5-AzaC) were previously found to exhibit an early flowering phenotype observed to be heritable over nine generations. The line ‘RE2’, was found to be least photoperiod sensitive. A Recombinant Inbred Line (RIL) population was developed by crossing ‘RE2’ with ‘Royal’. This mapping population was grown at the Kernen Crop Research Farm, University of Saskatchewan, over three years - 2015, 2016 and 2017 using the modified augmented design type 2 (MAD2) and evaluated for a range of phenotypic traits including, days to- start of flowering, full flowering, maturity; and height. The observed phenotypic values were adjusted using a MAD2 statistical pipeline. The adjusted phenotypic values were used to estimate statistical and genetic parameters, such as, coefficient of variation (CV), genotypic coefficient of variation (GCV) and heritability. The average days to start of flowering in the 2016 crop season was 38- and 33 days for ‘Royal’ and ‘RE2’, respectively. However, the days to anthesis for the RILs ranged from 30 days to 52 days with a mean of 37 days. The flowering time trait was moderately heritable with a broad sense heritability value of 0.49. The individuals of the RIL population were ranked based on days to flowering. The *high* and *low* bulks were constituted from the RILs exhibiting early- and late flowering time. The early- and late flowering bulks consisted of 13 and 11 individuals, respectively from the distributional extremes of the segregating population, representative of a single meiotic event, which was used to identify the potential genomic region underlying early flowering trait employing QTL-seq, a novel mapping by sequencing methodology. QTL-seq is a modified Bulk Segregant Analysis (BSA) utilising next generation sequencing reads. The distinctness of the bulks for days to flowering were confirmed using an independent field experiment with lines constituting the bulks and checks, in a Randomized Complete Block Design (RCBD) in 2017.

### 3.2 Introduction

The expansion of flax production to the northern regions of the Canadian Prairies is currently not feasible due to the short growing season and early fall frost. The crop currently grown at northern latitudes is highly susceptible to frost damage before reaching physiological maturity. Hence, the development of early flowering and consequently early maturing cultivars suitable for the northern part of the grain belt is a prime objective of flax improvement. Early flowering and maturing cultivars are also suitable for current flax growing region, as a more determinant growth habit where there is no reflowering after late summer rainfall, and capsule and stem browning occur simultaneously will reduce the tangling of green stems in harvest equipment (i.e. address straw management issues in flax production).

In crop plants, most of the agronomically important traits are controlled by multiple genes, each with a minor effect on the phenotype (Holland 2007; Mackay 2009). Flowering time is one such quantitative trait. By conventional linkage mapping and quantitative trait locus (QTL) analysis, multiple quantitative loci governing flowering time have been mapped in many crops such as brassica (Liu et al. 2016), chickpea (Daba et al. 2016) and pearl millet (Kumar et al. 2017b). However, genetic control of flowering time in crop plants, especially candidate genes underlying flowering is not well understood from these conventional studies.

In recent times, next generation sequencing based mapping strategies such as MutMap (Abe et al. 2012) and QTL-seq (Takagi et al. 2013) have been developed. In QTL-seq, initially, the method involves evaluating a biparental mapping population, for the trait of interest in different environments. The individuals exhibiting extreme phenotypes, consistently in different environments, are chosen to constitute the bulks utilized for DNA sequencing followed by downstream analysis.

Varied kinds of genetic mapping populations including $F_2$, recombinant inbred lines (RIL), doubled haploid (DH), nested association mapping (NAM) population and multiparent advanced generation inter-cross (MAGIC) population, each with its own advantages (Bazakos et al. 2017) are available. Among the biparental mapping populations, although developing $F_2$ population is less time consuming, the heterozygous nature and restricted seed availability are its limitations. Recombinant inbred populations are widely used because of their homozygosity, higher number of recombination events and immortality, despite the long generation time (Keurentjes et al.

2007). Doubled haploids are a rapid way of attaining homozygosity. However, there is possibility for only one recombination event and hence, the mapping resolution is limited (Zhang et al. 2017). The NAM population is developed by crossing a single reference line to multiple inbreds and finally combining families from several resulting biparental populations. Since the reference line is common, and it involves multiple parents, NAM combines the advantages of linkage mapping and association mapping (Cockram and Mackay 2018). The MAGIC population is derived by crossing a set of founders in a series of two-way, four-way and eight-way crosses. This not only serves as the mapping population for linkage studies but also serves as the reservoir of prebreeding lines representing diverse genetic background of founder lines and hence, suitable for selection (Huang et al. 2015). Both NAM and MAGIC populations provide high mapping resolution. The major disadvantage is the huge amount of resources needed for the generation of these multiparent lines. Since multiple parents are involved, identifying parental origin of alleles in the segregating population and mapping them requires special statistical methods.

In flax, a RIL mapping population was developed by crossing the early flowering line ‘RE2’ with ‘Royal’. Using this mapping population and the reference genome assembly generated as a part of the Total Utilization Flax GENomics (TUFGEN) project as resources, QTL-seq approach was deployed to map potential genomic region(s) responsible for early flowering phenotype. Identification of flowering time loci will help to unearth the candidate gene(s), which will help to deduce the molecular mechanisms involved in flowering, and towards development of diagnostic molecular markers that would assist in marker assisted breeding for this trait.

This chapter discusses the comprehensive phenotypic characterization of the ‘Royal’ x ‘RE2’ mapping population using MAD2, to identify individuals that constitute the early- and late flowering bulks for flowering time.

## 3.3 Materials and methods

### 3.3.1 Plant material

Three early flowering lines, named ‘RE1’, ‘RE2’ and ‘RE3’ were derived from the cultivar ‘Royal’ (Figure 3.1) upon treatment with the DNA methylation inhibitor 5-AzaC (Fieldes et al. 1994). The epimutant lines flowered 7 to 13 days earlier than ‘Royal’ (Fieldes and Harvey 2004).

Among the three accessions, ‘RE2’ was observed to be earliest flowering and least photoperiod sensitive (Sun 2015, M. Sc., Thesis, University of Saskatchewan). The early flowering phenotype was found to be stably inherited through meiosis over nine generations. A RIL population was developed by crossing ‘RE2’ and its progenitor genotype ‘Royal’, followed by advancing the $F_2$ population by single seed descent method for nine generations at the Crop Development Centre (CDC), University of Saskatchewan. The final mapping population consisted of 656 lines of which 288 lines were from the cross made using ‘RE2’ as the male parent, and the remaining 368 lines were derived from its reciprocal, where ‘Royal’ donated the pollen gamete.

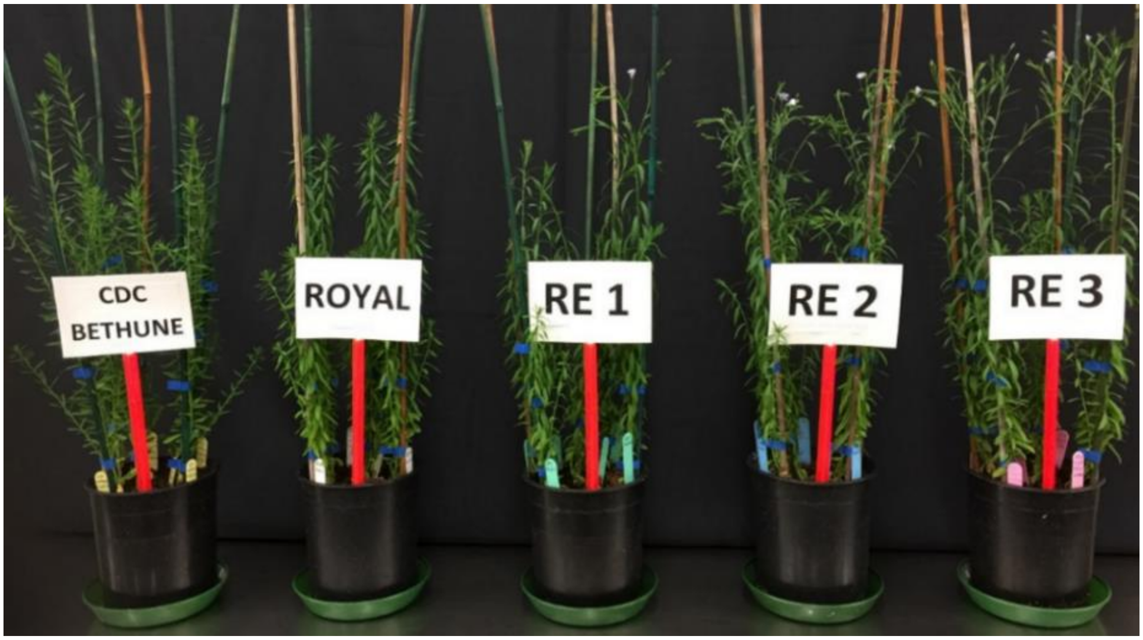

**Figure 3.1** Cultivar ‘Royal’ and its three early flowering derivatives ‘RE1’, ‘RE2’ and ‘RE3’ and the widely grown flax cultivar ‘CDC Bethune' (Cabinet grown, under long day- 16 hours light, eight hours dark conditions; 31 days after seeding)

### 3.3.2 Field trial

The field experiments to evaluate the RILs for selected agronomically important traits at different phenological stages of the crop were carried out at the Kernen Crop Research Farm, Saskatoon (52º 09’ 02.2” N and 106º 32’ 36.7” W; Elevation: 511m; Soil type: Silty clay) in 2015, 2016 and 2017. RILs were evaluated in the field using the MAD2 (Figure 3.2; Lin and Poushinsky 1985) to correct for potential soil heterogeneity. The 7 x 7 lattice design consisted of seven rows and seven columns with a total of 49 whole plots. Further, each main plot was subdivided into 15 subplots along the rows. The cultivar ‘CDC Bethune’ was used as the main plot control and was seeded in the middle subplot of each whole plot. A second cultivar ‘CDC Sorrel’ was used as the subplot control and was planted in two subplots of five randomly chosen whole plots. The 656 test entries and parents (‘Royal’ – 8 subplots; ‘RE2’ – 8 subplots) were randomized among the remaining subplots. The experiments were seeded (50 seeds/hill) on the 12th May and on the 17th May in 2015 and 2016, respectively. In 2017, the trial was initially seeded on the 19th May, but was extensively damaged by cutworms (*Agrotis orthogonia*), and hence, reseeded on the 12th June.

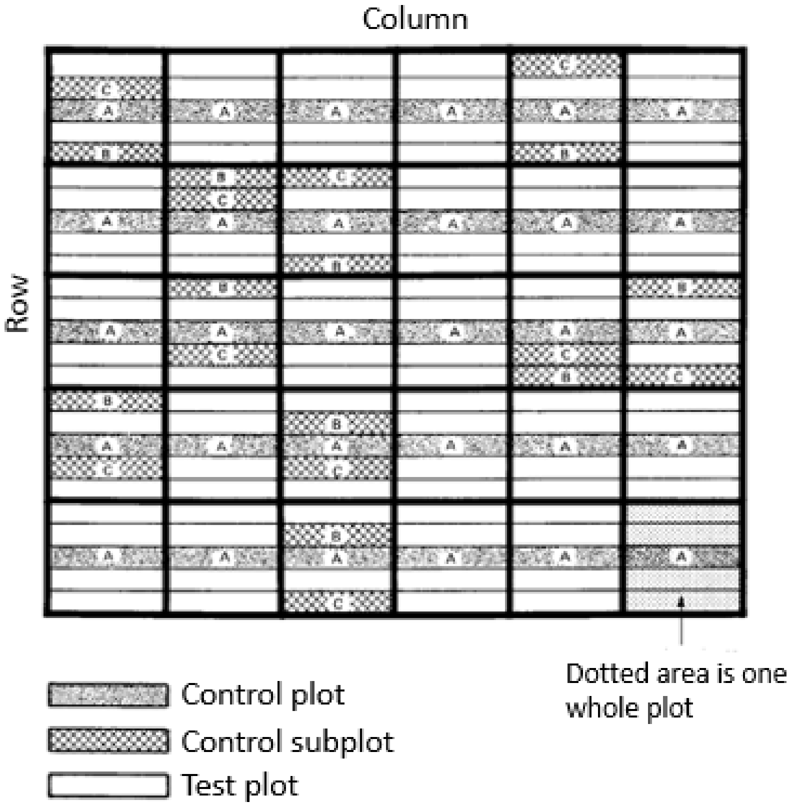

**Figure 3.2** Model layout of modified augmented design type 2; the plot control is seeded in middle subplot of all whole plots and sub plot controls are seeded in two subplots of five randomly selected whole plots. Figure adapted and used with permission from Lin and Poushinsky 1985.

### 3.3.3 Phenotypic characterization in the field

Standard flax descriptors were used to phenotype the different traits. The date of emergence of individual hills were marked using coloured tags representing specific days. Emergence was marked when the seeds in the hill germinated, and the seedlings were visible above the ground. The following traits were recorded: early vigor, start of flowering, full flowering, maturity date, maturity score and height. Specifically, early vigor rating was taken when ‘CDC Bethune’ was at a height of 10-15 cm. It was scored on a scale of one to nine where one indicated extremely weak vigor, five indicated average vigor represented by ‘CDC Bethune’ and a score of nine represented extremely vigorous plant growth. Start of flowering was recorded as the number of Julian days when 5% of the plants in the hill had reached anthesis and full flowering was recorded when 95% of plants in the hill had reached anthesis. Maturity date was noted as the Julian day when the seeds rattle in the capsule (75% of the plants) upon shaking. Days to- start of flowering, full flowering and maturity from the date of emergence were estimated for plants grown in each individual hill. Maturity score was recorded on a single day using an ordinal scale of one to five, representing immature to mature plants, respectively. The height (in centimeters), measure from ground surface to upper most portion of the plant was recorded at the completion of flowering phase. The phenotypic evaluation of the RILs in 2015 field season was carried out by Dr. Raja Ragupathy, a former senior colleague in the flax team.

### 3.3.4 Phenotypic characterization in the growth cabinet

The 288 RILs from the ‘Royal’ x ‘RE2’ cross, along with the parents, were grown and evaluated in the growth cabinet in 2015-2016 . Each RIL was replicated four times and randomized. Four randomly allocated single plants were grown in each pot. Each pot was filled with nearly 3 litres of Sungro propagation mix (Sungro Horticulture, Massachusetts, USA) as the growing media. The plants were grown under long day conditions with 16 hours of light and eight hours of night and the day and night temperatures were 22 $^{o}$C and 17 $^{o}$C, respectively. Phenological traits such as days to- start of flowering, full flowering, maturity; height and maturity score were recorded using the definitions described in section 3.3.3. This data was kindly provided by Dr. Raja Ragupathy.

### 3.3.5 Data analysis

Phenotypic data was analysed using the statistical pipeline involving PERL and SAS scripts developed specifically for the MAD2 (You et al. 2013). In brief, the SAS scripts of the pipeline perform two discrete ANOVA for plot and subplot controls. Based on the ANOVA results, the soil heterogeneity, both additive (along the rows and columns of the plot) and non-additive (multi directional), is corrected by adjustment of phenotypic values by three methods: (a) When the observed row or column effects were found to be higher than the row x column interactions, the plot control means along the rows and columns are used for adjustment (Method 1); (b) When the value of the main plot control was significantly greater than the value of subplot control, adjustment was based on the regression of the test plots on the plot control (Method 3); (c) In cases where both row (or) column effects and their interaction effects were significant, a method of adjustment combining methods 1 and 3 was used (Method 1+3). Finally, the phenotypic data were adjusted accounting for soil heterogeneity, and the relative efficiencies of different methods of adjustment were estimated using a PERL script. As the final step, the values from the most appropriate adjustment method were exported (You et al. 2013).

### 3.3.6 Identification of phenotypic extremes of the segregating population

The field phenotypic data for the years 2015, 2016 were analysed using the MAD2 pipeline (You et al. 2013). Based on the values for days to start of flowering in the years 2015, 2016, and the greenhouse data from 2015-2016, the RILs were sorted by giving the earliest flowering lines the highest rank. The individuals with extreme phenotypic values for flowering time from the cross 'Royal' x 'RE2' were only used. Among the 288 RIL individuals, 26 early flowering and 27 late flowering lines with consistent performance in different environments were identified. Finally, a subset of 13 early flowering and 11 late flowering lines, derived from a single meiotic event, were used for DNA sequencing.

### 3.3.7 Field trial for validation of phenotypic extreme lines

In addition to the main field trial, in 2017, the 26 early flowering and 27 late flowering lines identified from the 288 RILs derived from the 'Royal' x 'RE2' cross were planted in a randomized complete block design with three replicates. The trial was seeded on the 19th May (50 seeds/hill) at the Kernen Crop Research Farm, University of Saskatchewan. The three early flowering accessions ‘RE1’, ‘RE2’, ‘RE3’; the cultivars ‘Royal’, ‘CDC Bethune’, ‘CDC Plava’ and ‘Prairie Thunder’ were all used as check cultivars in the trial. The range of phenotypic traits including, start of flowering, full flowering, days to maturity, maturity score and height were recorded following the scoring procedure described above in section 3.3.3.

### 3.3.8 Estimation of genetic parameters

The analysis of phenotypic data was carried out using the statistical model

$$y_{ij} = \mu_{ij} + G_i + Y_j + E_{ij}$$

where,
$\mu$ = population mean,

$G_i$ = genotypic variance

$Y_j$ = year variance

$E_{ij}$ = error variance.

The broad sense heritability of the traits was estimated as the ratio of genetic variance to the total phenotypic variance observed in the population, determined using the equation

$$H^2 = \frac{\sigma_g{}^2}{\left(\sigma_g{}^2 + \left(\frac{\sigma_e{}^2}{n_y}\right)\right)}$$

where,

$\sigma_g{}^2$ = genetic variance

$\sigma_e{}^2$ = error variance

$n_y$ = number of years in which the trait was evaluated.

The coefficient of variation (CV) and genotypic coefficient of variation (GCV), expressed in percentage, were calculated as follows:

$$CV = \frac{\sigma_p}{x} \text{ x } 100$$

$$\text{GCV} = \frac{\sigma_g}{x} \text{ x } 100$$

where,

$\sigma_p$ = phenotypic standard deviation,

$\sigma_g$ = genetic standard deviation.

x = mean

### 3.3.9 Analysis of RCBD trial data

The data from the RCBD trial comprising of 53 individuals exhibiting extreme phenotypic values, with three replications was analysed. The PROC MIXED procedure of SAS version 9.3 (Copyright © 2011, SAS Institute Inc., Cary, NC, USA) was used. Significant differences among the mean values were declared at $P < 0.05$. The model used for analysis was

$$y_{ij} = \mu_{ij} + B_i + T_j + E_{ij}$$

where,

$y_{ij}$ = value of the dependent variable (the trait of interest),

$B_i$ = random effect of the blocks,

$T_j$ = fixed effect (early- and late flowering lines)

$E_{ij}$ = error.

## 3.4 Results

### 3.4.1 Growing conditions

As anticipated, there was considerable variation in local environmental conditions during 2015-2017 growing season. To highlight these differences the mean environments over 10 years was used as a base-line for comparison. The environmental conditions observed over a period of 10 years (2008-2017) at the Kernen Crop Research Farm are depicted in Figure 3.3. In the year 2015, there was minimal rain during the initial phase of the growing season with only 25.5% and 32.3% of the 10-year average rainfall received, for May and June, respectively. In contrast, in

2016, the month of May received 21.8% more rainfall than the 10-year average and June received 72.8% of the average rainfall. However, in 2017, the growing season with delayed seeding, the initial growing phase in the months of June and July received only 68.38% and 49.29% of the 10-year average rainfall. In addition, the seeding date in the 2017 field season was deferred by nearly one month in comparison to other two years. Hence, there was a significant difference in the photoperiod to which the plants were exposed after emergence (Figure 3.4).

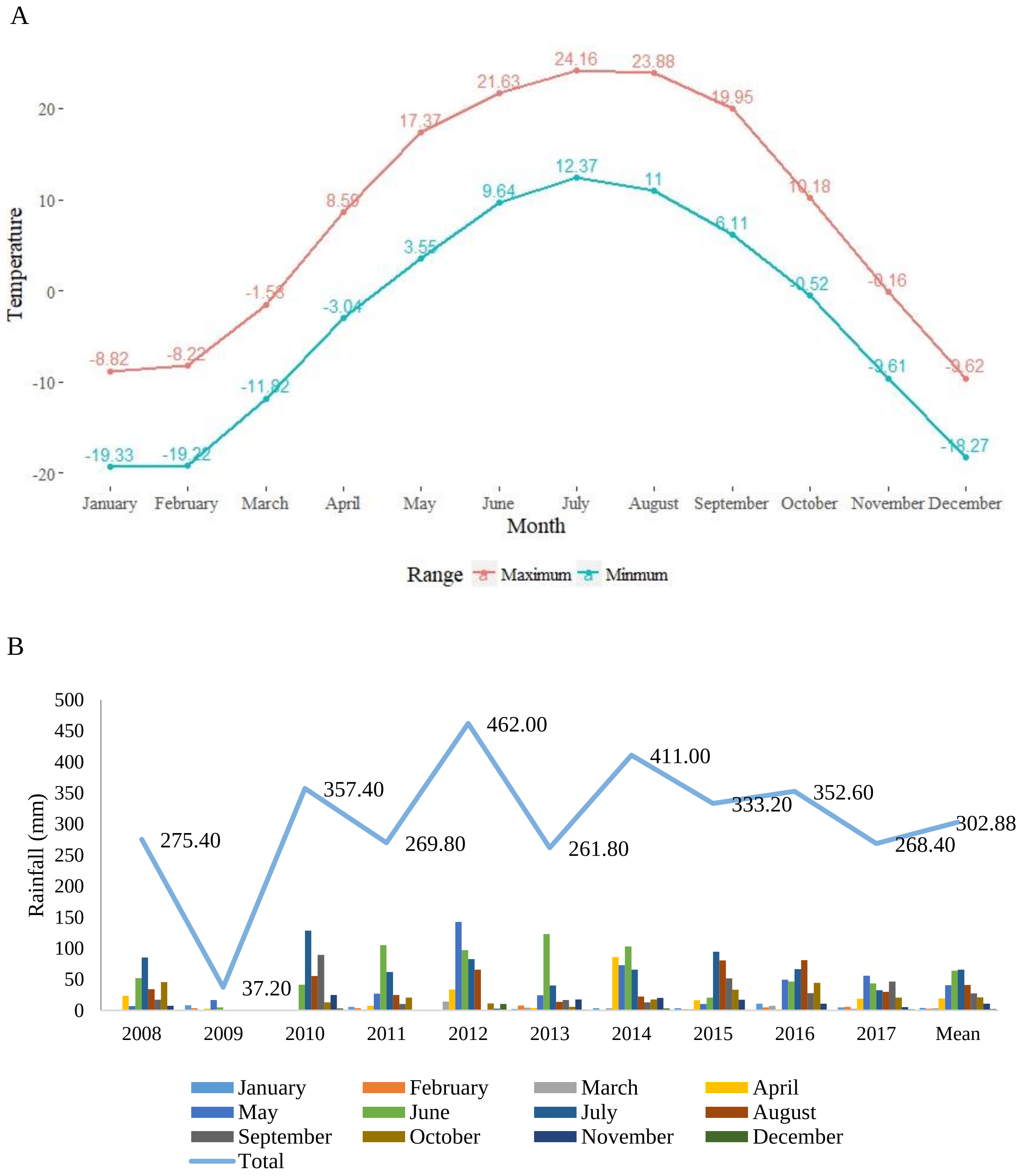

**Figure 3.3** Weather data at the Kernen Crop Research Farm, University of Saskatchewan, Saskatoon, Saskatchewan: (A) 10-year average (2008-2017) monthly temperature (B) 10-year (2008-2017) total annual and monthly distribution rainfall

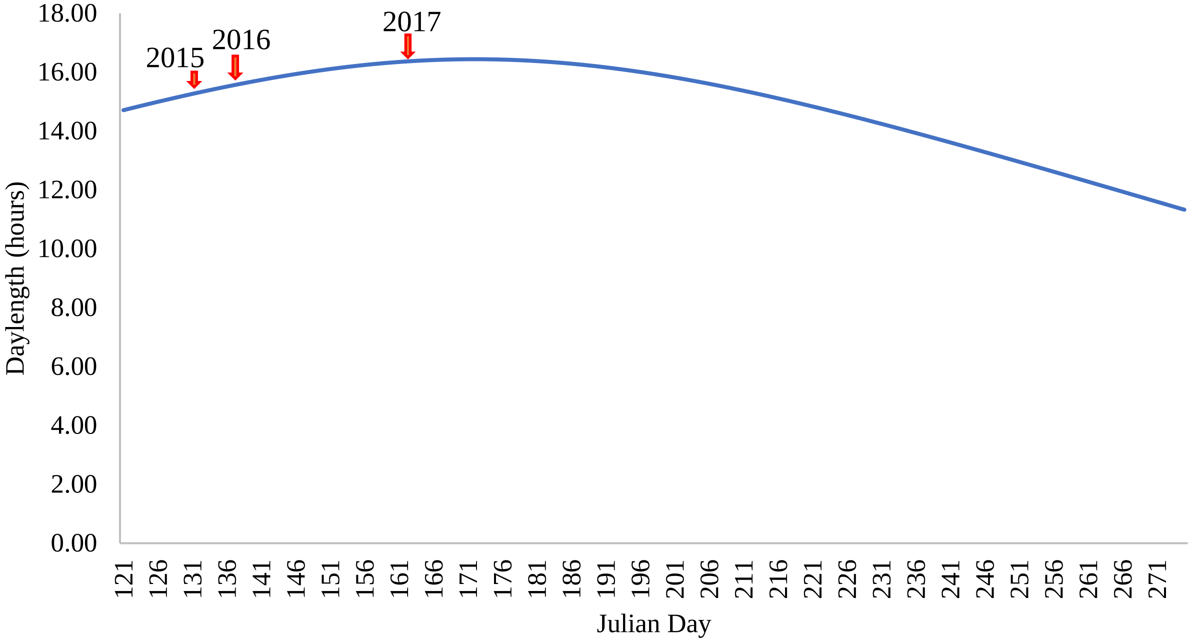

**Figure 3.4** The distribution of daylength (in hours) during the crop season (May to September) at Kernen Crop Research Farm, Saskatoon, Saskatchewan. Daylength was estimated using a formula described in Kirk 1994. The red arrows depict the seeding date in the three field seasons.

### 3.4.2 Genetic parameter estimates

The RIL population was planted in the field and scored for selected range of agronomic traits, and the data was summarized using the summary statistics. The RIL population was found to be segregating for flowering time in all three years (Figure 3.5). The mean, standard deviation, range, CV, GCV and heritability (Table 3.1) were estimated. The average days to flowering for the RIL population was observed to be 50.6 days in 2015, 37.0 days in 2016 and 35.4 days in 2017 (Table 3.2). The plot control ('CDC Bethune') and sub plot control ('CDC Sorrel') took the longest average-duration for the start of flowering in 2015, with corresponding values 54.2 days and 52.9 days, respectively. The least number of days to reach the start of flowering was observed in one of the RILs (EF RIL-280-23) in 2017, taking 25 days from emergence. This contrasted to maximum number of days to start of flowering recorded in 2015 field season as 64 days from emergence (EF RIL-271-34, EF RIL-288-2). Days to flowering had a relatively high CV of 17.67%. The CV was the lowest for days to maturity (3.07%) and highest for maturity score (27.12%). The GCV was relatively low for all traits, with maturity score having the lowest

GCV of 0.004%. The estimated CV and GCV values suggest the extent of phenotypic- and genetic variabitlity, respectively, in the population for the individual traits. Among the traits under study, days to start of flowering, days to full flowering and height exhibit moderate heritability values of 0.49, 0.42 and 0.43, respectively, while other traits present very low broad sense heritability values. The parents had no variability for the traits including early vigor, days to maturity and maturity score. Hence, these traits did not segregate in the RIL population, further resulting in the very low broad sense heritability values.

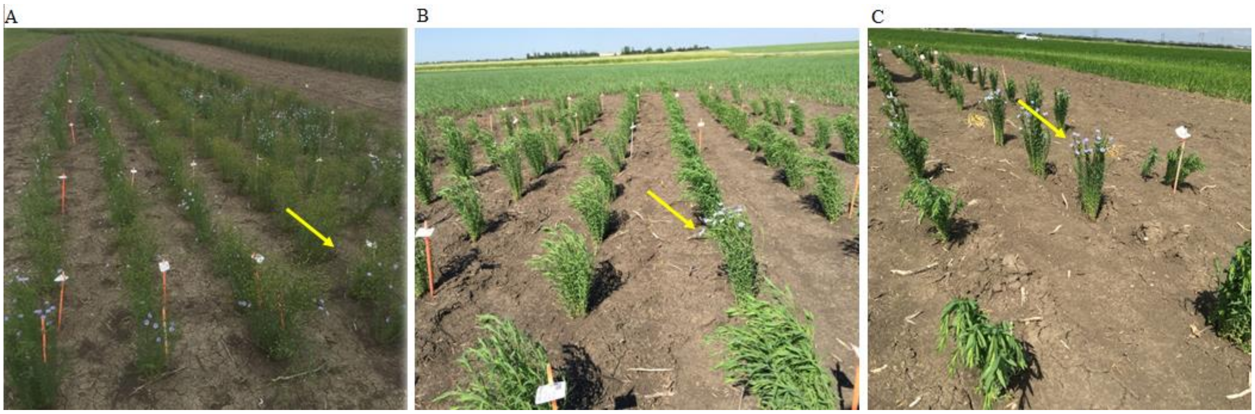

**Figure 3.5** The RIL population segregating for flowering time in three field seasons (A) 2015, (B) 2016, (C) 2017, at the Kernen Crop Research Farm, University of Saskatchewan, Saskatoon, Saskatchewan. (Used with permission from Dr. Raja Ragupathy).

**Table 3.1** Genetic parameters and phenotypic estimates for different agronomic traits in the 'Royal' x 'RE2' RIL population over the years 2015-2017

| **Trait** | **Mean ± SD** | **Min** | **Max** | **CV (%)** | **GCV (%)** | **$H^2$** |
|---|---|---|---|---|---|---|
| Early vigor | 5.90 ± 1.44 | 1.00 | 9.00 | 24.41 | 0.01 | $7.73E^{-07}$ |
| Days to start of flowering | 41.04 ± 7.25 | 25.00 | 64.00 | 17.67 | 2.85 | 0.49 |
| Full flowering | 48.21 ± 7.85 | 32.00 | 73.00 | 16.28 | 2.43 | 0.42 |
| Height (cm) | 55.22 ± 6.41 | 19.57 | 78.90 | 11.61 | 4.71 | 0.43 |
| Maturity score | 3.65 ± 0.99 | 1.13 | 5.47 | 27.12 | 0.004 | $1.36E^{-07}$ |
| Days to maturity | 98.05 ± 3.01 | 76.41 | 110.00 | 3.07 | 0.23 | 0.02 |

*Mean - population mean; SD - standard deviation; Min - Minimum; Max - Maximum; CV - coefficient of variation; GCV - genetic coefficient of variation; $H^2$ - broad sense heritability.*

**Table 3.2** Mean values of parental lines checks cultivars and RIL population in 2015, 2016 and 2017

| | | | 2015 | | |
|---|---|---|---|---|---|
| **Traits** | **CDC Bethune**$^{\$}$ | **CDC Sorrel**$^{\$}$ | **Royal**$^{@}$ | **RE2**$^{@}$ | **Royal x RE2 RILs*** |
| Early Vig | 7.00 | 7.55 | 7.30 | 5.85 | 6.80 (4.00 - 9.28) |
| Startflwr | 54.22 | 52.86 | 50.50 | 47.25 | 50.62 (47.00 - 64.00) |
| Fullflwr | 62.63 | 60.93 | 58.75 | 55.75 | 58.51 (52.00-73.00) |
| Height | 58.22 | 63.21 | 53.50 | 38.50 | 51.34 (28.00 - 63.00) |
| Mat Days | 102.10 | 104.57 | 100.75 | 96.75 | 98.78 (92.00 - 110.00) |
| | | | **2016** | | |
| **Traits** | **CDC Bethune**$^{\$}$ | **CDC Sorrel**$^{\$}$ | **Royal**$^{@}$ | **RE2**$^{@}$ | **Royal x RE2 RILs*** |
| Early Vig | 4.65 | 4.93 | 4.57 | 6.38 | 5.22 (1.00 - 9.00) |
| Startflwr | 42.32 | 42.64 | 37.65 | 33.35 | 37.00 (29.62 - 52.08) |
| Fullflwr | 49.65 | 50.68 | 44.91 | 41.01 | 44.22 (37.06 - 59.32) |
| Height | 69.47 | 70.46 | 55.51 | 54.65 | 57.22 (19.57 - 78.90) |
| Mat Days | 103.42 | 104.07 | 98.43 | 93.38 | 97.93 (90.00 - 108.00) |
| | | | **2017** | | |
| **Traits** | **CDC Bethune**$^{\$}$ | **CDC Sorrel**$^{\$}$ | **Royal**$^{@}$ | **RE2**$^{@}$ | **Royal x RE2 RILs*** |
| Early Vig | 5.00 | 5.14 | 5.00 | 5.00 | 5.66 (3.00 - 9.00) |
| Startflwr | 39.40 | 39.64 | 36.38 | 32.63 | 35.39 (25.00 - 46.00) |
| Fullflwr | 45.17 | 44.71 | 42.75 | 39.25 | 41.81 (32.00 - 52.00) |
| Height | 64.33 | 71.45 | 58.05 | 48.03 | 57.15 (37.52 - 68.19) |
| Mat Days | 95.13 | 99.58 | 98.60 | 98.71 | 97.38 (76.14 - 106.84) |

$^{@}$*Parents (‘Royal’ and ‘RE2’);*
$^{\$}$*Check cultivars (CDC Bethune – main plot control; CDC Sorrel – subplot control);*
**Mean (min – max) values of segregating RIL populations;*
*Early vig-early vigor; Startflwr-days to start of flowering; Fullflwr-days to full flowering; Mat sco-maturity score; Mat days-days to maturity.*

### 3.4.3 Correlation between phenological traits

The correlation among the phenological traits in the field data from three years and 2015-2016 growth cabinet data are presented in the Tables 3.3, 3.4, 3.5 and 3.6. The days to start of flowering is highly correlated with days to full flowering in all the three seasons - 2015 (0.87, P < 0.01), 2016 (0.89, P < 0.01), 2017 (0.84, P < 0.01) in the field; and the relationship was strongest in the growth cabinet (0.96, P < 0.05). Height was strongly correlated with days to start of flowering (0.83, P < 0.05) and full flowering (0.80, P < 0.05) in the 2015-2016 growth cabinet data, and a moderate correlation was observed only in the 2015 field data (0.33, P < 0.05). In the year 2016, days to maturity exhibited moderate correlation with days to start of flowering (0.57, P < 0.05) and days to full flowering (0.52, P < 0.05). In 2016 field season, the correlation between maturity score and maturity date was moderate and negative (-0.55, P < 0.05), whereas in 2017, it was weak and negative (-0.31, P < 0.05).

**Table 3.3** Pearson correlation coefficient for phenological traits in the RIL population-2015 field data

| | Startflwr | Fullflwr | Height | Mat sco | Mat days |
|---|---|---|---|---|---|
| Early vig | 0.20** | 0.21** | 0.50** | -0.17** | NS |
| Startflwr | | 0.87** | 0.33** | -0.43** | 0.19** |
| Fullflwr | | | 0.34** | -0.38** | 0.14** |
| Height | | | | -0.39** | NS |
| Mat sco | | | | | NS |

** Indicates significance at 5% level   ** Indicates significance at 1% level; NS- Not significant. Early vig-early vigor; Startflwr-days to start of flowering; Fullflwr-days to full flowering;*
*Mat sco-maturity score; Mat days-days to maturity.*

**Table 3.4** Pearson correlation coefficient for phenological traits in the RIL population-2016 field data

| | Startflwr | Fullflwr | Height | Mat sco | Mat days |
|---|---|---|---|---|---|
| Early vig | -0.18** | -0.18** | -0.22** | 0.20** | -0.26** |
| Startflwr | | 0.89** | 0.22** | -0.22** | 0.57** |
| Fullflwr | | | 0.31** | -0.15** | 0.52** |
| Height | | | | NS | 0.16** |
| Mat_sco | | | | | -0.55** |

** Indicates significance at 5% level  ** Indicates significance at 1% level; NS- Not significant. Early vig-early vigor; Startflwr-days to start of flowering; Fullflwr-days to full flowering;*
*Mat sco-maturity score; Mat days-days to maturity.*

**Table 3.5** Pearson correlation coefficient for phenological traits in the RIL population-2017 field data

| | Startflwr | Fullflwr | Height | Mat sco | Mat days |
|---|---|---|---|---|---|
| Early vig | -0.10* | -0.16** | NS | NS | 0.13** |
| Startflwr | | 0.84** | NS | 0.22** | 0.12** |
| Fullflwr | | | NS | 0.25** | 0.09* |
| Height | | | | NS | NS |
| Mat sco | | | | | -0.31** |

** Indicates significance at 5% level  ** Indicates significance at 1% level; NS- Not significant. Early vig-early vigor; Startflwr-days to start of flowering; Fullflwr-days to full flowering;*
*Mat sco-maturity score; Mat days-days to maturity.*

**Table 3.6** Pearson correlation coefficient for phenological traits in the RIL population-2015 growth cabinet data

| | Fullflwr | Height | Mat sco | Mat days |
|---|---|---|---|---|
| Startflwr | 0.96** | 0.83** | 0.11* | -0.29** |
| Fullflwr | | 0.80** | 0.08* | -0.27** |
| Height | | | 0.20** | -0.39** |
| Mat sco | | | | -0.19** |

** Indicates significance at 5% level  ** Indicates significance at 1% level. Startflwr-days to start of flowering; Fullflwr-days to full flowering; Mat sco-maturity score; Mat days-maturity days.*

### 3.4.4 Identification of individuals with extreme phenotypes

The 'RE2' x 'Royal' RIL population was primarily designed to investigate flowering time and assess its heritability after crossing. The distribution of the phenotypic values for days to start of flowering from 2016 field season is displayed in Figure 3.5 where, the mean of the population was 37 days and the mode was 38 days. On ranking the phenotypic values for start of flowering as described earlier, individuals showing extreme values were selected (Figure 3.6). A total of 26 early flowering and 27 late flowering lines from each tail of the distributional extreme were identified among the 288 RIL population. The subset of 13 early flowering and 11 late flowering lines (Table 3.7) when grown in the growth cabinet in 2017, were found to exhibit their corresponding phenotypes

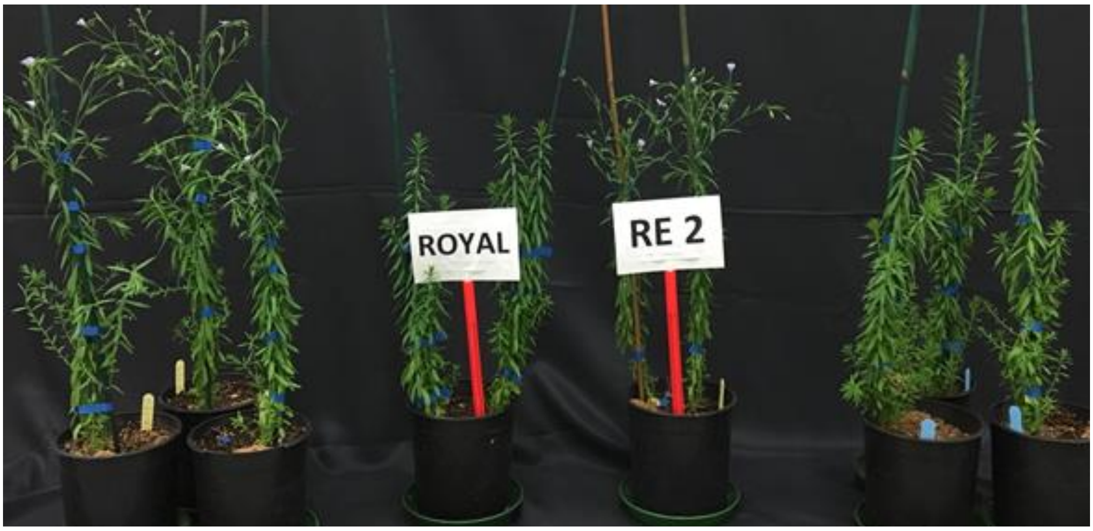

**Figure 3.6** The early- and late flowering bulks grown along with parents 'Royal' and 'RE2' in the growth cabinet under long day conditions with 16 hours of light.

**Table 3.7** List of early- and late flowering lines chosen to constitute the bulks

| **Lines constituting early flowering bulk** | **Days to flowering (2016 field season)** | **Lines constituting late flowering bulk** | **Days to flowering (2016 field season)** |
|---|---|---|---|
| EF RIL-279-28 | 31 | EF RIL-279-8 | 40 |
| EF RIL-279-14 | 32 | EF RIL-280-22 | 40 |
| EF RIL-279-2 | 33 | EF RIL-280-29 | 40 |
| EF RIL-280-26 | 33 | EF RIL-280-5 | 43 |
| EF RIL-279-1 | 33 | EF RIL-281-6 | 40 |
| EF RIL-279-16 | 33 | EF RIL-282-14 | 40 |
| EF RIL-279-7 | 31 | EF RIL-281-9 | 40 |
| EF RIL-281-28 | 32 | EF RIL-281-17 | 40 |
| EF RIL-282-20 | 32 | EF RIL-281-15 | 40 |
| EF RIL-281-25 | 32 | EF RIL-281-12 | 40 |
| EF RIL-281-30 | 33 | EF RIL-281-27 | 41 |
| EF RIL-282-10 | 33 | | |
| EF RIL-282-29 | 31 | | |

*Population mean=37±2.34*

### 3.4.5 Replicated field test of lines chosen for early- and late flowering bulks

From the analysis results of the RCBD trial data (Table 3.8), a significant difference was observed between the early- and late flowering lines for days to- start of flowering (P=0.0197), full flowering (P=0.0072) and maturity (P=0.0350). Also, a highly significant difference (P < 0.0001) between the extreme bulks was observed for height. The early- and late flowering bulks differ significantly for seed yield.

**Table 3.8** Analysis of phenotypic values from the RCBD trial in 2017

| | Days to start of flowering | Days to full flowering | Days to maturity | Height (cm) | Seed Yield (g) |
|---|---|---|---|---|---|
| **CDC Bethune** | 42.00 | 47.67 | 87.67 | 54.33 | 26.88 |
| **CDC Plava** | 42.50 | 48.00 | 88.00 | 55.00 | 22.74 |
| **Prairie Thunder** | 39.67 | 45.00 | 83.67 | 45.67 | 25.94 |
| **Royal** | 41.67 | 47.00 | 85.00 | 50.00 | 32.54 |
| **RE1** | 45.00 | 48.00 | 91.33 | 44.67 | 14.07 |
| **RE2** | 41.00 | 46.00 | 84.00 | 41.00 | 11.92 |
| **RE3** | 43.00 | 47.00 | 85.33 | 40.67 | 13.93 |
| **Check means** | 42.12 | 46.95 | 86.43 | 47.33 | 21.14 |
| | | | | | |
| **Early flowering bulk** | 40.95 | 45.29 | 85.16 | 44.01 | 20.39 |
| **Late flowering bulk** | 41.83 | 46.24 | 86.61 | 48.59 | 32.79 |
| ***P value*** | 0.0197 | 0.0072 | 0.0350 | < 0.0001 | < 0.0001 |

### 3.4.6 Phenotypic performance of- RILs for flowering time and bulks across environments

The flowering time range observed during the 2015 season was 17 days, with flowering starting 47 days after emergence and finishing 64 days after emergence. In 2016, flowering lasted for 22 days, commencing 30 days after emergence and went on till 52 days. Similarly, in 2017, flowering lasted for 27 days with the earliest flowering observed 25 days after emergence and the last day to start of flowering was 52 days after emergence. Maximum number of RILs started flowering on 49, 38, 36 days after emergence, in 2015, 2016 and 2017, respectively. In 2016, the average number of days to start of flowering for the parents ‘Royal’ and ‘RE2’ was 37.7 and 33.4, respectively. The main plot control ‘CDC Bethune’, and subplot control ‘CDC Sorrel’ had the values 42.3 and 42.6, respectively. The position of the chosen bulks at the distributional extreme of the segregating population in 2016, is depicted in Figure 3.7. The individuals constituting the early flowering bulk had days to start of flowering values ranging between 31 and 33 days, while those constituting the late flowering bulk had values ranging between 40 to 43 days. The distribution of phenotypic values of the early- and the late flowering bulks over different field seasons and in the growth-cabinet is given in Figure 3.8.

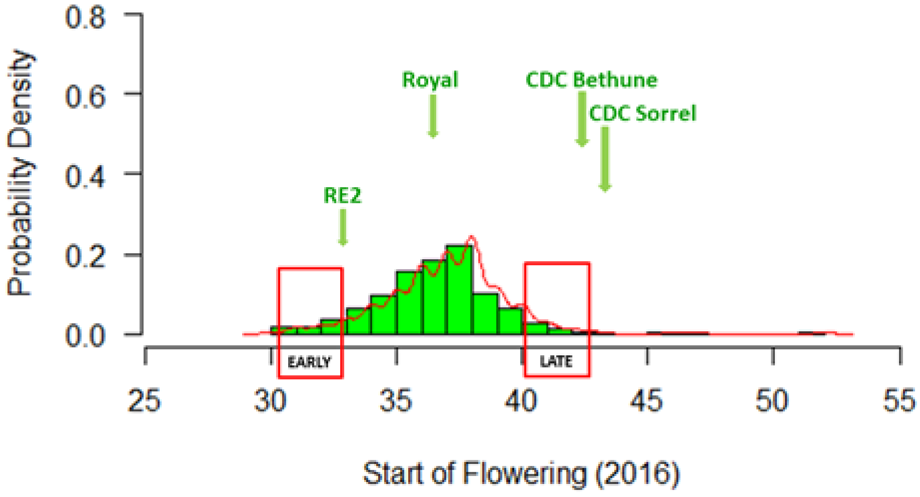

**Figure 3.7** Histogram with kernel density plot for start of flowering 2016

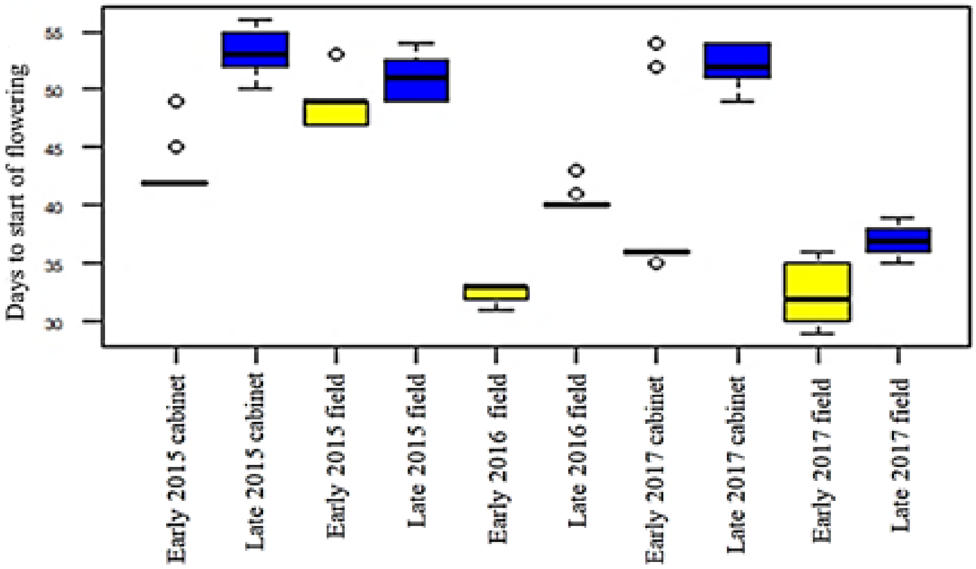

**Figure 3.8** Boxplot representing the performance of the bulks in different environments. Yellow represents early flowering bulk and blue represents late flowering bulk.

### 3.5 Discussion

A total of 29 flax cultivars have been registered in Canada since 2000. Among the cultivars, ‘2126’, a low linoleic acid containing variety is the earliest, taking ~95 days to maturity (Dribnenki et al. 2005). However, it has been suggested that a variety with a duration of 90 days to maturity is considered suitable for the northern prairie region (Duguid 2009). ‘Royal’, is an old flax cultivar with medium to late maturity (McGregor 1953). Treatment of ‘Royal’ with an epimutagenic chemical 5-Azacytidine resulted in three early flowering mutant lines – ‘RE1’, ‘RE2’ and ‘RE3’ (Fieldes 1994; Fieldes and Amyot 1999). Among the accessions, ‘RE2’ was the least photoperiod sensitive. The stable transmission of the early flowering trait through meiosis for nine generations, enabled the development of the RIL population, by crossing ‘Royal’ and ‘RE2’ (Sun 2015, M.Sc. thesis, University of Saskatchewan).

Biparental mapping population such as RILs are widely used for linkage mapping and QTL analysis besides serving as a breeding population (Morell et al. 2012). Phenotyping and genotyping of the mapping population are the basis for QTL analysis. Bulked segregant analysis is a resource efficient approach in gene tagging as it involves genotyping only of the pooled extremes, constituted based on the phenotypic data capturing the variation across the spectrum (Michelmore et al. 1991). The QTL-seq methodology is a modern version of BSA in which next generation sequencing is employed instead of markers, and its success depends on various factors such as size of mapping population, genetic architecture of the trait and accuracy of phenotyping. The precision of phenotyping of individuals with lesser quantity of seeds, and thus not amenable for replicated trials can be improved by the analysis of the phenotypic data using a modified version of augmented design such as MAD2 (Lin and Poushinsky 1985), which involves the replication of control throughout the design. Modified augmented design type 2 increases the signal to noise ratio by eliminating the influence of potential soil heterogeneity. This phenotyping methodology facilitated the selection of the most appropriate individuals for constituting the bulks, for DNA sequencing and genotyping.

The seeding dates were 12th May, 17th May and 19th May in 2015, 2016 and 2017, respectively. However, in 2017 reseeding was carried out on 12th June because of the cutworm damage. Hence, the growing environment of field trials were significantly different from each other in 2015, 2016 and 2017, because of differences in daylength, growing degree days (cumulative heat units) and moisture regimes. Specifically, in 2017, the RILs were exposed to shortening day length after planting, as a consequence of delayed seeding. The difference in seeding date has been reported to influence several agronomic traits in crop species including seed emergence, yield, pod fertility in chickpea (*Cicer arietinum* L.; Auld et al. 1998; Gan et al. 2002); yield in drypea (*Pisum sativum* L.; Gan et al. 2002); plant density, seed size, vigour and quality of produced seed in canola (*Brassica napus* L.; Gusta et al. 2004); shoot dry weight, yield in soybean (*Glycine max* L.; Matsuo et al. 2016); and plant height, harvest index, flowering time in camelina (*Camelina sativa* (L.) Crantz; Sintim et al. 2016). The relatively high mean values for days to start of flowering in 2015, in the RILs and the control cultivars can be potentially due to the undesirable or stressful environmental conditions, such as, reduced precipitation, prevalent at the initial stages of crop development (Dash et al. 2014).

The estimated genetic parameters provide an insight about the genetic architecture of the underlying phenotypic traits. The high CV for early season vigor and maturity score indicates the greater phenotypic variability for these traits in the RIL population. Also, the other traits show relatively high phenotypic variation, except for days to maturity. However, the low GCV for all traits demonstrate reduced level of variability at the genetic level, as expected in a cross involving a parent and its derivative line. The broad sense heritability values observed for days to- start of flowering, full flowering and height signify the moderate effect of environment on the expression of the phenotype. Whereas, for early vigor, days to maturity and maturity score, the very low heritability values imply, high G x E interaction and hence, most of the phenotypic variation was attributed to the environment. The moderate to low heritability values for the studied agronomic traits suggest a potential low selection response for these traits within this population. However, this RIL population is suitable for detection of QTL governing flowering time since, QTL-seq based study has successfully identified the genomic regions governing traits with low heritability such as, fruit weight in tomato (Illa-Berenguer et al. 2015).

The correlation values between days to- start of flowering and full flowering reveal the high positive association among the traits, as expected. A line chosen for early flowering will also complete flowering earlier than rest of the lines. The strong correlation of height with days to flowering, observed only in the growth cabinet condition is indicative of significant differences in the field condition in comparison with the controlled environment (reviewed in Poorter et al. 2016). However, there was also a significant relation between early- and late flowering bulks in the replicated field test, with the former being nearly 38% shorter than the latter. As expected, significant negative relationship between the maturity score and days to maturity was observed since higher maturity score was assigned to the early maturing lines.

A large body of evidence suggests flowering time is governed by nearly 300 genes (Bouché et al. 2015), and the impact of the organelle genome on the trait is not reported. Hence, when the bulks were constituted, the individuals exhibiting, early- and late flowering phenotypes were selected only from the 288 RILs derived from the cross involving 'RE2' as the male parent, the donor of only nuclear genome to the progeny. As described earlier, from the 288 RIL individuals, a total of 26 early flowering and 27 late flowering lines were chosen, of which 13 early flowering and 11 late flowering lines were used for next generation sequencing using Illumina platform. Previous studies using QTL-seq analysis for mapping different traits such as, flowering time in cucumber (Lu et al. 2014); 100 seed weight in chickpea (Das et al. 2015); fruit weight and locule number in tomato (Illa-Berenguer et al. 2015); and subgynoecy in cucumber (Bu et al. 2016) have used 10 individuals from each extreme, and the size of their mapping populations ranged from 191 to 232 individuals. Similarly, QTL-seq based mapping of 100 seed weight and root/plant total dry weight ratio in chickpea (Singh et al. 2016), flowering time in broccoli x cabbage cross (Shu et al. 2018) have used 15 individuals from each tail of the distribution. Though there are reports of other studies using higher number of individuals per bulk (reviewed in Zou et al. 2016), the above instances support that the choice of 11 individuals per bulk would be adequate for elucidating the potential genomic region(s) associated with flowering time.

The success of BSA is mainly dependent on the distinct nature of the two bulks for the target phenotype (Zou et al. 2016). From the results of the RCBD trial in 2017, a statistically significant difference (P=0.0197) between the early- and late flowering group of individuals for days to start of flowering was observed. The bulks were also distinct for days to full flowering (P=0.0072), which is substantiated by the strong correlation between days to- start of flowering and full flowering. In addition, significant differences between the bulks were observed for days to maturity (P = 0.0350), height (P<0.0001) and yield (P<0.0001).

### 3.6 Conclusion

The field experiments in 2015, 2016 and 2017 were carried out with the objective of phenotypic characterization of the RIL mapping population generated by crossing 'Royal' with RE2. The evaluation was carried out using the MAD2 and the phenotypic variation contributed by potential soil heterogeneity was corrected using the MAD2 statistical pipeline. Using phenotypic values with improved accuracy, the individuals from the distributional extremes were chosen for constituting the early- and late flowering bulks for the QTL-seq analysis to identify potential genotype-phenotype association for flowering time. In addition, validation of the bulks in an independent RCBD field trial provided additional evidence on their suitability for utilization in sequencing based BSA.

Understanding of the early flowering phenotype and development of markers would help in the breeding of early flowering cultivars of flax and consequent expansion of flax production into the northern part of the grain belt of the Canadian prairies. The outcome of this study would be helpful beyond flax for the improvement of other prairie crops since, development of early maturing cultivar is a universal objective in plant breeding.

# Chapter 4 QTL-seq for the Identification of the Potential Loci Governing Early Flowering Phenotype Observed in the Mutant ‘RE2’

## 4.1 Abstract

The potential genetic basis of the early flowering trait in ‘RE2’, a mutant derived by the treatment of cultivar ‘Royal’ with 5-Azacytidine (5-AzaC), was examined using the next generation sequencing (NGS) based bulked segregant analysis (BSA). The phenotypic characterization of the ‘Royal’ x ‘RE2’ recombinant inbred population was followed by identification of distributional extremes for flowering time (early- and late flowering bulks). The parents and the individuals constituting the bulks were sequenced with unique adapter indices for each line. The DNA sequencing data was analyzed using the QTL-seq pipeline by *in silico* pooling of *high* and *low* bulks, reaping the benefits of both NGS and BSA. The pipeline generated a secondary reference for ‘Royal’ to which the sequencing reads from the early- and late flowering bulks were aligned, and single nucleotide polymorphisms (SNP) were identified. Two parameters, namely SNP-index and ΔSNP-index were estimated, and their moving window averages were plotted. Significant association between genomic region and early flowering phenotype was not observed. However, removal of ambiguous SNP and those common between the bulks identified 363 SNP specific to the early flowering bulk, with a preponderance of transitions. Functional and the positional annotation of SNP using SnpEff suggested that majority of the SNP belonged to the upstream, downstream and intergenic region and had modifier effect, implying variation in the non-coding region. A SNP was identified in the upstream region of the flax gene (*Lus10040921*), homologous to Arabidopsis *LUMINIDEPENDENS* (*LD*), involved in the autonomous flowering pathway. Missense variants with SNP-index of one were predominantly associated with flax genes whose Arabidopsis homologues encode proteins of unknown function localized to the membrane.

## 4.2 Introduction

Breeding short duration cultivars is a universal goal in crop improvement. Identifying genotypes that flower early is important because of its association with maturity (Zhang et al. 2015). Flowering time is under complex genetic control which has been studied in detail in Arabidopsis, with recent evidence describing additional epigenetic regulation (Bloomer and Dean 2017). Hundreds of genes have been described as playing a role in flowering time regulation in Arabidopsis, many have functional orthologues in several crop species and others have diversified to control additional traits (Blümel et al. 2015). For instance, in soybean (*Glycine max* L.), *GIGANTEA (GI)*, a gene involved in photoperiod and circadian clock pathways, also has pleiotropic effect on maturity (Watanabe et al. 2011). In addition, 11 other genes have been identified in soybean, which control both flowering time and maturity (Kong et al. 2018). In maize (*Zea mays* L.), a *FLOWERING LOCUS T-LIKE* gene *ZEA CENTRORADIALIS 8 (ZCN8)* and the recessive *DELAYED FLOWERING 1 (dfl1)* gene which control flowering time were also found to influence leaf number (Li et al. 2016). Hence, studying of genetic basis of flowering time will identify genomic regions underlying this important adaptive trait, as well as generate knowledge on other correlated agronomic traits.

Flax exhibits photoperiod sensitivity and is a long-day plant (Nuttonson 1948), and the majority of the flax cultivars grown in the Canadian Prairies are not influenced by vernalization (Darapuneni et al. 2014). In flax, an early flowering mutant 'RE2' was identified by treating a traditional cultivar 'Royal' with 5-Azacytidine (5-AzaC; Fieldes et al. 1994). The origin of the underlying variation controlling the early flowering trait in 'RE2' is unclear with potential for it to be genetic (single nucleotide polymorphisms-SNP, translocation or deletion) or epigenetic (chromatin variation induced by an altered DNA methylation pattern) or a combination of the two. Despite the uncertainty of its origin, association of flowering time variation with allelic polymorphism can locate controlling factors using quantitative trait locus (QTL) mapping approaches. However, the presence of genetic polymorphism makes interpretation of any epigenetic variation more difficult.

Bulked Segregant Analysis (BSA) is a strategy of associating polymorphic genomic regions (often DNA markers linked to a target genotype) with phenotypic variation in segregating populations (Michelmore et al. 1991). With the advent of next generation sequencing (NGS) technologies, novel variants of BSA have been proposed. MutMap (Abe et al. 2012) identifies SNP underlying the phenotype of interest by combining the principles of both BSA and NGS. The SNP associated with the desired phenotype can be easily identified among the progeny obtained from mutant x wild-type crosses since the number of segregating loci responsible for the phenotype would be minimal because of their near-isogenic nature. In other words, the unlinked SNP are in equilibrium and are expected to segregate in a 1:1 ratio for mutant and wild type alleles. In contrast, the causal SNP (internal to the gene) or those in linkage disequilibrium with the phenotype will not be segregating in a 1:1 ratio and will be preferentially enriched from one of the parents.

QTL-seq, another variant of MutMap methodology was proposed by Takagi et al. (2013). QTL-seq follows the same principle of MutMap using segregants from a mutant x wild-type cross. However, the target trait is quantitative in nature whereas MutMap focuses on qualitative traits and hence, QTL-seq is more appropriate to examine material derived from breeding populations. QTL-seq involves sequencing of pools of segregants formed from individuals at the tails of phenotypic distribution allowing the localization of genomic regions harbouring potential candidate genes influencing the trait of interest (Takagi et al. 2013). In addition, QTL-seq is more efficient than traditional linkage mapping followed by QTL analysis because of the direct identification of genomic regions (tagging) associated with the trait of interest due to the potential enrichment of alleles from a given parent and reduced cost in terms of resources. The NGS based BSA was first used in yeast (*Saccharomyces cerevisiae*) to dissect the sensitivity to 17 chemical substances measured as a quantitative trait (Ehrenreich et al. 2010). In crop plants, QTL-seq has been widely deployed for accelerated identification of agronomically important genomic regions in rice (*Oryza sativa* L.; Takagi et al. 2013, Kadambari et al. 2018), cucumber (*Cucumis sativus* L.; Lu et al. 2014), chickpea (*Cicer arietinum* L.; Das et al. 2015, Singh et al. 2016), tomato (*Solanum lycopersicum* L.; Illa-Berenguer et al. 2015), pigeonpea (*Cajanus cajan* (L.) Millsp.; Singh et al. 2015), brassica (*Brassica napus* L.; Wang et al. 2016), groundnut (*Arachis hypogaea* L.; Pandey et al. 2017) and soybean (*Glycine max* L.; Zhong et al. 2018).

The present study dissecting flowering time in flax is a further modification of this NGS-BSA approach. Here we use an *epi*RIL mapping population derived from a ‘Royal’ x ‘RE2’ cross. In previous analyses, crosses between different genotypes ensure adequate levels of polymorphism segregating in the population. However, the ‘Royal’ x ‘RE2’ derived *epi*RIL mapping population involves a cross between ‘RE2’ and its original progenitor genotype Royal, where the level of allelic polymorphism is expected to be extremely low. Despite this, QTL can still be located to an interval defined by polymorphic markers.

### 4.3 Materials and Methods

### 4.3.1 Sample collection

The ‘Royal’ x ‘RE2’ derived RIL population was phenotypically evaluated and the distributional extremes constituting the early- and late flowering bulks were identified as described in Chapter 3. Single plants of the parents, ‘Royal’ and ‘RE2’, and that of the 13 early flowering and 11 late flowering lines were grown in the growth cabinet under 16 hours light condition and with a day- and night temperature of 22 °C and 17 °C, respectively. The leaf tissues were collected from individual plants 29 days after seeding and were immediately frozen in liquid nitrogen before being stored at -80°C.

### 4.3.2 DNA extraction and quantification

The DNA was extracted from the tissues, using the modified Cetyl Trimethyl Ammonium Bromide (CTAB) protocol (Porebski et al. 1997; Healey et al. 2014). The details of reagent preparation are provided in Appendix A. One-gram of frozen leaf tissue was ground to a fine powder using pestle and mortar chilled with liquid nitrogen. The ground tissue was transferred to a 50 ml falcon tube to which 10 ml of extraction buffer pre-heated at 65°C was added. The sample was incubated in a water bath at 65°C for 1 hour with frequent inversion every 15 minutes. The incubation step was followed by centrifugation at 4800 x g for 6 minutes to remove the debris. The supernatant was transferred to a fresh falcon tube and an equal volume of 24:1 chloroform:isoamyl alcohol was added and mixed to form an emulsion. The sample was centrifuged at 4800 x g for 6 minutes and the aqueous phase was transferred to a new falcon tube. A total of 5 µl of RNAse A (10mg/ml) was added to the solution and incubated at 37°C for 15 minutes. This step was followed by another round of chloroform:isoamyl alcohol extraction. The DNA present in the aqueous phase was precipitated by adding 1/10th volume 3M Sodium acetate (pH 5.2) and two volumes of ice cold 95% ethanol. The sample was incubated at -20°C for exactly 1 hour. After incubation, the sample was centrifuged at 4800 x g for 11 minutes. The pellet was washed with 3 ml of 70% ethanol by centrifugation at 4800 x g for 11 minutes. Finally, the DNA pellet was air dried and the pellet was dissolved in 200 µl of resuspension buffer (RSB; Illumina Inc., USA) and stored at -20°C.

The extracted DNA was quantified using Qubit 2.0 fluorometer (ThermoFisher Scientific, MA, USA) with a broad range (BR) DNA assay kit using the protocol described in Appendix B. The DNA concentration in the samples (ng/ml) were noted. Finally, the concentration in the DNA stocks (ng/µl) were estimated.

### 4.3.3 Shearing of DNA

The isolated genomic DNA was diluted to a concentration of 4 ng/µl in a volume of 60 µl of RSB, and transferred to 0.5 ml Bioruptor microtubes (Diagenode Inc., NJ, USA). Prior to sonication, the samples were vortexed for 10 seconds, centrifuged at 1000 rpm for 10 seconds and incubated on ice for 15 minutes. Initial optimization indicated that nine cycles of sonication in the Bioruptor, where each cycle pulsed for 30 seconds separated by 90 seconds pauses, generated a population where the majority of fragments were ~550 bp in size. Finally, all of the DNA samples were subjected to sonication-based shearing. The quality of fragmented DNA samples was analyzed using 1.5% agarose gel electrophoresis and visualized and documented on Bio-Rad Geldoc XR+ (BioRad, CA, USA). The protocol followed is described in Appendix C.

#### 4.3.4 Sequencing-library preparation

The sheared DNA was used in the construction of sequencing libraries using TruSeq Nano DNA library preparation kit (Illumina Inc., San Diego, USA) following the manufacturer's instructions. The protocol followed is described below.

#### 4.3.4.1 Cleaning-up of fragmented DNA

The sheared DNA sample was centrifuged at 280 x g for 5 seconds to collect the sample. Fifty micro-litres of sheared DNA was placed in a 1.5 ml Eppendorf tube for library preparation. The resuspension buffer (RSB) and sample purification beads (SPB) were incubated at room temperature for 30 minutes. Bead-based cleaning-up of the fragmented DNA occurred with the SPB being thoroughly mixed using a vortex to disperse the beads uniformly in the solution. Each of the samples were incubated with 80 µl of SPB and mixed thoroughly by pipetting. The samples were incubated for 5 minutes at room temperature. After incubation, they were placed on a magnetic stand for 8 minutes to stabilize. The supernatant from each sample tube was removed completely. The settled beads were washed with freshly prepared 80% (v/v) ethanol as follows: with tubes still on magnetic stand, 200 µl of freshly prepared 80% (v/v) ethanol was added and incubated for 30 seconds. After removal of ethanol, the wash was repeated for a second time. Finally, all residual ethanol was discarded using a 20 µl pipette. The beads were air dried for 5 minutes, and then after removal from magnetic stand, 62.5 µl of RSB was added and mixed well by pipetting. The resuspended beads were incubated for 2 minutes at room temperature. The samples were placed on the magnetic stand and incubated for 5 minutes for the beads to settle. A volume of 60 µl of clear supernatant was transferred to 0.2 ml PCR tube.

### 4.3.4.2 End repair and library size selection

The end repair mix (ERP) was centrifuged at 600 x g for 5 seconds. Each of the sample had 40 µl ERP added, and the final volume of 100 µl was mixed well by pipetting. The samples were then placed on a PCR machine and the program with the following temperatures was run: preheat lid option at 100°C; incubation at 30°C for 30 minutes and final hold at 4°C.

For the removal of large DNA fragments, the following clean-up steps were carried out. A volume of 92 µl of SPB was diluted in 92 µl of nuclease free water, per sample, as the desired insert size was 550 bp. The diluted SPB was well mixed using a vortex. Each 100 µl of end repaired sample was transferred to a 1.5 ml Eppendorf tube, to which 160 µl of diluted SPB was added and mixed thoroughly by pipetting. The solution was incubated for 5 minutes at room temperature. The sample was then placed on a magnetic stand and the solution was allowed to clear with a 5 minutes incubation. After the beads were settled and stabilized using a magnet, 250 µl of clear supernatant was transferred to a fresh 1.5 ml Eppendorf tube.

For the removal of small DNA fragments, the following clean-up was performed. The undiluted SPB was mixed using a vortex and 30 µl was added to the 250 µl of collected supernatant and thoroughly mixed. The solution was incubated for 5 minutes at room temperature and then placed on a magnetic stand for 5 minutes. After the beads were settled, all the supernatant was discarded. The beads were washed with freshly prepared 80% (v/v) ethanol as follows: 200 µl ethanol was added to every sample and incubated for 30 seconds on the magnetic stand and then ethanol was removed. The wash was repeated, and ethanol was removed completely using 20 µl pipette. The beads were air dried and suspended in 20 µl of RSB. The samples were removed from the magnetic stand and mixed well by pipetting and samples were incubated for 2 minutes and then transferred back to magnetic stand and incubated for 5 minutes for the beads to settle down. Finally, 17.5 µl of supernatant was transferred to a 0.2 ml PCR tube.

### 4.3.4.3 Adenylation of 3’ ends

The thawed A-tailing mix (ATL) was centrifuged at 600 x g for 5 seconds. The end repaired, size selected DNA was added with 12.5 µl of ATL and mixed thoroughly by pipetting. The sample was centrifuged at 280 x g for a minute. Incubation was carried out in a thermocycler with the following program setting: preheat lid option at 100°C; 37°C for 30 minutes; 70°C for 5 minutes; 4°C for 5 minutes; the total volume was 30 µl.

#### 4.3.4.4 Adapter ligation

The DNA adapters were thawed at room temperature and then centrifuged at 600 x g for 5 seconds. The ligation mix - 2 (LIG 2) was removed from the -20$^{o}$C freezer. Each of the samples had 2.5 µl RSB added, followed by 2.5 µl LIG 2 and finally 2.5 µl of the appropriate DNA adapters. The samples were mixed thoroughly. The 0.2 ml tubes were placed on the thermocycler and incubated with the following setting: preheat lid option set at 100$^{o}$C; 30$^{o}$C incubation for 10 minutes; final cool down to 4$^{o}$C. The total volume was 37.5 µl. The stop ligation buffer (STL) was thawed at room temperature and then centrifuged at 600 x g for 5 seconds. To end the ligation process, 5 µl of STL was added to each sample and mixed well by pipetting.

For cleaning-up the adapter ligated DNA fragments, 42.5 µl of vortexed SPB was added to each sample and pipetted to mix the samples which were incubated at room temperature for 5 minutes and then placed on a magnetic stand for another 5 minutes. After the liquid was clear, the supernatant was discarded. The beads were washed by adding 200 µl of 80% (v/v) ethanol as described earlier. The beads were air dried for five minutes and the samples were removed from the magnetic stand, when 52.5 µl RSB was added and mixed well. The supernatant (50 µl) was transferred to a fresh 1.5 ml Eppendorf tube. The above-mentioned steps for cleaning-up the adapter ligated DNA was repeated using 50 µl of SPB initially, and 27.5 µl of RSB for suspending the beads. Finally, 25 µl supernatant was transferred to 0.2 ml PCR tubes.

#### 4.3.4.5 Amplification of DNA fragments

The samples were placed on ice and added with 5 µl of PCR primer cocktail. A volume of 20 µl of enhanced PCR mix (EPM) was added to each sample and pipetted carefully to mix, and the samples were spun down at 280 x g for 1 minute. The following program was run on the thermal cycler: preheat lid option at 100$^{o}$C; incubation at 95$^{o}$C (3 minutes) followed by 8 cycles of 98$^{o}$C for 20 seconds, 60$^{o}$C for 15 seconds, 72$^{o}$C for 30 seconds; 72$^{o}$C for 5 minutes; finally maintained at 4$^{o}$C. The amplified DNA was spun down at 280 x g for 1 minute and transferred to 1.5 ml Eppendorf tubes. Each sample was added with 50 µl of uniformly suspended SPB and mixed thoroughly by pipetting. After the samples were incubated for 5 minutes, they were placed on a magnetic stand for another 5 minutes. Once the beads settled, the supernatant was discarded, and the beads were washed with 80% (v/v) freshly prepared ethanol. A total of 32.5 µl of RSB was added to the air-dried beads, after removal from the magnetic stand and mixed by pipetting. The samples were incubated for 2 minutes at room temperature and again placed on the magnetic stand for the liquid to clear (~5 minutes). For each sample, 30 µl supernatant containing the sequencing libraries, was transferred to a fresh 0.2 ml tube and stored at -20$^{o}$C.

### 4.3.5 Quantification of sequencing libraries and quality assessment

The prepared DNA sequencing libraries were quantified using a Qubit high-sensitivity (HS) DNA assay kit on a Qubit 2.0 fluorometer (Thermofisher Scientific, MA, USA). The quality of library, based on the insert size, was analyzed using a Bioanalyzer, an automated electrophoresis system, with the high sensitivity (HS) DNA assay chip (Agilent technologies, Germany). The methodology followed for quality assessment using Agilent Bioanalyzer is described in Appendix D.

### 4.3.6 Sequencing of DNA libraries

The prepared libraries were sequenced at the NRC Aquatic and Crop Resources Centre, Saskatoon. Briefly, the libraries were quantified using the KAPA library quantification kit for the Illumina platforms (Kapa Biosystems, MA, USA). Real time PCR was employed, and the concentration of DNA fragments flanked by oligonucleotide sequences P5 and P7, which facilitate the attachment of the library to the flow cell, were determined. After estimating the concentrations, the libraries were diluted to 2 nM using 10 mM Tris-HCl (pH 8.0) with 0.1 % Tween 20. The indexed libraries were pooled by adding equal volumes of each library and sequenced on a single lane. A final volume of 10 µl of the pooled libraries were denatured and diluted using 0.1 N sodium hydroxide (NaOH) and hybridization buffer, respectively, based on the cBot clustering protocol (Illumina). Finally, libraries at a concentration of 20 pM were sequenced on an Illumina HiSeq 2500 platform (Illumina Inc., San Diego, USA) utilizing the HiSeq SBS v4 chemistry with 2 x 125 bp cycles. The binary base call (BCL) files generated by the sequencer were converted to standard FASTQ format using the software bcl2fastq.

### 4.3.7 Analysis of sequencing data using the QTL-seq pipeline

The QTL-seq pipeline developed at the Iwate Biotechnology centre, Japan, was used for the analysis (http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap; Takagi et al 2013). The pipeline required the following programmes and tools to be installed: Perl (v5.8.8), Perl module Math :: Random :: MT :: Auto 6.14, R (version 2.15.0), BWA (version 0.5.9 - r16; Li and Durbin 2009, SAM tools (0.1.8 or before; Li et al. 2009) and FASTX - toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The sequencing data generated from 11 early- and 11 late flowering individuals representing the segregants from the extremes of the RIL population distribution were used in the analysis.

QTL-seq analysis using the pipeline involved six steps. In the first step, the FASTQ files of 'Royal' were loaded into a sub-directory and the sequencing read files were named according to the naming convention of the pipeline (**_ (0-9) *_(1or2) _sequence.txt.gz*; where, 0-9 were unique numbers provided and, 1 and 2 were given for forward and reverse reads, respectively). The reference sequence of 'CDC Bethune' (You et al. 2018) was added to the qualify_read directory. The FASTQ files of the individuals constituting the early- and late flowering bulks were added to the directories named early and late, respectively. The configuration file (*config.txt*) was edited, to adapt the pipeline for the QTL-seq analysis of early flowering time in the 'Royal' x 'RE2' RIL population. The name of the bulks (early and late), parent used to generate the secondary reference ('Royal'), were assigned in the relevant config fields. The score type was set as 'Sanger' and the file name of the available genome reference was also provided. The option, *key3_mode_reference_FASTA* was set to zero since, the secondary sequence for 'Royal' was generated in the pipeline. The number of individuals in each bulk was set to 11 and the type of population defined as RIL. Finally, a shell script provided with the package was run to create the *common.fnc* file.

The second step filters the sequencing reads for quality, and also the quantity of sequence data between the bulks is equalized. For a sequencing read to pass this quality filter, 90% of each read must have a Phred quality score of 30 or above. This step is carried out by running the shell script in the directory named qualify_read as *Run_all_Bats.sh <number>* where, <number> would correspond to 9, 0, 1 for parent used for secondary reference, early-, late flowering bulk, respectively. After filtering, the FASTQ files containing paired reads and reads with broken pairs during processing were generated and their corresponding statistics files were produced for 'Royal'. Also, the files containing the equalized reads for the bulks were generated.

In the third step, the sequencing reads of 'Royal' were aligned to the reference sequence of 'CDC Bethune' using BWA (Li and Durbin 2009). The alignment was processed using Coval (Kosugi et al. 2013), a software to filter spurious alignments and to improve the confidence of the detected variants. The high confidence SNP were replaced in the reference sequence to generate the reference-guided assembly of ‘Royal’. The sequence reads of ‘Royal’ were again aligned to the secondary reference and the SNP were detected, which were mainly due to alignment errors, and were placed in a separate pileup file to remove them in downstream analysis in the pipeline.

The fourth step involved the calculation of SNP-index values across genomic positions for both the bulks. The SNP-index was estimated by the analysis pipeline as the ratio of number of reads with an alternate base relative to the total number of reads aligning to the position. The equalized reads of the early- and late flowering bulk were aligned to the secondary reference using BWA. The Coval (Kosugi et al. 2013) software was used for filtering the detected SNP, and any spurious SNP identified in third step were removed. Separate text files for the early- and late flowering bulks were generated for mismatch filter values 2, 3 and 4, defined as the maximum number of mismatched bases tolerated per read. The files contain SNP-index values for each position along with additional parameters defining the number of reads covering the site and read bases and base quality.

In the fifth step, the SNP data and the other parameter information of both the bulks was merged into a single file with filtering for quality. The SNP positions with depth of coverage less than seven and SNP-index values less than 0.3 in both the early- and late flowering bulks were excluded. However, if the SNP-index values were less than 0.3 in only one of the bulks, the positions were considered to have true SNP.

The final step involved the estimation of a ΔSNP-index and the generation of graphical representation of the data. The ΔSNP-index value is defined as the difference between the SNP-index value of the early- and late flowering bulks. The confidence intervals at 90%, 95% and 99% were estimated by the pipeline using computer simulation. For every read depth, a set of alleles in the given data were sampled, and the ΔSNP-index was calculated. This process was then repeated 10,000 times using a bootstrapping algorithm to obtain the confidence intervals. The ΔSNP-index plots were generated for each chromosome, with chromosomal position plotted along the X-axis and ΔSNP-index plotted on the Y-axis. The data were smoothed by applying sliding window sizes of 4 Mb and 2 Mb with an increment of 50 kb.

### 4.3.8 Annotation of called variants

The file which contained SNP information for both the early- and late flowering bulks along the different positions in the genome was used for further analysis. The SNP in common between both the bulked samples, as well as ambiguous SNP were removed since these are not expected to be associated with the trait. The remaining polymorphic loci were annotated using the SnpEff tool version 4.3 where the functional significance of the detected variation is predicted (Cingolani et. al. 2012). Firstly, a database for flax genome was built in SnpEff using the reference FASTA file and the GFF file. Secondly, the variants present in the input file were annotated using the 'ann' command. The resulting annotation was added to the final field along the row in the given input file (.vcf), for every position. The effect of each variant including, upstream and downstream gene variation, missense variants, synonymous variants were predicted. The different effects of observed variations were further categorized as high, moderate, low and modifier based on functional impact. The variations detrimental to proper functioning of the gene were considered to potentially have high functional effect. The variations which change the amino acid sequence are categorized as having a moderate effect. The low effect variants were predicted to have no effect on protein structure but might affect expression through modifying gene regulation. The variations mostly located in the non-coding region of the genome and the influence of which were hard to determine were considered as having a modifier impact. The homologues of the genes associated with the variations were identified using homology search employing the Basic Local Alignment Search Tool (BLAST) alignment algorithm (tBLASTx; Altschul et al. 1997).

## 4.4 Results

### 4.4.1 Re-sequencing of parents and bulks

In total, ~22 million paired-end reads (125 bp) were generated for the 'Royal' genotype. The alignment of these sequence reads to the reference genome using BWA (Li and Durbin 2010) resulted in an average coverage of 4.48, with ~71% of the reference sequence being covered. A total of 293.7 million and 285.6 million paired-end reads were produced for the early- and late flowering bulks, respectively. The alignment of the short reads of early- and late flowering bulks resulted in ~82% breadth of coverage for both the bulks. In terms of absolute genome size, ~230 Mb of the total 317 Mb in the improved flax reference sequence was covered. The average depth of coverage was observed to be 44.26 and 46.12 for early- and late flowering bulks, respectively (Table 4.1).

**Table 4.1** Summary of reads generated, percentage of genome covered and average depth of coverage for parents and the bulks

| **Parent/Bulk** | **Total number of reads generated** | **Genome coverage (%)** | **Mean depth of coverage** |
|---|---|---|---|
| Royal | 22719458 | 71.29 | 4.48 |
| *Early* | 293682557 | 82.00 | 44.26 |
| *Late* | 285573907 | 82.03 | 46.12 |

### 4.4.2 ΔSNP-index analysis

After aligning the sequencing reads from the early- and late flowering bulks to the ‘Royal’ secondary reference sequence, the SNP-index values across 275,571 polymorphic positions throughout the genome were estimated. During this process, the loci with SNP-index values less than 0.3 in either of the two bulks were excluded to remove potential artifacts resulting from low alignment depth. After filtering the SNP on the basis of quality and minimum read coverage, the ΔSNP-index was estimated for the 243,393 (88%) filtered positions. The distribution of SNP identified between the bulks and the ‘Royal’ secondary reference, across each chromosome, using a window of size 2 Mb is presented in Figure 4.1. The ΔSNP-index plots, representing the difference between the bulks, were developed using the average ΔSNP-index values using a window size of 2 Mb with an increment of 50 kb (Figure 4.2). Based on the ΔSNP-index graphs (Figure 4.2), ΔSNP-index was observed to be zero for most of the regions.

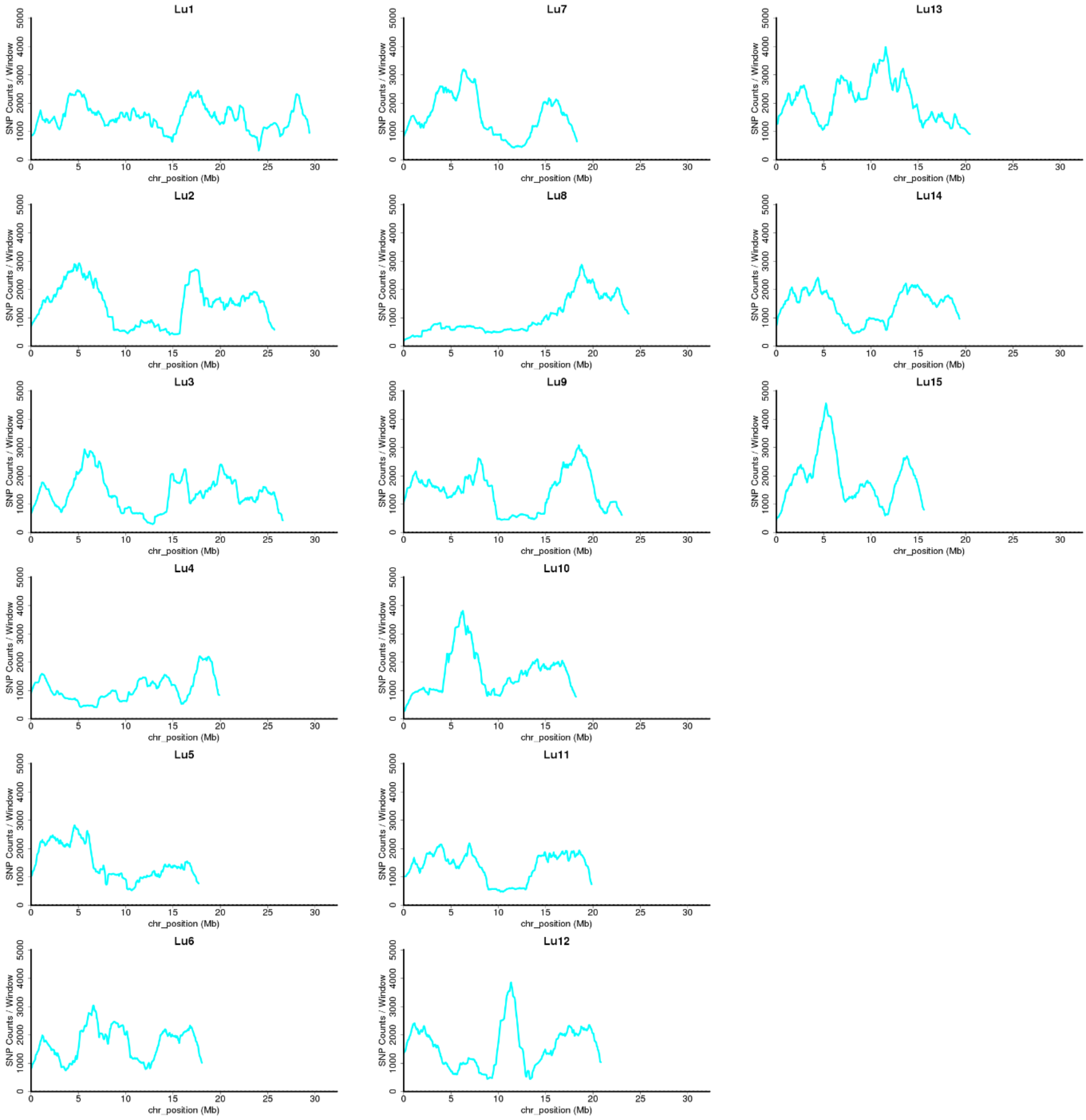

**Figure 4.1** Distribution of SNP across 15 chromosomes of the flax genome. The X-axis represents the chromosome position in Mb and Y-axis represents the SNP count in the window size of 2 Mb.

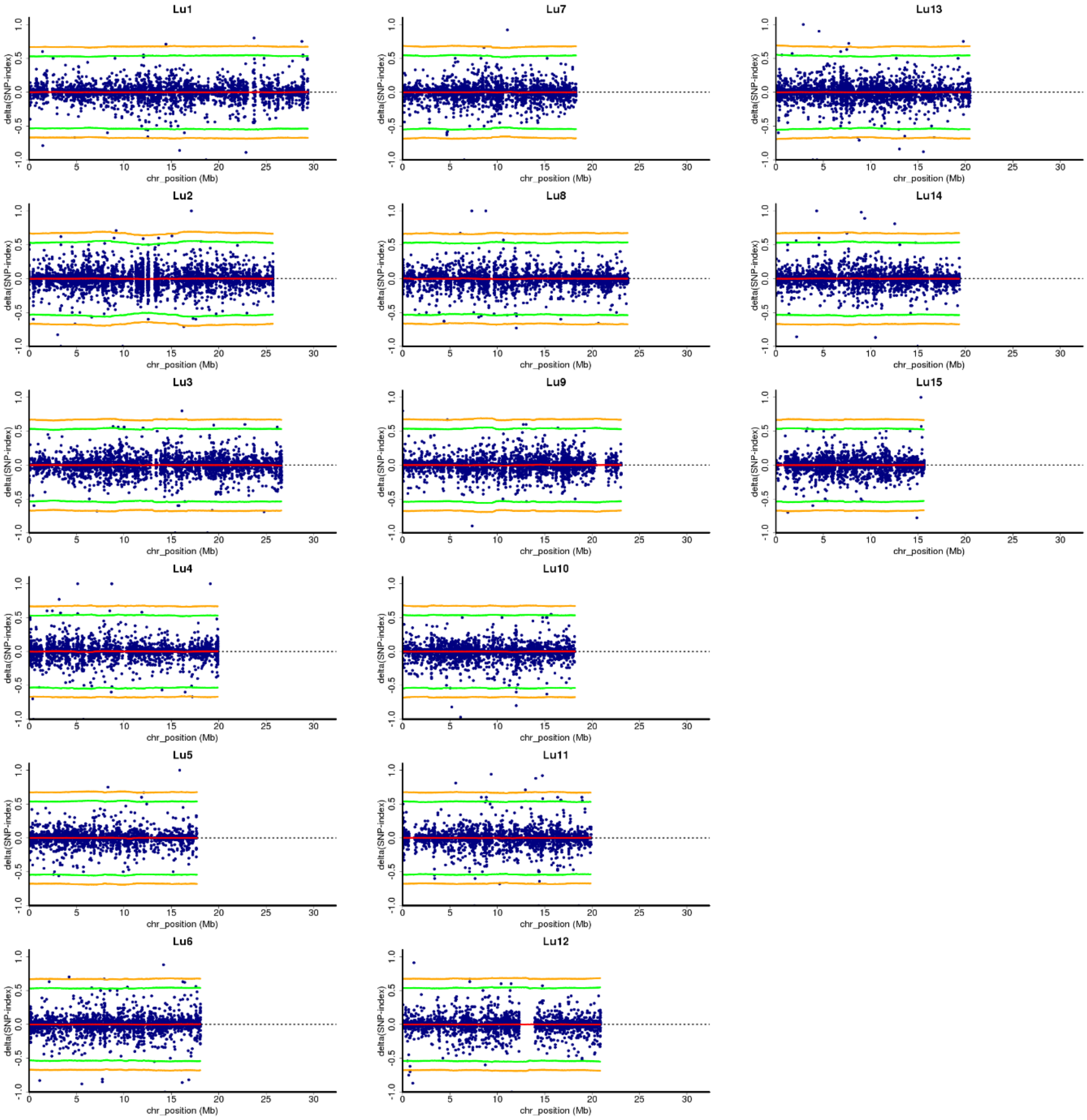

**Figure 4.2** ΔSNP-index plot – the blue dots represent the ΔSNP-index along various positions on the chromosome. The red line indicates the moving window average of ΔSNP-index. The orange and green lines represent the level of significance at P<0.01 and P<0.05.

### 4.4.3 Characterization of SNP

The total of 275,571 SNP positions were identified using the QTL-seq pipeline. These polymorphisms are detected from comparisons to the ‘Royal’ reference sequence. These data were further filtered to remove SNP loci with common and ambiguous SNP between the bulks. The common SNP were defined as possessing the same allele in both the early- and late flowering alignment i.e. they were only polymorphic to the reference sequence, whereas the ambiguous loci were defined as those where multiple alleles were identified within one of the bulked samples. These loci can not be responsible for the trait variation. The number of SNP loci surviving stringent filtering was 724. Among the SNP loci, 281 and 361 positions were polymorphic in only one of the early- or late flowering bulks, respectively. The remaining 82 loci exhibited polymorphism to the reference allele in both of the bulks. Further examination revealed that 61 of the polymorphic loci identified in the early-flowering bulk possessed a SNP-index of one and zero in the late flowering bulk, indicting that these loci were polymorphic and every individual within each bulk had identical alleles. The 724 single nucleotide positions were distributed all through the genome (Figure 4.3), at a density of 1 SNP/1.0 Mb in early-flowering bulk, and 1 SNP/0.8 Mb in the late flowering bulk. As anticipated due to structural similarities of bases, functional annotation of the SNP loci revealed that the total number of transition mutations (purine to purine (or) pyrimidine to pyrimidine) outnumbered the transversion mutations (purine to pyrimidine (or) pyrimidine to purine) detected (Figure 4.4). The most frequently observed mutation was the, Cytosine to Thymine transitions (126 in total) whereas, the most frequent transversion class observed was the Thymine to Adenine (73 in total).

### 4.4.4 Annotation of SNP

Functional annotation of the 724 SNP loci was predicted using the SnpEff variant annotation tool (Figure 4.5; Cingolani et al. 2012). In the early-flowering bulk, the effect of the mutation was defined as, moderate, low and modifier in 11, 17, 335 cases respectively. No SNP were predicted to result in a truncated protein resulting from the introduction of a stop codon (high impact). The majority of the SNP were detected in non-coding potential regulatory regions (upstream gene - 156 SNP), whereas SNP in intron sequences were identified at a similar rate as coding sequences (15 SNP). Additionally, a small number (14) of synonymous SNP predicted to have low mutagenic effect were identified. The SNP loci with the greatest level of mutagenic potential were the mis-sense class (moderate impact) where a total of 11 were detected.

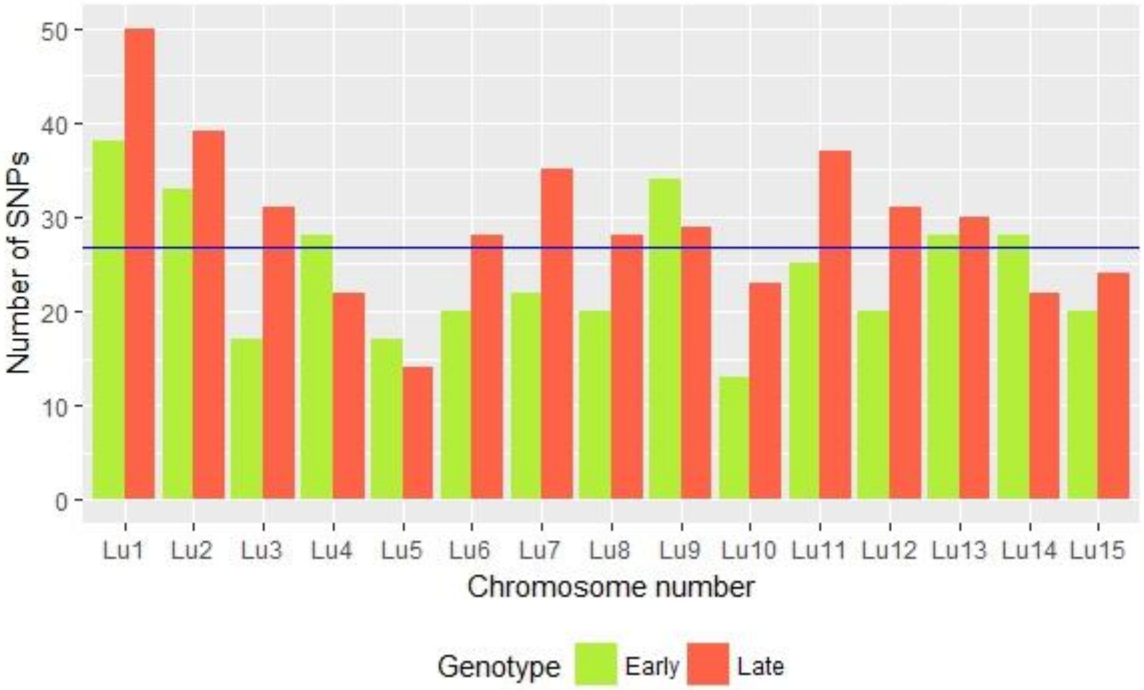

**Figure 4.3** Distribution of SNP across the chromosomes in the early- and late flowering bulks. Green represents the early flowering bulk, red represents the late flowering bulk and the blue line represents the average (~27 SNP).

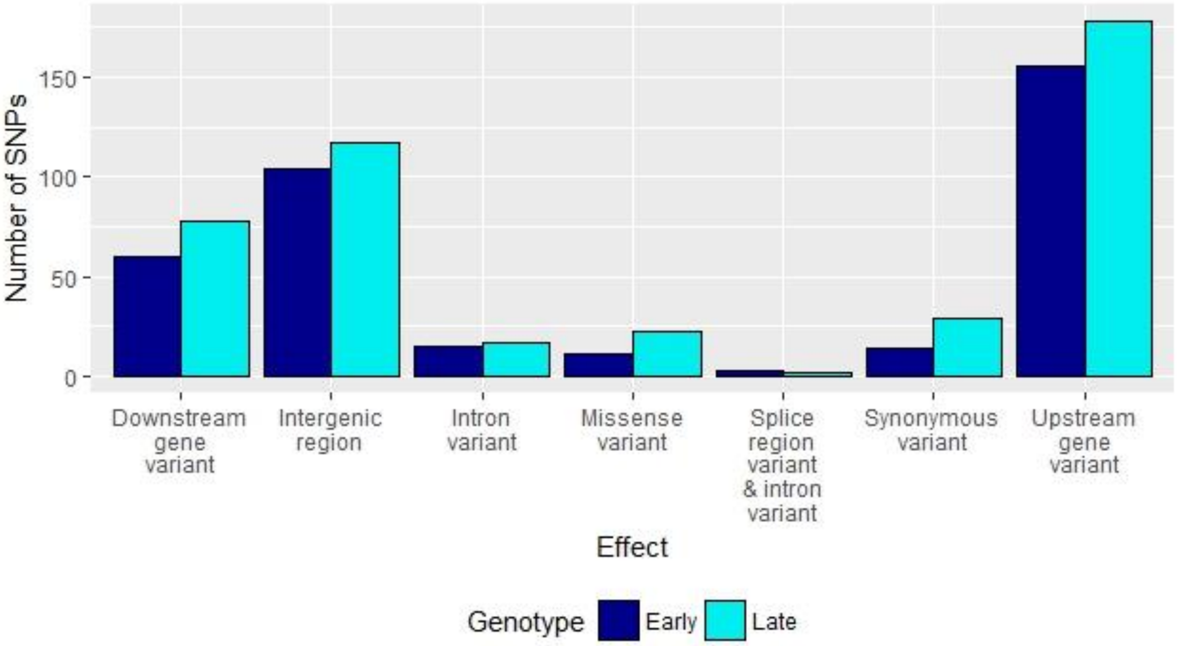

**Figure 4.4** Nucleotide variations categorized based on the positional and functional impact. Blue represents the early flowering bulk and cyan represents the late flowering bulk.

Similar functional analysis was performed on the late flowering bulk. A total of 22 moderate, 31 low and 390 modifier variants were detected. Similar to the early flowering bulk, upstream nucleotide variants (modifier effect) were the most common variation detected (178 SNP in total). In addition, 29 synonymous coding variants were identified. Single nucleotide polymorphisms at only two positions were found to be non-synonymous substitutions.

The flax genes were assigned Arabidopsis functional annotations based on sequence homology that allowed the identification of gene potentially involved in the transition to flowering. Among the 363 SNP specific to the early flowering bulk, 113 genes (including the flanking genes for variations found in intergenic regions) had Arabidopsis homologues based on the homology parameters used in the BLAST alignment. Two of the eleven missense variants in the early flowering bulk were associated with flax gene identifiers *Lus10018444* and *Lus10011571*, which had Arabidopsis homologues. In the tBLASTx based homology search results, gene *Lus10018444* was the homologue of the Arabidopsis genes *AT1G61260* (percent identity=59%; E-value=1e$^{-51}$), *AT4G04990* (percent identity=61%; E-value=1e$^{-15}$) and *AT5G54300* (percent identity=58%; E-value=8e$^{-37}$) and gene *Lus10011571* was homologous to *AT2G22795* (percent identity=14%; E-value=5e$^{-43}$) and *AT3G28770* (percent identity=14%; E-value=5e$^{-14}$). The SNP present in the upstream region of gene *Lus10040921* on chromosome 15, a homologue of the Arabidopsis *LUMINIDEPENDENS (LD)* gene *AT4G02560* and the annotation suggests that in Arabidopsis, recessive mutants are late flowering.

## 4.5 Discussion

QTL mapping is a method for dissecting traits with complex inheritance. It allows the position of genomic regions underlying quantitative trait variation to be identified and commenced with analysis of traits including fruit weight, total soluble solids and fruit pH in tomato by Paterson et al. in 1988. QTL mapping has been applied to other systems where, several QTLs have been mapped in crop plants including rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), maize (*Zea mays* L.) and soybean (*Glycine max* L.; Price 2006). The first plant QTL was cloned only in 2000 in tomato (*Solanum lycopersicum* L.; Frary et al. 2000). The QTL underlying flowering time, an important adaptive trait, has been studied in several crop plants. In self-pollinating plants including rice and Arabidopsis, flowering time is largely governed by a few QTL with large effects in contrast to cross-pollinated species like maize where, several minor QTL with additive effect control the trait (*Zea mays* L.; Buckler et al. 2009). Similarly, in soybean (*Glycine max* L.), four QTLs were mapped to different linkage groups, of which one major QTL explained 69.7% of the observed phenotypic variance (Yamanaka et al. 2001). The effect of polyploidy further complicates quantitative traits exemplified by the control of flowering time in canola (*Brassica napus* L.) where QTL were mapped to 10 of the 19 chromosomes in a doubled haploid mapping population (Raman et al. 2013). Further, map-based cloning of positional QTLs identified SNP that were found to be associated with loss of seed shattering during domestication (Konishi et al. 2006), cold tolerance (Ma et al. 2015) and blast resistance (Li et al. 2017a) in rice.

The detection of QTLs controlling a trait of interest by conventional QTL mapping requires identification of markers distributed in the genome, polymorphic between the parents differing for the trait of interest (Simon et al. 2008). Also, classical QTL mapping methodology is resource intensive since, all individuals in a mapping population must be genotyped. Bulked segregant analysis (BSA) is a technically less intensive and rapid methodology developed to identify genomic regions conditioning the phenotype segregating in a mapping population (Michelmore et al. 1991). Bulked segregant analysis is less sensitive to the possible random errors in phenotyping (Schneeberger et al. 2009) in contrast to classical QTL mapping. The rapid development of genomics resources has led to the generation of reference genome sequences for a range of species as next generation sequencing (NGS) technologies have become more accessible (Goodwin et al. 2016). The combination of BSA and NGS offers new advantages that can be exploited by the QTL-seq strategy. The QTL-seq methodology combines the power of recombination and positioning of QTL to a single step called ‘gene-tagging’. Several qualitative traits such as disease resistance are effectively mapped using BSA (reviewed in Zou et al. 2016). In addition, quantitative traits including flowering time have been identified by employing BSA. In cucumber (*Cucumis sativus* L.), a QTL with a major effect which was homologous to the *FLOWERING LOCUS T* (*FT*) gene of Arabidopsis was detected by employing QTL-seq (Lu et al. 2014). Using a mapping population developed from broccoli x cabbage, a QTL harbouring a homologue (*BolGRF6*) of an Arabidopsis flowering time control gene was identified using the same strategy (Shu et al. 2018).

The availability of a high-quality flax reference sequence (You et al. 2018) is a foundation for future genomics analyses in flax. The cost efficiencies associated with NGS technology enable QTL-seq analysis examining flowering time to be performed. A strategy was developed using two bulks (early- and late flowering) identified from a recombinant inbred line (RIL) mapping population derived from a ‘Royal’ x ‘RE2’ cross. The RIL population was evaluated in three field seasons and once in the growth cabinet to define the early- and late- flowering bulks (Chapter 3). Individual lines belonging to the two bulks along with their parents were sequenced on an Illumina Hi-seq 2500 platform to generate a wealth of sequence data. The analysis pipeline developed a ‘Royal’ reference sequence by aligning to the reference sequence of ‘CDC Bethune’ (You et al. 2018), and the sequencing reads of early- and late flowering bulks were aligned to the secondary reference to identify the sequence polymorphism. The breadth of coverage observed on aligning the sequencing reads from ‘Royal’ to the ‘CDC Bethune’ reference (~71%) and the early- and late flowering bulk reads to the ‘Royal’ secondary reference (~82%) potentially implies ~20% of the genome cannot be covered, which might be due to the absence of these regions in ‘Royal’ in comparison to ‘CDC Bethune’ or the improper alignment of sequencing-reads in the repetitive region. However, this observation could partially be due to variations at the levels of sequencing library preparation, batch effect of DNA sequencing and read alignment (Sims et al. 2014). In chickpea (*Cicer arietinum* L.), candidate genes underlying seed weight have been identified using genome-wide coverage as low as ~3x for the bulks (Singh et al. 2016). Hence, in the present study, the relatively high depth of coverage observed for the bulks (~44x for early flowering bulk, ~46x for late flowering bulk) was sufficient to dissect potential genomic region(s) associated with the early flowering phenotype. We implemented an improved strategy of *in silico* pooling in contrast to the bulking of individuals at the DNA level followed in the chickpea study, which may be responsible for the comparatively higher depth of coverage.

According to the principle of BSA, the genomic region controlling the trait of interest will be uniform in the individuals within the bulks and different between the two bulks and hence will exhibit unequal representation of the parental genomes, while the other regions being equally contributed due to recombination and random chromosome assortment. The moving-window average of ΔSNP-index when plotted with the chromosomal position along the X-axis, would help to visualize these region(s) polymorphic between the parents. However, no genomic interval was found to underlie the phenotypic difference between the bulks for flowering time. The output of the pipeline in the form of ΔSNP-index plot (Figure 4.2) is generated by analysis of the genomic region as bins (2 Mb) and hence, did not have sufficient resolution to examine the effect of individual SNP on the phenotype.

Whole genome resequencing of mutant lines is a proven approach to detect potential mutations in genes resulting in modified phenotypes (Shirasawa et al. 2016). The SNP detected by the pipeline were found to be distributed across the genome without bias. The chromosome 1 is the largest in size by assembly (29.4 Mb) and has the maximum number of single nucleotide variations in both the early- and late flowering bulks. Higher number of transitions when compared to transversions was as expected (Lyons and Lauring 2017) and similar to that observed in *Lotus japonicus* L. (Mohd-Yusoff et al. 2015) and *Solanum lycopersicum* L. (Shirasawa et al. 2016) on treating with chemical mutagens.

SnpEff (Cingolani et al. 2012) is a tool used to infer the positional and functional impact of SNP in resequencing studies. Out of the total 363 SNP specific to the early flowering bulk, 104 SNP were present in the intergenic region, 156 and 60 SNP were in the upstream and downstream regions, respectively (an interval of 5,000 bp from the gene was demarcated as upstream and downstream region by the SnpEff annotation tool). Based on the functional impact, the majority of the SNP were classified as modifier type by SnpEff, suggesting their presence in the non-coding region of the DNA. The role of regulatory-element mutations in domestication and their potential to generate alleles favoured in crop breeding is well established (Swinnen et al. 2016). The changes in the regulatory regions in the genome is preferred over the variation in the exonic region because the former results in comparatively less deleterious effects. In maize (*Zea mays* L.), the insertion of a transposable element in the promoter region of a gene controlling photoperiod response has been shown to repress the gene expression (Yang et al. 2013a). This reduced the photoperiod responsiveness of the crop and facilitated its adaptation to different geographical regions. Similarly, in soybean (*Glycine max* L.), a SNP in the promoter modified the motif of a *cis*-element resulting in determinate growth habit (Liu et al. 2010). In another study, the flowering time differences among maize accessions were found to be controlled by genetic polymorphisms in a distant regulatory region ~70 kb upstream of *AP2-like* gene (Salvi et al. 2007). Also, the compact panicle phenotype of modern rice (*Oryza sativa* L.) cultivars is the effect of a single SNP present ~11 kb upstream of a ligule development controlling gene (Zhu et al. 2013). Hence, it is evident that modification in even distant regulatory region, by possible association with transcriptional regulators, transcription factor binding sites and other means, can consequently influence gene expression resulting in modified phenotype.

A SNP was observed upstream of the flax gene *Lus10040921,* and *LUMINIDIPENDENS* (*LD*) was its Arabidopsis homologue. *LD* is involved in the autonomous pathway and likely controls the timing of transition into the floral meristem in Arabidopsis by regulating the expression of *LEAFY* (*LFY*). It is a transcriptional regulator confined to the nucleus and its expressed only in the shoot apical meristem and primordial leaves (Lee et al. 1994; Aukerman et al. 1999). The QTLs harbouring candidate genes for flowering time in flax have not been reported in literature. In our study, SNP were found in the non-coding region of the genome as discussed earlier. Hence, further investigation is required to determine the molecular function of the identified SNP upstream to *LD* in the present study using state-of-the-art technology such as clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system.

Missense variants have been observed to make remarkable changes in the phenotype as in loss of resistance to powdery mildew in Arabidopsis (Wawrzynska et al. 2008) and development of seedless grapes (Royo et al. 2018). In this study, among the genes associated with missense variation, two had homologues in Arabidopsis. The flax gene *Lus10018444,* based on homology search using BLAST (tBLASTx), was homologous to *AT1G61260* (percent identity=59%; E-value=1e$^{-51}$), *AT4G04990* (percent identity=61%; E-value=1e$^{-15}$) and *AT5G54300* (percent identity=58%; E-value=8e$^{-37}$). While *AT1G61260* encodes a protein similar to cotton fiber protein localized in the chloroplast, *AT5G54300* encodes a variant of cotton fiber-like protein present in the chloroplast and membrane. *AT4G04990* is a serine/arginine repetitive matrix like protein also present in the membrane. The gene *Lus10011571* had two Arabidopsis homologues, *AT2G22795* (percent identity=14%; E-value=5e$^{-43}$) and *AT3G28770* (percent identity=14%; E-value=5e$^{-14}$). *AT2G22795* is a hypothetical protein present in the golgi apparatus and *AT3G28770* is a putative transmembrane protein confined to membrane and nucleus. The specific molecular function of these genes and their involvement in different pathways related to flowering time is unknown.

Beyond the absence of any variation at the nucleotide level between the bulks, there are other conceivable reasons for any genetic basis being undetected by the pipeline. Firstly, the annotation of flax genes was carried out training AUGUSTUS (Stanke and Morgenstern 2005) pipeline using Arabidopsis gene models. However, there may be genes specific to flax involved in flowering time since, both the species diverged nearly 106 million years ago (http://www.timetree.org/; Hedges et al. 2006). Secondly, upon whole genome resequencing using the NGS technology, not all regions of the genome have uniform depth of coverage. Low sequencing depth in certain portion of the genome may result in non-identification of

SNP in flanking region. These intervals can be subjected for targeted sequencing at higher depth to analyse their contribution to the early flowering phenotype. Breeding populations have been already developed using ‘RE2’ as a parent in different genetic backgrounds (Dr. Helen Booker personal communication). These populations can be harnessed to further study the potential genetic basis of the variation for flowering time between ‘Royal’ and ‘RE2’ since, different genetic backgrounds are found to modify the gene expression (Chandler et al. 2013).

## 4.6 Conclusion

Early flowering is an important trait for flax improvement for its adaptation to short growing season in the northern region of the prairies. QTL-seq, a proven methodology for mapping of quantitative traits was employed to identify the potential genetic basis of the early flowering phenotype segregating in the ‘Royal’ x ‘RE2’ recombinant inbred population. Since no specific genomic region was found to be associated with the flowering time trait using the pipeline, the SNP specific to the early-flowering bulk was characterized *in silico*. Recently, a statistical package in R, namely QTLseqr for NGS based BSA analysis was proposed (Mansfeld and Grumet 2018). Making use of this improved tool, analysis was carried out to validate the results presented in this chapter.

# Chapter 5 Whole Genome Resequencing Based BSA Analysis Using QTLseqr Package

## 5.1 Abstract

The DNA sequencing data from the segregants representing extreme phenotypic values for flowering time in the mapping population (‘Royal’ x ‘RE2’) were reanalysed using QTLseqr package to validate the result from the older QTL-seq pipeline. The sequencing-reads from the recombinant inbred lines were aligned to the ‘CDC Bethune’ reference assembly employing Bowtie2, and best hits were extracted with custom PERL scripts. Using the alignment files from the early- and late flowering bulks, variants were called using Genome Analysis Tool Kit (GATK) HaplotypeCaller. Filtering of variants using the parameters such as read depth and reference allele frequency, and statistical analysis identified two genomic regions on chromosomes 9 and 12 associated with early flowering phenotype with significant ΔSNP-index. The flax genes harbouring in the region delimited by significant variants were homologous to LATE EMBRYOGENESIS ABUNDANT (LEA) HYDROXYPROLINE-RICH GLYCOPROTEIN FAMILY, MAINTENANCE OF MERISTEMS-LIKE (MAIL), CYTOCHROME P 450 87A3 and PHLOEM PROTEIN 2-A12 encoding genes. QTLseqr algorithm functions by taking a weighted average of ΔSNP-index to account for linkage disequilibrium, and hence, genes in the flanking region up to the closest variant with a ΔSNP-index of one were also analysed, and most of the genes involved in abiotic stress response with indirect association with flowering time were identified. In addition, a few genes with no homologues were also identified, suggesting the potential role of these flax specific genes in flowering time control.

## 5.2 Introduction

Quantitative trait locus (QTL) mapping, as stated by Prof. Trudy F. C. Mackay (2001) is based on the principle that

> *“if a QTL is linked to a marker locus, there will be a difference in mean values of the quantitative trait among individuals with different genotypes at the marker locus”*.

Conventional QTL mapping involves the following steps: crossing two parents with contrasting phenotypes to generate a mapping population, extensive genotyping of the segregating population with markers and the statistical analysis for detecting the QTL. Once the QTL interval has been identified, further fine-mapping of the region with additional markers using a much larger mapping population to increase recombination in the interval will identify the candidate gene underlying the phenotypic variation. The limitation of this approach is the need for polymorphic DNA markers distributed across the genome. However, the advent of next generation sequencing (NGS) and the availability of reference genomes for several plant species led to the identification of huge number of markers across the genome and the development of the mapping-by-sequencing strategy, which not only identifies the genomic region but also unravels the sequence variation associated with the altered phenotype (Schneeberger and Weigel 2011; Doitsidou et al. 2016). Use of isogenic mapping populations derived from mutants and their progenitors for mapping complex traits is being achieved with the prevalence of NGS platforms (Schneeberger 2014).

ShoreMap, initially developed in Arabidopsis, was the first strategy for applying mapping-by-sequencing, based on the principle of allele frequency differences between mutant and its wild type (Schneeberger et al. 2009). Later, Abe et al. (2012) developed a modified mapping-by-sequencing methodology called MutMap and used a novel statistic named the SNP-index, defined as the number of reads that harbour a variant allele relative to the total number of reads aligning at that locus. Since, MutMap is more suitable for qualitative traits, Takagi et al. (2013) developed the QTL-seq strategy in rice, for mapping quantitative traits combining the principles of bulked segregant analysis (BSA) and NGS. The QTL-seq strategy can be adapted using biparental breeding population for mapping polygenic traits. The individuals exhibiting extreme phenotypes were chosen from a segregating population which were pooled and sequenced. During sequencing, if the lines constituting the bulks are sequenced separately with individual indices, recombination events across the genome can also be investigated in future studies (Candela et al. 2014). In QTL-seq, in addition to SNP-index, another statistic namely the ΔSNP-index representing the difference between the SNP-index of the high- and low bulks was used. Additionally, RNA sequencing data has been used in the mapping of expressed regions in maize (Liu et al. 2012) and wheat (Trick et al. 2012) to complement DNA sequence reads approaches.

The QTL-seq (Takagi et al. 2013) pipeline provides the complete bioinformatic workflow for identifying the genomic region controlling the quantitative trait of interest. Briefly, the sequencing reads for one of the parents used to develop the biparental mapping population is aligned to the reference genome to generate a secondary reference. This is achieved by replacing high confidence single nucleotide polymorphisms (SNP) into a FASTA sequence retaining the space and thus the structure of any associated annotation. The sequencing reads of the bulks, constituted by lines with extreme phenotypic values will be aligned to the secondary reference and SNP loci can be detected. The SNP-index and ΔSNP-index are estimated for high confidence SNP loci, and the genomic region potentially underlying the trait of interest will be visualized with a moving window average.

The results obtained from the QTL-seq pipeline (Chapter 4) was validated by analysis of the data using the recently developed alternate algorithm named QTLseqr (Mansfeld and Grumet 2018) and are presented in this chapter. Further, this investigation was carried out for increasing the robustness of the analysis. The sequencing reads were aligned to the flax reference genome using Bowtie2 (Langmead and Salzberg 2012) instead of Burrows Wheeler Aligner (BWA; Li and Durbin 2009) used in the previous pipeline. The alignment files belonging to the respective bulks were merged and variants were called using Genome Analysis Tool Kit (GATK) HaplotypeCaller (De Pristo et al. 2011). These genotypic variants were investigated for their impact on phenotypic variation using the highly user-configurable QTLseqr package, employing an improved statistic called tricube weighted moving average of ΔSNP-index.

## 5.3 Materials and methods

### 5.3.1 Alignment of sequence reads to the reference genome

The raw DNA sequence reads of the early-flowering and late flowering lines (Chapter 4 - methodology) were processed using Trimmomatic (Bolger et al. 2014) for the removal of low-quality sequence reads and Illumina adapter indices used for multiplexing and sequencing. Processed paired-end sequence reads resulted in four output files, among which two files contained the forward and reverse reads as paired output. The other two files comprised reads, with broken pairs that were removed from further analysis. The paired-reads were aligned to the flax reference genome using the Bowtie 2 algorithm (Langmead and Salzberg 2012). The mixed and discordant alignment options were disabled, and local alignments were performed. With parameter –K set to 50, Bowtie 2 looked for 50 discrete and valid alignments for every read. The output in sequence alignment map (SAM) format from the Bowtie 2 was parsed with a custom PERL script, developed at Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre to extract unique alignments or in the case of multiple alignments only the best hit. The SAM file was converted to its binary version (BAM format), sorted and indexed using SAMtools (Li et al. 2009).

### 5.3.2 Variant calling

The sorted-BAM files of 11 early-flowering lines and ‘RE2’ were combined using the SAMtools *merge* function, to a single BAM file. Similarly, 11 late flowering lines and ‘Royal’ were merged into a single BAM file and both files were again sorted and indexed.

The metadata associated with the library including sequencing platform unit and sample name were added to the header of each BAM file using the *AddOrReplaceReadGroups* function of Picard tools (http://broadinstitute.github.io/picard/) to make them compatible for variant calling utilizing the GATK HaplotypeCaller (De Pristo et al. 2011). The output format was set as genomic Variant Call Format (gVCF), which contains comprehensive information for all sites irrespective of presence or absence of variation to facilitate the downstream analysis combining samples. The ploidy level for the analysis was set at four to enable the variant caller algorithm to detect all possible alternate alleles at a single locus within the bulked sample. Quality control of the sequence read is performed automatically by HaplotypeCaller and reads were removed where Phred scores were less than 20.

The *combineGVCF* tool of GATK was employed to merge the gVCF files of the early- and late flowering bulked samples into a single file (combined GVCF), from which the *genotypeGVCF* tool generated appropriate genotype likelihoods after traversing across the samples and each locus. This VCF file was passed through the *VariantToTable* tool of GATK, and the relevant fields required for analysis by the QTLseqr pipeline were extracted into a table. The relevant fields include: chromosome ID, nucleotide position, reference allele, alternate allele(s), genotyping quality, allele depth and depth of coverage.

### 5.3.3 Analysis using QTLseqr package

The QTLseqr (Mansfeld and Grumet 2018; https://github.com/bmansfeld/QTLseqr) analysis was carried out using the early- and late flowering phenotypes as the *high* and *low* bulked samples, respectively. The GATK output table containing the SNP information of the bulks was imported using the *importFromGATK* function. In addition to importing the data, this function estimates the reference allele frequency, SNP-index and ΔSNP-index for all the detected polymorphic nucleotide positions. Reference allele frequency was determined as the ratio between reference allele count and the total read depth in combined bulks. SNP-index was estimated as the ratio of number of reads with alternate alleles at the particular position to the total number of reads aligned at the position and ranges from zero to one. ΔSNP-index is the difference between the SNP-index of the high and low bulks and can vary between -1 and +1.

The distribution of read depth and reference allele frequencies in the raw data was graphically represented as a histogram to examine alignment coverage of the genome. Variants were filtered based on sample- and total- read depth and the reference allele frequency was used to remove low confidence data such as SNP at positions with very low coverage as well as SNP with dense coverage representing multiple alignment artefacts to repeat regions. The parameters used for analysis were as follows:

*refAlleleFreq* = 0.20 (filters for SNP with reference allele frequency > 0.2 and less than 0.8)
*minTotalDepth* = 14 (minimum total coverage)
*maxTotalDepth* = 400 (maximum total coverage)
*minSampleDepth* = 7 (minimum coverage in individual bulks)
The population structure (popStruc) was set as “RIL” and the bulk size was set to 12

Using the *runQTLseqAnalysis* function of the pipeline, the weighted average of ΔSNP-index across the chosen window size of 1 Mb was estimated. Within each window, higher weights were allocated to SNP closer to the variation in focus, and confidence intervals were generated using bootstrap computer simulations using 10,000 iterations. The average read depth across each window was combined with the 95th and 99th quantile of the simulated Δ-SNP index values to determine the confidence intervals across the genome.

Finally, the Δ-SNP index values across the genome was plotted along with confidence intervals at 95% and 99% to identify the location of potential QTL. The *getQTLTabl*e function exported the regions considered significant, at the specified confidence interval.

### 5.3.4 Variant annotation

The nucleotide variants were annotated using the SnpEff algorithm (Cingolani et al. 2012) to identify the potential positional and functional impact of SNP on genes closely linked to the variation. The genes localized in the interval between the variants with a tricube ΔSNP-index above the 95% confidence interval threshold and the SNP with a ΔSNP-index value of one, adjacent to the significant region identified by the algorithm were further investigated. The coordinates of the genes in the region were extracted from the General Feature Format (GFF) file and the corresponding DNA sequence information was extracted from the flax reference assembly using the *getfasta* utility of the BEDTools suite (Quinlan and Hall 2010).

Homology searches of the extracted genes against the National Center for Biotechnology Information (NCBI) non-redundant protein database were carried out using BLASTx (Altschul et al. 1997) to assign a putative gene function with an E-value threshold of E< $e^{-10}$.

## 5.4 Results

### 5.4.1 Sequencing data

Whole genome sequencing of the parents and the lines constituting the bulks, generated on an average ~26 million paired-end reads for each line (Table 5.1). While aligning the reads to the 'CDC Bethune' draft reference, the coverage for 'Royal' and 'RE2' was 81.97% and 82.60%, respectively. The mean breadth of coverage for the RILs was 82.27 %. The average depth of coverage was estimated to be the lowest for the early flowering line Plant 6-E1 (~8x) and the highest for the Plant 6-E3 at ~18x (Table 5.1). The maximum depth of coverage observed for combined dataset was ~8,023.

### 5.4.2 Filtering of input SNP data

The input variant table consisted of information for 608,426 polymorphic loci. A histogram depicting the read depths observed for these data is presented in Figure 5.1. The read depth in the raw data ranged between 'zero' and '987' in the *high* bulk, and 'zero' and '873' in the *low* bulk. The total read depth varied from 1 to 1,657. The distribution of reference allele frequency in the input data (Figure 5.2), indicate that maximum number of SNP had a reference allele frequency of zero.

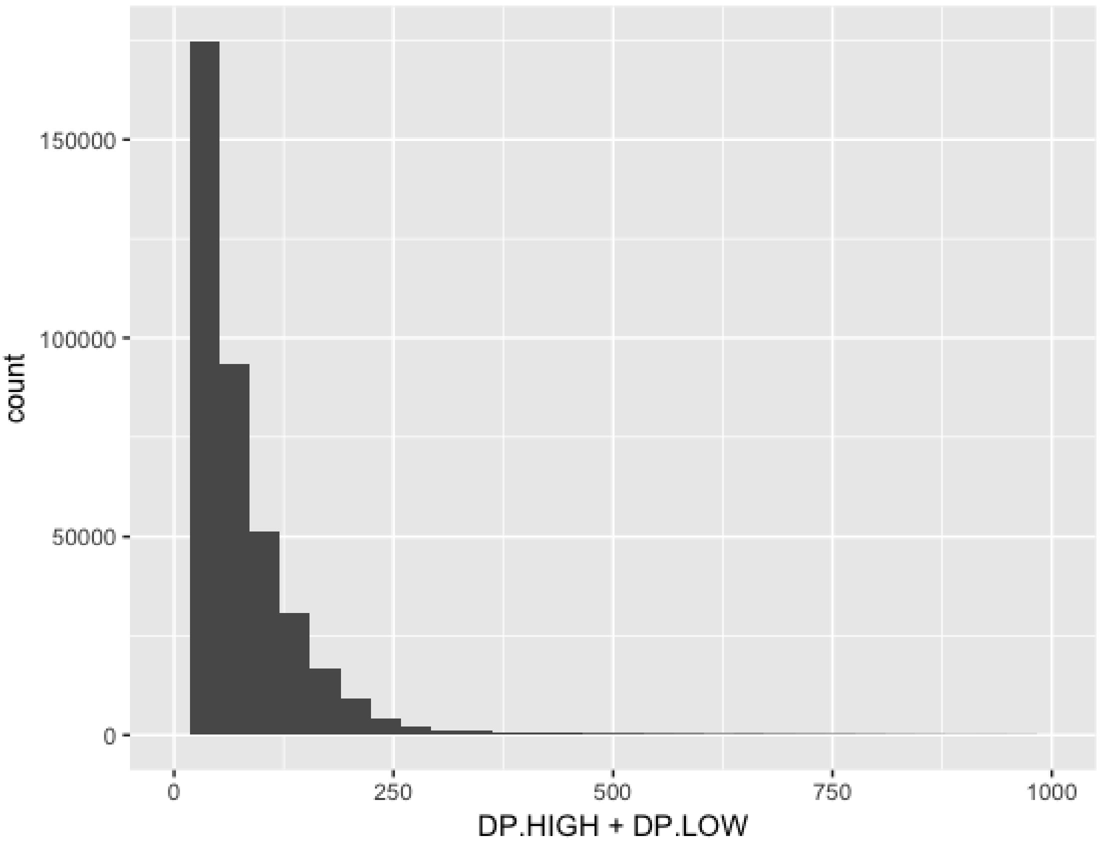

**Figure 5.1** The distribution of total read depths at different polymorphic positions in the input data. The X-axis represents the sum of read depth of the high and low bulks, and the Y-axis represents the number of variants

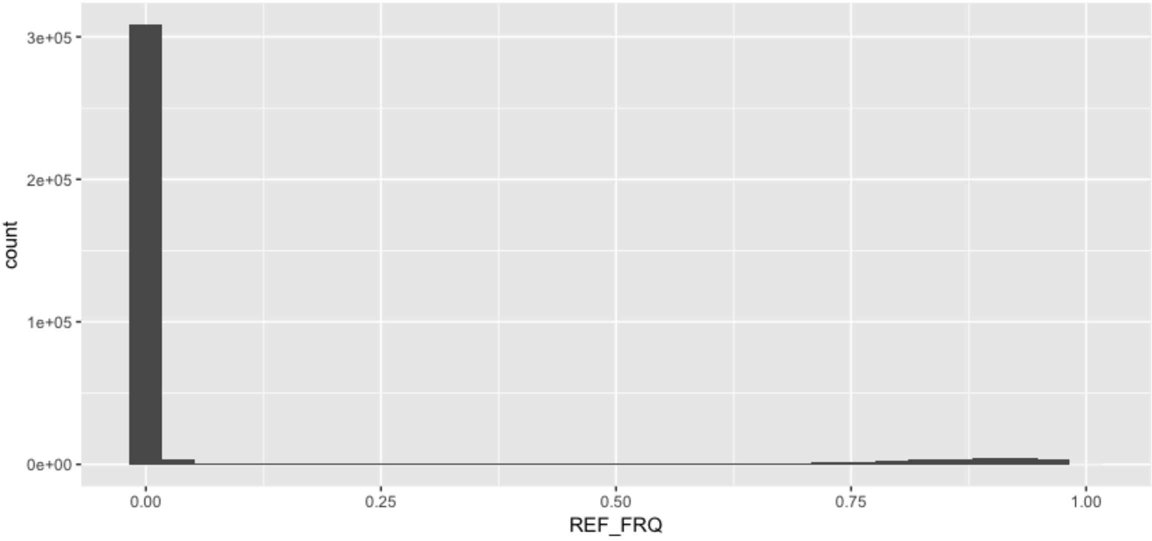

**Figure 5.2** Distribution of reference allele frequency (REF_FRQ) in the input variant data for the QTLseqr algorithm. The X-axis represents the reference allele frequency which is the ratio of number of reads with reference alleles to the total number of reads align

**Table 5.1** Number of reads generated for parents and the lines of the early- and late flowering bulks

| Sl. No | Sample Name | Parent/line constituting the Bulk (early or late) | Number of forward reads | Genome coverage (%) | Average depth of coverage |
|---|---|---|---|---|---|
| 1 | Royal | Parent | 22,719,458 | 81.97 | 10.94 |
| 2 | RE2 | Parent | 50,396,258 | 82.60 | 21.99 |
| 3 | Plant 6-E1 | Early | 15,285,185 | 80.98 | 07.59 |
| 4 | Plant 6-E2 | Early | 32,140,103 | 82.36 | 15.09 |
| 5 | Plant 6-E3 | Early | 38,409,712 | 82.46 | 17.76 |
| 6 | Plant 6-E4 | Early | 27,769,151 | 82.23 | 13.24 |
| 7 | Plant 6-E5 | Early | 22,332,380 | 81.93 | 10.40 |
| 8 | Plant 6-E6 | Early | 22,654,934 | 81.99 | 10.88 |
| 9 | Plant 6-E7 | Early | 25,377,600 | 82.30 | 12.80 |
| 10 | Plant 7-E1 | Early | 24,725,460 | 82.35 | 12.40 |
| 11 | Plant 7-E2 | Early | 19,309,978 | 81.89 | 09.79 |
| 12 | Plant 7-E3 | Early | 33,858,977 | 82.53 | 16.68 |
| 13 | Plant 7-E4 | Early | 34,816,234 | 82.57 | 17.05 |
| 14 | Plant 7-E5 | Early | 33,392,771 | 82.48 | 16.62 |
| 15 | Plant 7-E6 | Early | 26,195,457 | 82.38 | 13.10 |
| 16 | Plant 6-1L | Late | 22,887,089 | 82.21 | 11.44 |
| 17 | Plant 6-2L | Late | 20,323,812 | 82.02 | 10.14 |
| 18 | Plant 6-3L | Late | 29,735,797 | 82.42 | 14.93 |
| 19 | Plant 6-4L | Late | 31,356,835 | 82.53 | 16.15 |
| 20 | Plant 7-1L | Late | 27,514,137 | 82.46 | 14.36 |
| 21 | Plant 7-2L | Late | 26,607,022 | 82.44 | 14.00 |
| 22 | Plant 7-3L | Late | 25,045,358 | 82.38 | 13.18 |
| 23 | Plant 7-4L | Late | 26,119,848 | 82.43 | 13.74 |
| 24 | Plant 7-5L | Late | 24,001,370 | 82.35 | 12.46 |
| 25 | Plant 7-6L | Late | 25,551,569 | 82.40 | 13.51 |
| 26 | Plant 7-7L | Late | 26,431,070 | 82.40 | 13.98 |

The informative SNP are those that are able to differentiate the ‘Royal’ from the ‘RE2’ genotype and thus the polymorphisms between ‘Royal’ and ‘CDC Bethune’ are uninformative in this analysis. The ‘Royal’-‘CDC Bethune’ polymorphisms were filtered from the data using the reference allele frequency of 0.2 (0.2 < reference allele frequency < 0.8). As anticipated, the majority of the detected SNP distinguish ‘Royal’ and ‘CDC Bethune’ genotypes and filtering removed 575,046 (95%) loci from the total detected (Figure 5.3). Further filtering for retaining high confidence SNP loci using minimum coverage depth removed an additional 8,261 SNP. The total read depth in the filtered dataset ranged between 16 and 400 (Figure 5.4). Finally, after passing through the minimum genotyping quality filter, 11,385 SNP remained for QTL-seqR analysis at an average of ~750 per chromosome.

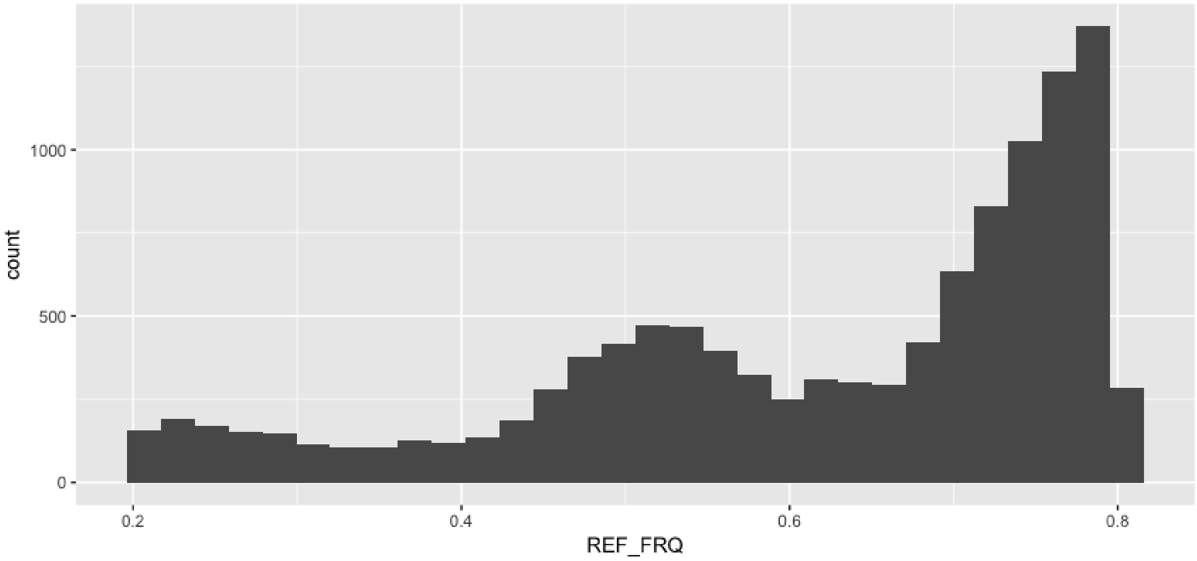

**Figure 5.3** Distribution of reference allele frequency (REF_FRQ) in the filtered variant data for the QTLseqr algorithm. The X-axis represents the reference allele frequency which is the ratio of number of reads with reference alleles to the total number of reads aligning to that position, and Y-axis represents the number of variants.

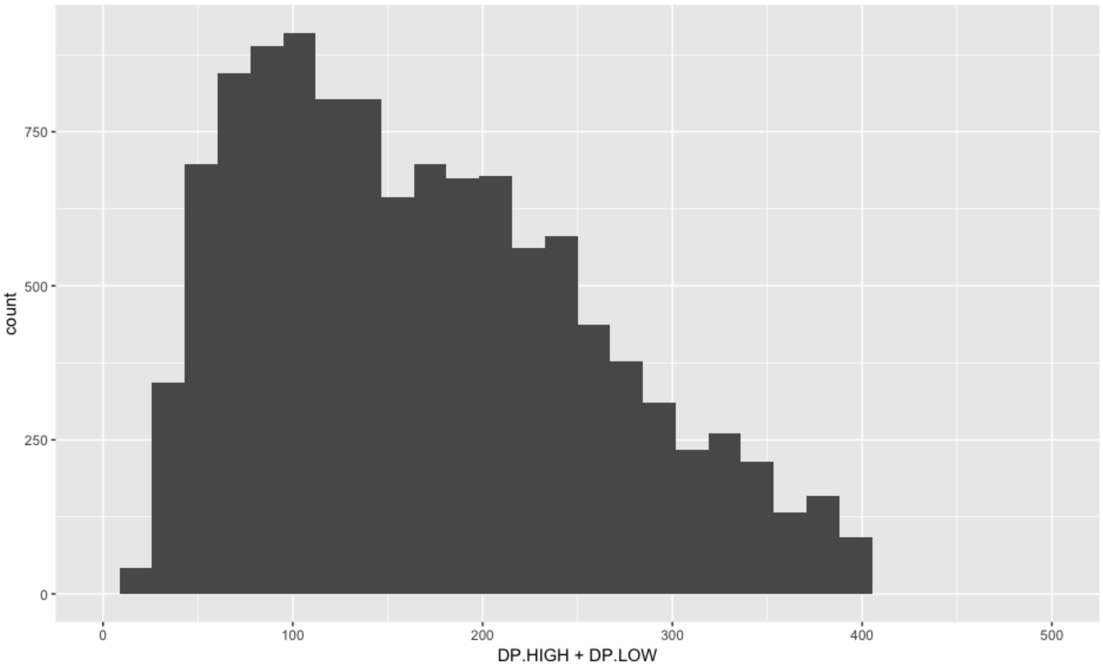

**Figure 5.4** Distribution of total read depth in the filtered data in which regions with sparse and highly dense coverage were removed. The X-axis represents the sum of read depth in the high and low bulks, and the Y-axis represents the number of variants.

### 5.4.3 Detection of candidate genomic region

After the estimation of the weighted average of ΔSNP-index values for the filtered SNP as described in the methodology, and confidence intervals in the QTL analysis, the ΔSNP-index plot was generated for regions across the 15 chromosomes of flax (Figure 5.5).

QTL analysis using QTLseqr identified five major peaks throughout the genome indicating the polygenic nature of flowering time in flax. Two of these five QTL satisfied statistical rigour at the 95% confidence level with three loci falling slightly under this threshold. Interestingly, the SNP at co-ordinate 7,455,755 on the chromosome 9, and the region between the variation at 9,793,543 and 9,798,720 on the chromosome 12 (spanning 5,177 bp), with a tricube smoothed ΔSNP-index value above the 95% confidence interval threshold were found to have significant association with the phenotype.

Functional annotation of the SNP variants using the SnpEff algorithm suggested that the SNP on chromosome 9 was present in the intergenic region between the flax genes *Lus10024495* and *Lus100024494*. The homologue of *Lus10024495* in Arabidopsis belongs to Cytochrome P450 gene family, whereas that of *Lus100024494* is annotated as a phloem protein. Similarly, with reference to the nucleotide variation identified on the chromosome 12, the positions

9,793,543 and 9,798,720 were localized to upstream and downstream, of the flax gene *Lus10024264*, respectively. The corresponding Arabidopsis homologue of *Lus10024264* is annotated as LATE EMBRYOGENESIS ABUNDANT HYDROXYPROLINE-RICH GLYCOPROTEIN family.

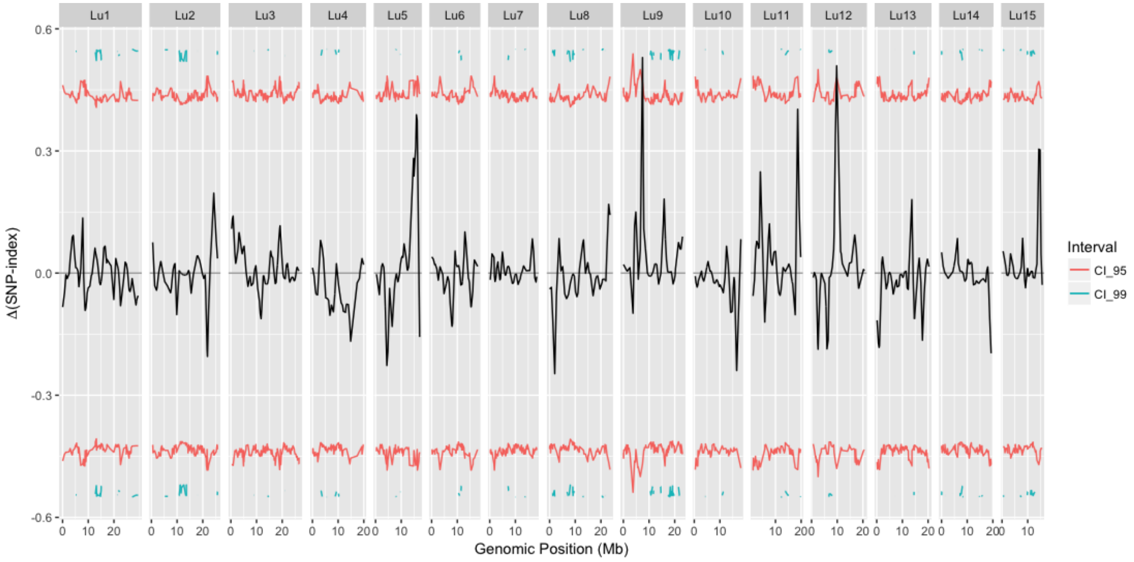

**Figure 5.5** The ΔSNP-index plot generated with a moving bin window of 1 Mb. The red line depicts the confidence interval at 95%. The genomic position in Mb is depicted by the X-axis, and Lu1 to Lu15 represent the linkage groups. The Y-axis represets the ΔSNP-index values.

In addition, the interval spanning the SNP with a ΔSNP-index value of one and the significant region (peak) identified by the QTLseqr algorithm was also characterized. Between the coordinates 7,455,755 and 7,956,993 on chromosome 9, a total of 29 genes were present. In the chromosome 12, the interval between the loci 9,305,376 and 9,798,720 encompassed 41 genes. The results of homology search employing BLASTx for all 70 genes are listed in the Table 5.2. Among the 70 genes, 60 had homologues of which 20 genes were homologous to uncharacterized or hypothetical proteins. Several abiotic stress response related proteins including DEHYDRATION RESPONSIVE ELEMENT-BINDING PROTEIN, MAINTENANCE OF MERISTEMS-LIKE, ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR-LIKE and HEAT STRESS TRANSCRIPTION FACTORS were encoded by genes in the region. Majority of the genes exhibited homology to those in *Populus trichocarpa* (Torr. & Gray), *Ricinus communis* L. and *Jatropha curcas* L.

**Table 5.2** Homology of 70 genes present in the regions of interest on flax chromosomes 9 and 12

| Gene | Start position | End position | Length (bp) | Strand | Identity (%) | Homologous gene | E value |
|---|---|---|---|---|---|---|---|
| | | | | | | **Genes on Lu9** | |
| Lus10024495 | 7447896 | 7450216 | 2321 | - | 85 | Cytochrome P450 87A3 in *Populus trichocarpa* | 3e$^{-99}$ |
| Lus10024485 | 7598742 | 7602363 | 3622 | - | 87 | Predicted: homeobox protein knotted-1 like 7 of *Cucumis melo* | 5e$^{-71}$ |
| Lus10024486 | 7591538 | 7593043 | 1506 | + | 57 | UDP-glycosyltransferase 1 of *Linum usitatissimum* | 0 |
| Lus10024487 | 7551935 | 7556451 | 4517 | + | 74 | Hypothetical protein CISIN 1g0081631mg in *Citrus sinensis* | 5e$^{-90}$ |
| Lus10024488 | 7547478 | 7550243 | 2766 | + | 36 | Uncharacterized protein LOC105645885 in *Jatropha curcas* | 9e$^{-21}$ |
| Lus10024489 | 7535253 | 7536890 | 1638 | - | NA | NA | NA |
| Lus10024490 | 7523972 | 7524339 | 368 | + | NA | NA | NA |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10024491 | 7518840 | 7519514 | 675 | - | 56 | Dehydration-responsive element-binding protein 1A in *Jatropha curcas* | $2e^{-71}$ |
| Lus10024492 | 7497993 | 7498670 | 678 | + | 51 | Predicted: ethylene-responsive transcription factor ERF-027-like in *Gossypium hirsutum* | $2e^{-55}$ |
| Lus10024493 | 7482933 | 7485112 | 2180 | - | 35 | Predicted: ribosomal RNA processing protein 36 homolog isoform X2 in *Glycine max* | $3e^{-19}$ |
| Lus10024494 | 7480914 | 7482110 | 1197 | + | 52 | Phloem protein 2-A12 in *Arabidopsis thaliana* | $3e^{-118}$ |
| Lus10001191 | 7660633 | 7661616 | 984 | + | 53 | Uncharacterized protein LOC110631412 in *Manihot esculenta* | $4e^{-97}$ |
| Lus10001192 | 7666667 | 7669970 | 3304 | + | 64 | FIZZY-RELATED 2-like protein in *Manihot esculenta* | $3e^{-90}$ |
| Lus10001193 | 7674851 | 7675581 | 731 | + | 35 | PKS-NRPS hybrid synthetase CHGG 01239-like in *Chenopodium quinoa* | $2e^{-54}$ |
| Lus10001194 | 7678782 | 7679993 | 1212 | - | 82 | Predicted: protein STAY-GREEN, chloroplastic-like isoform X1 in *Glycine max* | $4e^{-68}$ |
| Lus10001195 | 7688921 | 7689492 | 572 | + | 72 | Homeobox protein HD1 of *Vigna radiata* var *radiata* | $3e^{-43}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10002989 | 7864564 | 7866433 | 1870 | - | 75 | Conserved hypothetical protein of *Ricinus communis* | $2e^{-121}$ |
| Lus10002990 | 7859330 | 7863009 | 3680 | + | 53 | Pentatricopeptide repeat-containing protein AT1G10270 of *Manihot esculenta* | 0 |
| Lus10002991 | 7855373 | 7858047 | 2675 | + | 49 | Uncharacterized protein LOC110624913 isoform X3 of *Manihot esculenta* | $4e^{-75}$ |
| Lus10002992 | 7853002 | 7853739 | 738 | + | 73 | Triphosphate tunel metalloenzyme 3 *Jatropha curcas* | $4e^{-104}$ |
| Lus10002993 | 7814575 | 7816536 | 1962 | + | 72 | Protein kinase and PP2C-like domain-containing protein isoform X1 *Manihot esculenta* | 0 |
| Lus10002994 | 7798482 | 7799380 | 899 | - | 65 | Putative methyltransferase DDB G0268948 in *Populus trichocarpa* | $2e^{-101}$ |
| Lus10002995 | 7793794 | 7796149 | 2356 | - | 67 | Subtilisin-like protease SBT1.2 of *Jatropha curcas* | 0 |
| Lus10002996 | 7784972 | 7788671 | 3700 | - | 71 | Calcium permeable stress-gated cation channel 1-like in *Manihot esculenta* | $5e^{-123}$ |
| Lus10002997 | 7778495 | 7779223 | 729 | - | 76 | Expansin-A7 of *Jatropha curcas* | $4e^{-116}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10002998 | 7772561 | 7773583 | 1023 | - | 51 | Hypothetical protein CUMW 074040 in *Citrus unshiu* | 5e$^{-86}$ |
| Lus10002999 | 7756204 | 7756766 | 563 | + | NA | NA | NA |
| Lus10007433 | 7926180 | 7929557 | 3378 | + | 73 | Predicted: chloride channel protein CLC-c in *Lupinus angustifolius* | 0 |
| Lus10007434 | 7937900 | 7945160 | 7261 | + | 77 | Hypothetical protein GOBAR DD20035 of *Gossypium barbadense* | 3e$^{-67}$ |
| Lus10007435 | 7951900 | 7964803 | 12904 | + | 32 | Predicted: protein FAR-RED ELONGATED HYPOCOTYL 3-like of *Beta vulgaris* subsp. *vulgaris* | 4e$^{-38}$ |
| **Genes on Lu12** | | | | | | | |
| Lus10034901 | 9720433 | 9720711 | 279 | + | 57 | Predicted: uncharacterized protein LOC107260870 in *Ricinus communis* | 1e$^{-17}$ |
| Lus10034902 | 9714256 | 9715008 | 753 | + | 45 | Dehydration-responsive element-binding protein 2C-like in *Manihot esculenta* | 9e$^{-75}$ |
| Lus10034903 | 9712026 | 9713286 | 1261 | - | 53 | Hypothetical protein GLYMA 08G038900 in *Glycine max* | 2e$^{-45}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10034904 | 9703666 | 9704622 | 957 | + | 47 | Uncharacterized protein LOC110610983 in *Manihot esculenta* | 6e$^{-56}$ |
| Lus10034905 | 9699655 | 9702907 | 3253 | + | 46 | Hypothetical protein POPTR 018G079500v3 in *Populus trichocarpa* | 2e$^{-77}$ |
| Lus10034906 | 9693141 | 9696510 | 3370 | + | NA | NA | NA |
| Lus10034907 | 9684716 | 9685728 | 1013 | - | 43 | Heat stress transcription factor C-1 in *Populus trichocarpa* | 3e$^{-72}$ |
| Lus10034908 | 9650449 | 9652354 | 1906 | - | 65 | Predicted: indole-3-pyruvate monooxygenase YUCCA6-like in *Gossypium hirsutum* | 1e$^{-34}$ |
| Lus10034909 | 9635746 | 9637186 | 1441 | - | 91 | 40s ribosomal protein s7-1-like in *Trifolium pratense* | 5e$^{-38}$ |
| Lus10034910 | 9634226 | 9634768 | 543 | + | 42 | Uncharacterized protein LOC111461329 in *Cucurbita moschata* | 4e$^{-135}$ |
| Lus10034911 | 9600632 | 9601177 | 546 | + | 30 | Uncharacterized protein LOC110713760 in *Chenopidium quinoa* | 1e$^{-11}$ |
| Lus10034912 | 9596366 | 9600031 | 3666 | + | 65 | Uncharacterized protein LOC110611313 in *Manihot esculenta* | 6e$^{-167}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10034913 | 9591695 | 9594503 | 2809 | + | 74 | Uncharacterized protein LOC105637720 isoform X2 in *Jatropha curcas* | $1e^{-125}$ |
| Lus10034914 | 9589782 | 9590210 | 429 | - | NA | NA | NA |
| Lus10034915 | 9581960 | 9583512 | 1553 | + | 51 | Predicted: phosphatidylinositol 4-kinase alpha 1 in *Ricinus communis* | $1e^{-34}$ |
| Lus10034916 | 9574515 | 9575112 | 598 | - | NA | NA | NA |
| Lus10034917 | 9568151 | 9569825 | 1675 | - | 62 | Predicted: Chlorophyll (ide) b reductase NYC1, chloroplastic *Cicer arietinum* | $7e^{-117}$ |
| Lus10034918 | 9552111 | 9554512 | 2402 | + | 79 | Subtilisin-like protease SBT6.1 isoform X2 in *Jatropha curcas* | 0 |
| Lus10034919 | 9544318 | 9546693 | 2376 | + | 53 | Subtilisin-like protease SBT6.1 in *Manihot esculenta* | $2e^{-46}$ |
| Lus10034920 | 9541889 | 9542708 | 820 | - | 42 | Protein MAINTENANCE OF MERISTEMS-like in *Chenopodium quinoa* | $3e^{-18}$ |
| Lus10034921 | 9523384 | 9526752 | 3369 | - | 47 | Predicted: uncharacterized protein LOC8268361 isoform X1 in *Ricinus communis* | $1e^{-173}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10034922 | 9521400 | 9522993 | 1594 | + | 78 | Hypothetical protein Mpv17 in *Manihot esculenta* | $4e^{-71}$ |
| Lus10034923 | 9515666 | 9516208 | 543 | + | NA | NA | NA |
| Lus10034924 | 9502697 | 9504544 | 1848 | - | 59 | Pentatricopeptide repeat-containing protein AT5G21222 in *Jatropha curcas* | 0 |
| Lus10034925 | 9499966 | 9501632 | 1667 | - | 42 | Peptidase S26A, signal peptidase I in *Corchorus capsularis* | $6e^{-46}$ |
| Lus10034926 | 9498165 | 9499826 | 1662 | + | 54 | Hypothetical protein L484 001883 in *Morus notabilis* | $5e^{-60}$ |
| Lus10034927 | 9481012 | 9483443 | 2432 | - | 76 | Rop guanine nucleotide exchange factor 12 in *Solanum tuberosum* | $7e^{-116}$ |
| Lus10034928 | 9459570 | 9468716 | 9147 | + | 47 | Retrovirus related Pol polyprotein from transposon TNT 1-94 in *Morus notabilis* | 0 |
| Lus10034929 | 9449509 | 9449841 | 333 | + | 41 | Uncharacterized protein LOC9305531 isoform X2 in *Arabidopsis lyrate* subsp. *lyrata* | $2e^{-10}$ |
| Lus10034930 | 9436937 | 9440846 | 3910 | - | 44 | Predicted: probable plastidic glucose transporter 3 isoform X2 in *Vitis vinifera* | $2e^{-30}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10034931 | 9434794 | 9436104 | 1311 | + | 55 | Cyclin-dependent kinase F-1 in *Populus trichocharpa* | $5e^{-110}$ |
| Lus10034932 | 9428737 | 9430103 | 1367 | + | 100 | Rac-like GTP-binding protein ARAC7 in *Jatropha curcas* | $3e^{-44}$ |
| Lus10034933 | 9427240 | 9428159 | 920 | - | 51 | Vacuolar protein-8 like in *Manihot esculenta* | $6e^{-35}$ |
| Lus10034934 | 9386833 | 9390209 | 3377 | + | NA | NA | NA |
| Lus10034935 | 9365290 | 9365601 | 312 | - | 62 | Predicted: IRK-interacting protein in *Lupinus angustifolius* | $4e^{-16}$ |
| Lus10034936 | 9363389 | 9364160 | 772 | + | 63 | Hypothetical protein POPTR 012G087500v3 in *Populus trichocarpa* | $6e^{-61}$ |
| Lus10034937 | 9345218 | 9352134 | 6917 | + | 66 | Putative E3 ubiquitin-protein ligase LIN isoform X1 in *Jatropha curcas* | 0 |
| Lus10034938 | 9321624 | 9322265 | 642 | + | NA | NA | NA |
| Lus10024264 | 9796674 | 9797375 | 702 | + | 69 | Uncharacterized protein LOC105637702 in *Jatropha curcas* | $6e^{-81}$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lus10024265 | 9791500 | 9792324 | 825 | + | 42 | Protein MAINTENANCE OF MERISTEMS-like in *Spinacia oleracea* | $1e^{-23}$ |
| Lus10024266 | 9776048 | 9776717 | 670 | - | NA | NA | NA |

*NA - Not Available (No significant similarity found)*

## 5.5 Discussion

The original QTL-seq pipeline was developed by Takagi et al. (2013) for the identification of genomic region underlying the polygenic traits segregating in a mapping population. The pipeline takes sequencing-reads from bulks representing distributional phenotypic extremes in the population and that of one of the parents as the initial input and carries out the following steps: sequencing-read alignment, variant calling and statistical analysis to identify genotype-phenotype association. However, the pipeline possesses a few limitations such as the low ease of configuration at the variant calling step. Additionally, the pipeline employs legacy versions of the different tools for the analysis and hence has limited adaptability for the integration of upgraded tools and software. Whereas, the recent QTLseqr package (Mansfeld and Grumet 2018), developed based on the open source statistical computing platform-R (R core team 2018) with several biostatistics and computational biology packages, is more convenient for the next generation sequencing (NGS) based bulked segregant analysis (BSA), with provisions for user defined alignment and variant calling steps.

Burrows-Wheeler Aligner (BWA; Li and Durbin 2009) is the alignment tool used in the original QTL-seq pipeline (Takagi et al. 2013). However, plant genomes known for their complexity due to the presence of evolutionary genome duplication events, repetitive sequences, pseudogenes, paralogous genomic region, insertions and deletions-InDels (Schatz et al. 2012), are handled by Bowtie 2 with better efficiency (Langmead and Salzberg 2012; Scheben et al. 2017) and hence, was employed as an aligner in the analysis using QTLseqr. For filtering misaligned reads based on number of mismatches, and variant calling, the QTL-seq pipeline utilizes Coval (Kosugi et al. 2013). However, for the QTLseqr analysis, custom PERL scripts were used to pick the best hits from the alignment files based on alignment scores, thus reducing the level of heterozygosity across many loci. Further, the improved algorithm Genome Analysis Tool Kit (GATK) HaplotypeCaller (De Pristo et al. 2011), capable of handling pooled samples, was used to generate the Variant Call File (vcf ) file containing the SNP information for downstream analysis using QTLseqr (Mansfeld and Grumet 2018). Moreover, GATK HaplotypeCaller identifies the genomic region with variants and carries out local realignment improving the accuracy of variant identification.

In the QTL-seq pipeline, the minimum and maximum read depth for the samples is modifiable for filtering. However, filtering in QTLseqr employs additional parameters of minimum- and maximum total depth of coverage from both samples at each locus. Unlike pre-set parameters used for SNP calling in the QTL-seq pipeline, in QTLseqr, filtering is carried out on called variants making it more efficient.

The minimum required sequence depth in sample was set as seven to differentiate the true variants from false positives in the original QTL-seq study in rice (*Oryza sativa* L.; Takagi et al. 2013), chickpea (*Cicer arietinum* L.; Singh et al. 2016) and groundnut (*Arachis hypogaea* L.; Pandey et al. 2017). However, in QTLseqr, an additional parameter namely reference allele frequency (RAF) was used for removal of alleles which are minimally represented in a given locus and might arise as sequence or alignment errors. Reference allele frequency is defined as the number of reads containing the reference allele at the given position to the total number of reads aligning to that position from both the bulks. Hence, removing the variants with RAF < 0.2 filtered those variants represented mostly by the alternate allele in the consensus. In this study, since the alignments were made to the reference of ‘CDC Bethune’ and not to a secondary reference generated from one of the parents (‘Royal’ or ‘RE2’), these filtered variants (RAF < 0.2) are potentially representing the difference between ‘CDC Bethune’ and ‘Royal’ as they are common among all the sequenced individuals. As discussed earlier, by removing variants with a RAF > 0.8, poorly represented alleles were excluded from further analysis. Removal of low quality and uninformative loci aids in downstream data analysis along with improving statistical robustness.

SNP-index is calculated as the ratio of sequence depth of allelic variant (alternate allele) compared to the total read depth at a given position (Abe et al. 2012). ΔSNP-index is estimated as the difference between the SNP-indices of *high* and *low* bulks. However, in QTLseqr, instead of a simple ΔSNP-index, a modified statistic called tricube smoothed ΔSNP-index was estimated, wherein weightage was assigned for closeness of a SNP in linkage disequilibrium (LD) to the focal SNP within the sliding window. In other words, QTLseqr gives weightage for SNP in LD, since genomic recombination is a non-random event controlled by distribution of specific sites called hotspots (de Massy 2013; Choi and Henderson 2015). Setting up of statistical thresholds for parameters has impact on discovering genotype-phenotype associations.

In QTLseqr, association of specific genomic region with the phenotype is considered significant if the tricube ΔSNP-index is higher than the threshold which is determined by tens of thousands of bootstrapped simulations for read depths varying from one to 50. In order to avoid spurious associations, higher threshold values were assigned to the following parameters: minimum sample depth=7; minimum total depth=14; maximum total depth=400; RAF=0.2. However, under these stringent filtering criteria, certain real associations would have been missed.

QTLseqr identified two significant regions associated with flowering time on the chromosomes 12 and 9. Homology search (BLASTx - Altschul et al. 1997) was carried out for the genes present in this region. On chromosome 12, the significant region was delimited by two variants that spanned 5,177 bp (coordinates: 9793543 – 9798720), in addition, the region extending to the closest variant with a ΔSNP-index of 1 (coordinate: 9305376), encompassed 41 genes (Table 5.2). Based on the annotation using SnpEff, the two variants delimiting the significant region were located downstream of *Lus10024264* and *Lus10024265*. The flax gene *Lus10024264* was homologous to a gene encoding an uncharacterized *Jatropha curcas* L. protein (percent identity=69; E-value=6e$^{-81}$; the best hit) was homologous to *AT4G13270* in Arabidopsis (percent identity=52; E-value=4e$^{-68}$) which belongs to LATE EMBRYOGENESIS ABUNDANT (LEA) HYDROXYPROLINE-RICH GLYCOPROTEIN family and plays an important role in drought tolerance (Magwanga et al. 2018). The flax gene *Lus10024265* was homologous to *MAINTENANCE OF MERISTEMS-LIKE* (*MAIL*) in *Spinacia oleracea* L. (percent identity=42; E-value=1e$^{-23}$). Based on studies in Arabidopsis, *MAINTENANCE OF MERISTEMS* (*MAIN*) is involved in sustaining the meristem stability and retention of genome integrity (Wenig et al. 2013). The Arabidopsis gene *MAIL1* also has similar function, besides its involvement in cell differentiation (Ühlken et al. 2014).

On chromosome 9, the region flanked by the significant variant (coordinate: 7,455,755) and the site having a ΔSNP-index one (coordinate: 7,956,993), was found to harbour 29 genes (Table 5.2). SnpEff annotation indicated the presence of variation with significant association with flowering time in the intergenic region between *Lus10024495* and *Lus10024494*, which are homologous to *CYTOCHROME P450 87A3* in *Populus trichocarpa* (Torr. & Gray) (percent identity=85; E-value=3e$^{-99}$) and *PHLOEM PROTEIN 2-A12* in Arabidopsis (percent identity=52; E-value=3e$^{-118}$), respectively. The CYTOCHROME P450 (CYP) superfamily of proteins is a large group involved in diverse growth and developmental activities in plants, especially in synthesis of secondary metabolites. The CYP 87 belongs to the CYP 85 clan contributing to brassinosteroid and gibberelic acid biosynthesis (Nelson et al. 2004; Jun et al. 2015). Phloem protein 2, an abundant group of proteins distributed across plant species, suggested to have diversified function associated with different domains acquired over evolutionary time (Dinant et al. 2003). Specifically, Phloem protein 2-A12 belongs to *Nicotiana tabacum* L. agglutinin (Nictaba) family and contains F-box domain majorly associated with mechanisms related to abiotic stress response (Eggermont et al. 2017).

The additional 41 and 29 genes on both chromosomes 12 and 9 were functionally annotated using sequence alignment BLASTx (Altschul et al. 1997). One gene each on chromosome 9 (*Lus10024491*) and chromosome 12 (*Lus10034902*) were found to be homologous to DEHYDRATION-RESPONSIVE ELEMENT BINDING (DREB) protein family. *DEHYDRATION-RESPONSIVE ELEMENT BINDING* transcription factors belong to the *APETALA2/ETHYLENE RESPONSIVE TRANSCRIPTION FACTOR* (*AP2/ERF*) family and are involved in abiotic stress response in plants (Lata and Prasad 2011). Interestingly, mutant of *DWARF AND DELAYED FLOWERING 1*, related to *DREB* type of *AP2/ERF* family transcription factors is found to influence flowering time and reveal phenotype as in mutants lacking gibberellic acid, in Arabidopsis (Magome et al. 2004). Additionally, *CYCLING DOF FACTOR 3* (*CDF3*), a key gene of the photoperiodic flowering pathway is found to play a role as the master regulator of transcription factors including *DREB2A* in Arabidopsis, and consequently involved in tolerance to drought, salinity and temperature extremities (Corrales et al. 2017). The flax gene *Lus10024492* on chromosome 9 was homologous to *ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR (ERF)-027-LIKE* in *Gossypium hirsutum* L., belonging to the *AP2* superfamily of transcription factors. The latter involved in both biotic and abiotic stress response has been reported in barley (*Hordeum vulgare* L.; Guo et al. 2016), cotton (*Gossypium hirsutum* L.; Jin et al. 2010), cauliflower (*Brassica oleracea* L. var. *botrytis*; Li et al. 2017b) and sunflower (*Helianthus annuus* L.; Najafi et al. 2018). In addition, varied expression levels of different *ERFs* like *ERF96* (Wang et al. 2015) and *ERF019* (Scarpeci et al. 2016) are also suggested to alter flowering time in the model plant Arabidopsis. The other analysed flax genes exhibit higher degree of homology to poplar (*Populus trichocarpa* (Torr. & Gray)) and castor (*Ricinus communis* L.) because of the lesser evolutionary divergence than the model organisms as reported earlier (Venglat et al. 2011). Nearly 20 genes among the total 70 genes were homologous to uncharacterized or hypothetical proteins and hence, their function could not be determined. A total of ten flax genes (*Lus10024489*, *Lus10024490*, *Lus10002999, Lus10034906*, *Lus10034914*, *Lus10034916*, *Lus10034923*, *Lus10034934*, *Lus10034938*, *Lus10024266*) present in this region did not have orthologues identified, and these unique flax specific genes might be involved in flowering time regulation and need further investigation.

In addition, the QTLseqr analysis detected three other peaks on chromosomes 5, 11 and 15, although below the significant threshold. It could be possible that if more individuals were added to bulks then these peaks would be detected with greater significance. The peak observed on chromosome 15 did not overlap with *Lus10040921*, a homologue of Arabidopsis *LUMINIDEPENDENS* (*LD*), identified in the QTL-seq pipeline (Chapter 4). However, among the total 178 genes underlying these three peaks, *Lus10024163* on chromosome 5 was the homologue of Arabidopsis *SENSITIVITY TO RED LIGHT REDUCED 1* (*SRR1*), which is involved in the circadian clock pathway and regulates multiple *FT* repressors and plays a role in photoperiod- dependent and independent flowering (Johansson and Staiger 2014).

### 5.6 Conclusion

In order to identify the genomic region associated with early flowering phenotype, a segregating RIL population from 'Royal' x 'RE2' was used (Chapter 3). The heritability of the early flowering trait through at least 9 generations suggested that this trait was under genetic control. Sequencing of distributional extremes for the flowering time phenotype and analysis using QTL-seq pipeline followed by characterization of SNP identified a polymorphism upstream to a gene involved in flowering (*LD*), based on homology analysis (Chapter 4). However, investigation using an improved QTL analysis pipeline (QTLseqr: Mansfeld and Grumet 2018) detected two significant regions associated with the early flowering phenotype, one on chromosome 12 and another on chromosome 9. Additional minor QTL might also be present on chromosomes 5, 11 and 15 suggesting that this trait is controlled by five QTL. These data are in agreement with the original estimate made by Fieldes and Amyot (1999). In total, nearly 70 genes were investigated and majority of them were annotated to be associated with biotic and abiotic stress responses in other plants and an additional few uncharacterized genes that appear to be unique to flax. It is perhaps plausible that increased expression of stress related genes might promote early flowering. Expression data need to be generated to confirm if indeed these genes are in fact dysregulated. However, considering that the method of generation of the original mutant ‘RE2’ was exposure to 5-Azacytidine, and the inability to detect a strong mutation in genes associated with flowering time, the question as to the contribution of methylation variation remains open. DNA methylation variation might also contribute to dysregulation of gene expression. While the association of these genomic regions with the early flowering trait defines the causative regions, the molecular mechanisms controlling the early flowering trait in ‘RE2’ are more complex than originally anticipated.

# Chapter 6 Identification of Differentially Methylated Regions Using Whole Genome Bisulfite Sequencing Variation in DNA Methylation Associated with the Early Flowering Phenotype

## 6.1 Abstract

The early flowering mutant ‘RE2’ was obtained by treatment of cultivar ‘Royal’ with the hypomethylating chemical 5-Azacytidine (5-AzaC). The method of generation suggests the potential epigenetic basis of the trait. Hence, a subset of early- and late flowering bulks, derived from the segregants of the ‘Royal’ x ‘RE2’ recombinant inbred population, in addition to the parents, were subjected to whole genome bisulfite sequencing for inferring methylation patterns. Visualization of DNA methylation in all three sequence contexts (CG, CHG and CHH) at a global level suggested no genome-wide significant differences. However, investigation of specific chromosome bins identified significant Differentially Methylated Regions (DMRs). A total of 494,263 cytosines exhibited differential methylation patterns between the bulks, and 127 significant differentially methylated regions were distributed in the genic, upstream and intergenic regions. Homology search of the genes overlapping with DMRs as well as those located downstream to DMRs identified three genes homologous to Arabidopsis FASCILIN-LIKE ARABINOGALACTAN group, involved in biomechanics of stem development. A cluster of significant DMRs were localized on chromosome 12, the same linkage group on which candidate regions were identified using QTLseqr. Two significant DMRs were present upstream to flax genes *Lus10036234* and *Lus10015319* encoding proteins homologous to SUPPRESSOR OF FRI4 and FRIGIDA ESSENTIAL 1 of Arabidopsis, respectively, with a role in vernalization pathway.

## 6.2 Introduction

Flowering time is an important trait in crop breeding because of its association with adaptation (Sasaki et al. 2017) and early maturity (Kong et al. 2018). Several studies exploring the association between nucleotide variation and prime agronomic traits including flowering time are prevalent, and they have been adapted in crop improvement through the use of polymorphic DNA markers, linkage mapping and quantitative trait locus (QTL) analysis that together allow the application of marker aided breeding strategies (Bevan et al. 2017). Characterization of phenotypic variation controlled by underlying epigenetic differences is less well established.

Epigenetic information including DNA methylation changes, histone modification and non-coding RNA (siRNA) is emerging as topic of great interest where, DNA methylation has been studied in the greatest detail in plants (Seymour and Becker 2017). However, the utility of epigenetic variation as a source of useful phenotypic variability for crop breeding remains an open question with suggestion that it has not been exploited to its fullest potential (King et al. 2010).

The generation of reference quality genome sequences has opened the door to post-genome analyses that examine epigenetic variation. These assays measure the structure of chromatin and examine the potential of chromatin variation to influence gene expression, DNA replication and its effects on chromosome pairing and segregation. There are numerous reports describing the prevalence and extent of DNA methylation in plant genomes where positional information is obtained through alignment of bisulfite converted sequencing reads. Often common patterns are observed among these genomes (Takuno et al. 2016). In plants, DNA methylation occurs in three sequence contexts CG, CHG and CHH (where H refers to any nucleotide other than guanine). In all cases, the highest level of methylation occurs at CG positions with CHG sites being the next most abundant and a low level of methylation observed at CHH positions. The context of the cytosine residue indicates the underlying biochemistry responsible for transferring the information. Largely, at the symmetrical positions (CG and CHG) methylation status is either maintained through replication by enzymes encoded by the *METHYLTRANSFERASE 1* (*MET1*; Finnegan et al.1996) and *CHROMOMETHYLASE 3* (*CMT3*; Lindroth et al.2001) gene families. Methylation at the non-symmetrical positions (CHH) occurs through the action of the *DOMAINS REARRANGED METHLYLASE 2* (*DRM2*) gene family that is guided to the correct location by non-coding RNA molecules as part of the RNA dependent DNA methylation pathway (reviewed in Zhang and Zhu 2011).

The function of DNA methylation is unclear, but DNA methylation is highly abundant at repetitive regions of the genome, likely protecting genome integrity by acting to silence the movement of transposable elements (Feng and Jacobsen 2011). However, DNA methylation is also present, albeit at a lower level throughout gene body sequences whose function remains to be demonstrated (Bewick and Schmitz 2017). It is now accepted that DNA methylation patterns are faithfully transmitted through meiosis (Niederhuth and Schmitz 2014), a necessary process if it functions in transposon silencing. However, there is some variation with populations suggesting that the level of methylation over a short range is important rather than its exact location at a base-pair resolution. Variation at the level of DNA methylation has been observed in natural plant population perhaps as a source of geographical adaptation, opening the possibility for the generation of epialleles (Schmitz et al. 2013; Kawakatsu et al. 2016a). A number of differentially methylated regions (DMRs) between accessions are capable of modifying the phenotype even though the individual DMRs do not have a significant impact suggesting potential interaction among DMRs (Springer and Schmitz 2017). While DNA methylation variation is widely detected, the definition of an epiallele is stringent. However, several have been reported where they are the cause of observed phenotypes, some having agronomic value (Weigel and Colot 2012). For instance, the epigenetic modification through the methylation of specific cytosines in the upstream region of *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* gene was associated with delayed fruit ripening in tomato (*Solanum lycopersicum* L.; Manning et al. 2006). In oil palm (*Elaeis guineensis* Jacq.) trees regenerated from tissue culture, the hypomethylation of Karma, a Long Interspersed Nuclear Element (LINE) retrotransposon, in the intronic region of *DEFECIENS*, resulted in modified transcript due to abnormal splicing and consequent production of mantled fruits (Ong-Abdullah et al. 2015). These examples build on initial demonstration of epigenetic mechanisms controlling flowering time variation in Arabidopsis although both monocots and dicots use epigenetic mechanism for the regulation of flowering time (Dennis and Peacock 2007). An early study in Arabidopsis reported the atypical expression of the *FLOWERING WAGENINGEN (FWA)* gene caused by hypomethylation resulting in late flowering (Soppe et al. 2000). Although the gene silencing of the *FWA* locus occurs by DNA methylation of its promoter, the allele mechanism is mediated by non-coding RNA generated by the duplication of a repeat sequence in the *FWA* allele (Lippman et al. 2004; Chan et al. 2004). This provides an example of the complex interaction altering chromatin, using both epigenetic and genetic control. Perhaps the best characterized phenomenon under epigenetic regulation is vernalization, a key pathway regulating flowering time. Vernalization uses histone modification to regulate gene expression in a process that is heritable through mitosis but is reset during meiosis. Gene silencing is established by the binding of Polycomb repressive complex 2 (PRC2) which mediates trimethylation of nucleosomes at the Histone 3 Lysine 27 (H3K27me3) mark. Critically, PRC2 binding leads to the repression of

*FLOWERING LOCUS C* (*FLC*) acting to repress *FLOWERING LOCUS T* (*FT*) expression (Berry and Dean 2015; Hepworth and Dean 2015; Bouché et al. 2017). Silencing of the floral repressor *FLC* relives the *FT* inhibition, subsequently promoting floral meristem identity change. To re-establish floral repression in the following generation, preventing flowering from occurring during the winter, the H3K27me3 marks are removed during embryogenesis (reviewed in He and Li 2018). This occurs when *LEAFY COTYLEDON 1* (*LEC1*) encodes a transcription factor in the seed, which increases the H3K36me3 at the *FLC* locus and subsequently reducing the level of H3K27me3 (Tao et al. 2017). In monocot plants like rice, some of the pivotal genes underlying heading time such as *EARLY HEADING DATE 1* are controlled through epigenetic regulation of its repressors including *FUSCA 3-LIKE 1* (Jeong et al. 2015).

In the model plant Arabidopsis, a systematic search to identify epialleles was conducted from a population derived from two lines that were uniform at the nucleotide level with the exception of a mutation in the *DECREASE IN DNA METHYLATION 1 (DDM1)* gene that maintains CG methylation. Repetitive inbreeding led to the generation of an epigenetic recombinant inbred line (*epi*RIL) population varying only at their DNA methylation patterns throughout the genome (Johannes et al 2009). Phenotypic characterization of the population revealed a number of traits including flowering time and plant height that segregated in this *epi*RIL population. Based on a genetic map developed using recombination between chromosomes, methylation polymorphisms (DMR) co-segregating in the *epi*RIL population was observed (Colomé-Tataché et al. 2012). This has led to the identification of epigenetic QTLs (*epi*QTLs) underlying quantitative triats such as flowering time and root length (Cortijo et al. 2014). In *Brassica napus*, an allotetraploid, using methylation-sensitive amplified fragment length polymorphism (AFLP) markers, the epiQTLs underlying seven agronomic traits namely plant height, seed oil content, erucic acid content, protein content, seed development time, flowering time and maturity duration have been identified (Long et al. 2011). Hence, both naturally occurring as well as induced variation in methylomes can be associated with phenotypic variation (Schmitz 2014).

DNA methylation inhibitors including 5-Azacytidine (5-AzaC) and Zebularine are used to generate heritable hypomethylation associated variation which can be exploited in crop improvement (Boyko and Kovalchuk 2013). Interestingly, 5-AzaC has been reported to cause reduced DNA methylation levels across the genome without sequence-context specificity

(Griffin et al. 2016), as compared to the effects of mutation in genes underlying DNA methylation which are sequence-context specific (reviewed in Zhang et al. 2018).

The utility of this approach of using pharmacological drugs for generating variation in DNA methylation was demonstrated in flax where three early flowering lines named ‘RE1’, ‘RE2’ and ‘RE3’ were selected from a population of the variety ‘Royal’ after mutagenesis using 5-AzaC (Fieldes 1994; Fieldes and Amyot 1999). The early flowering trait identified after 5-AzaC treatment was found to be heritable and was observed to be stably transmitted through meiosis for at least nine generations (Sun 2015, MSc thesis). The heritability of this trait enabled the generation of three recombinant inbred line (RIL) populations developed by crossing each of the three early flowering derivatives to their progenitor genotype ‘Royal’. The crosses were made and RIL populations were developed at the Crop Development Centre, University of Saskatchewan. These RIL populations offered a unique genetic resource to elucidate the underlying genetic and epigenetic factors controlling the variation in these early flowering lines. After studying the possible genetic basis of the observed early flowering trait using the QTL-seq strategy as described in the previous chapters, the investigation into the methylation differences between the early- and late flowering bulks and their potential influence on flowering time are described in this chapter.

## 6.3 Materials and methods

### 6.3.1 Bisulfite conversion of DNA

The DNA samples of the parents ‘Royal’, ‘RE2’ and five constituent lines each of early- and late flowering bulks, were subjected to bisulfite conversion using the EZ DNA Methylation - Gold Kit (Zymo Research Corp., CA, USA). The CT conversion reagent was prepared by adding, 900 µl of nuclease free water, 50 µl of M-dissolving buffer and 300 µl of M-dilution buffer. The reagent was mixed thoroughly for 10 minutes by vortexing. DNA sample (100 ng of DNA in 20 µl), spiked-in with 0.27 ng of λ DNA, was added with 130 µl of CT conversion reagent and were mixed by pipetting up and down. The DNA was denatured at 98$^{o}$C for 10 minutes followed by the conversion step at 64$^{o}$C for 2.5 hours and the samples were cooled down to 4$^{o}$C. A volume of 600 µl of M-binding buffer was added to a Zymo-Spin IC column placed on a collection tube, later to which the samples were loaded. The columns were inverted to mix the sample and the buffer. The spin column was centrifuged at a speed of 11,000 x g for 30 seconds. The solution passing through the filter was discarded. Each spin column was added with 100 µl of M-wash buffer and centrifuged at 11,000 x g for 30 seconds. M-desulphonation buffer was added at the rate of 200 µl per column and incubated at room temperature for 20 minutes. Later, spin columns were centrifuged at 11,000 x g for 30 seconds and the filters were washed with 200 µl of M-wash buffer twice by centrifugation at 11,000 x g. Finally, 10 µl of M-elution buffer was added to the matrix and the DNA was eluted into a 1.5 ml Eppendorf tube by centrifugation at 11,000 x g for 30 seconds. The bisulfite converted DNA was quantified on a NanoDrop 2000 spectrophotometer (ThermoFisher Scientific, MA, USA) in the RNA setting, and the samples were stored at -20$^o$C.

### 6.3.2 Preparation of DNA methylation libraries

The methylation libraries using the bisulfite-treated DNA were constructed using the TruSeq DNA Methylation Kit (Illumina Inc., USA). The protocol followed is described below.

#### 6.3.2.1 Synthesis-primer annealing

DNA sample containing ~50 ng of bisulfite converted DNA (in 9 µl volume) was added with 2 µl of synthesis-primer. The samples were placed in a thermal cycler at 95$^o$C for 5 minutes and immediately after removal, were placed on an ice water bath.

#### 6.3.2.2 Synthesis of DNA

During this step, DNA fragments were added with random hexamer tags and amplified. A master mix containing 4 µl TruSeq DNA Methyl PreMix per sample, 0.5 µl each of 100mM Dithiothreitol (DTT) and TruSeq DNA Methyl Polymerase per sample was prepared. Sample tubes placed on the ice water bath were added with 5 µl of Master mix and homogenized by pipetting. With a final volume of 16 µl, the following program was run on a thermal cycler with preheated lid: 25$^o$C for 5 minutes, 42$^o$C for 30 minutes, 37$^o$C for 2 minutes and final cooling down to 4$^o$C. The sample tubes were removed one by one from the thermal cycler and added with 1 µl of Exonuclease I, mixed thoroughly, and placed again on the thermal cycler maintained at 4$^o$C. Then the following program with preheated lid option was run: 37$^o$C for 10 minutes, 95$^o$C for 3 minutes, 25$^o$C for 2 minutes and cool down to 4$^o$C.

### 6.3.2.3 Tagging DNA

In this step, the 3' end of the fragments were added with a complementary sequence, thus resulting in a di-tagged DNA. A master mix was prepared by mixing 7.5 µl of TruSeq DNA Methyl Term Tag Premix per sample and 0.5 µl DNA polymerase per sample. Each of the synthesized DNA sample was removed one after another from the thermal cycler held at 4$^{o}$C and added with 8 µl of master mix and pipetted up and down to mix and returned to the thermal cycler. With the total volume of 25 µl the following program was run: 25$^{o}$C for 5 minutes, 95$^{o}$C for 3 minutes and finally maintained at 4$^{o}$C.

### 6.3.2.4 Clean-up of di-tagged DNA

The di-tagged DNA samples were cleaned with AMPure XP beads (Beckman Coulter, CA, USA). A volume of 40 µl of AMPure XP beads was added to each sample and mixed well. The entire content was then transferred to a 1.5 ml tube and incubated for 5 minutes. The microcentrifuge tubes containing the samples were then placed on a magnetic stand for 5 minutes for the beads to settle down and, the supernatant was discarded. The beads were added with 200 µl of ethanol (80% v/v) allowed to stand for 30 seconds and then removed completely, and the ethanol wash was repeated. The samples were centrifuged at 1000 rpm for 10 seconds and placed on a magnetic stand. Residual ethanol was removed using 10 µl pipette after one-minute of incubation. The beads were air dried on the magnetic stand for 3 minutes. A volume of 24.5 µl of nuclease free water was added and the sample tubes were removed from the stand. The samples were mixed well by pipetting up and down ten times and then incubated for 2 minutes. The beads were allowed to settle down by placing the samples on a magnetic stand for 5 minutes. Later, 22.5 µl of supernatant was transferred to a fresh PCR tube.

### 6.3.2.5 Library amplification

Each of the di-tagged DNA sample as template was added with 25 µl of Failsafe PCR PreMix, 1 µl of corresponding adapter-index specific for a given sample and 0.5 µl of Failsafe PCR enzyme mix were added, and the total volume was 50 µl. The following PCR program was executed with preheated lid: 95$^{o}$C for 1 minute; 10 cycles of 95$^{o}$C for 30 seconds, 55$^{o}$C for 30 seconds, 68$^{o}$C for 3 minutes; 7 minutes at 68$^{o}$C and final cooling at 4$^{o}$C. The amplified libraries were cleaned-up as described earlier using AMPure XP beads at the ratio of 1:1 of beads to DNA sample. The beads were resuspended in a volume of 20 µl of nuclease-free water and finally 20 µl of each library was collected in a fresh tube.

#### 6.3.2.6 Library quantification and sequencing

The DNA libraries were quantified using Qubit high sensitivity (HS) DNA assay (Thermofisher Scientific, MA, USA). The quality was assessed using BioAnalyzer HS DNA kit (Agilent technologies, Germany). The quantity and quality assessment procedure were like that for DNA sequencing libraries. The libraries were sequenced at NRC Aquatic and Crop Resource Centre, Saskatoon. The protocol for library quantification and sequencing was similar to that of genomic DNA libraries. Briefly, the indexed-libraries were quantified using qPCR (Kapa Biosystems, MA, USA), diluted and pooled. A final concentration of 20 pM of the libraries were spiked-in with 5% Phix library as control and sequenced on the Illumina HiSeq2500 platform using HiSeq SBS v4 chemistry with 2 x 125 bp cycles.

### 6.3.3 Generation of secondary reference for 'RE2'

Whole genome bisulfite sequencing (WGBS) uses SNP to infer the methylation status and any SNP between 'RE2' and 'CDC Bethune' would confound the interpretation. Hence, secondary 'RE2' reference was generated using custom PERL scripts. The sequencing reads of 'RE2' from genomic DNA sequencing were aligned to the reference genome of 'CDC Bethune' using Bowtie 2. Each position of the genome was scanned. A minimum coverage of eight reads per locus was set as the threshold. The major allele present in the consensus ('RE2') was used to replace the allele in the reference sequence ('CDC Bethune'). If a single allele could not be distinguished as the major allele, an ambiguous base was inserted (as per IUPAC nomenclature). The regions with low or no coverage were annotated as they can be excluded from further analysis.

### 6.3.4 Analysis of bisulfite sequencing data

The FASTQ files containing the processed paired-reads were aligned to the secondary reference of 'RE2' using BSMAP 2.89 (Xi and Li 2009). The adapter sequences, and low-quality sequences (Phred score < 30) at the 3'end were trimmed. A minimum base quality filter was set at the default value of 33. A python script (*methratio*) available in BSMAP was used to extract the methylation ratios only from the reads that were uniquely aligned. Methylation ratio was estimated as the ratio of number Cytosines (C) to the total number of Cs and Thymines (T) at a given position of read alignment. Similarly, the bisulfite reads were also aligned to the λ DNA sequence using BSMAP, and the methylation ratio was determined. The λ DNA from a bacteriophage has only unmethylated Cs and therefore all the Cs are expected to be converted to Ts during bisulfite conversion. Based on the methylation ratio of Cs in the spiked-in lambda DNA, the rate of bisulfite conversion for each library was estimated.

The output files of the early flowering lines obtained from *methratio* were merged into a single file, and similarly, the output files were combined for late flowering lines as well. Methylation status of each of the C residue at single base pair resolution in the context of CG, CHG and CHH were called and statistical assessment was carried out assuming binomial distribution so as to determine whether methylation of a given locus occurred only by chance. Circos plots (Krzywinski et al. 2009) depicting the methylation status across the genome in ‘Royal’ and ‘RE2’ were developed by dividing the chromosomes into bins of 250,000 bp length and counting the number of methylated cytosines (5mC) in all three sequence contexts. The global differences in DNA methylation between the accessions (‘Royal’ vs ‘RE2’) were visualized from Circos plots.

The identification of differentially methylated regions (DMRs) between the bulks using custom R scripts included the following steps: firstly, the chromosomes were split into bins of size 1000 bp. A score of zero and one was given to the unmethylated and methylated sites, respectively. The number of Cs and 5mCs in each of the bins were counted. Fisher’s exact test was performed to infer significant differences in the proportion of methylated and unmethylated cytosines in each of the bins of the bulks.

The genes partially (or) fully overlapping the significant DMRs and those with the DMRs in the upstream region (upstream interval considered as 5kb) were identified by parsing the GFF annotation file. The homology search for the genes was carried out using protein basic local alignment search tool (BLASTP; Altschul et al. 1997). For protein homology identification, a 30% identity is the thumb rule, and an E-value less than 0.001 is dependable (Pearson 2013).

## 6.4 Results

### 6.4.1 DNA methylation analysis in the early- and late flowering bulks

From the Circos plots of ‘Royal’ and ‘RE2’, the methylation patterns were observed to be similar, globally (Figure 6.1). However, differential methylation patterns were observed at 494,263 positions among the total 57,192,166 Cs sequenced in both the bulks. Among the differentially methylated positions between the early flowering and the late flowering bulks, a total of 260,193 were transformed from methylated state in the late flowering bulk to the unmethylated state in the early flowering bulk.

Based on the distribution of DNA methylation ratio depicted in Figure 6.2, most Cs in the genome were unmethylated. A major fraction of the 5mCs in the bulks were present in the CG context. In the early flowering group, the total number of 5mCs with a methylation ratio of one, was 1,212,558 of which, a total of 1,082,602 were present in the CG context. The number of 5mCs in the CHG and CHH context were 125,464 and 4,492, respectively. A similar pattern was observed in the late flowering bulk in which, of the 1,376,082 positions with a methylation ratio of one, a total of 1,222,543 were in the CG context. The methylated sites in the CHG and CHH contexts were 149,610 and 3,929, respectively (Figure 6.3).

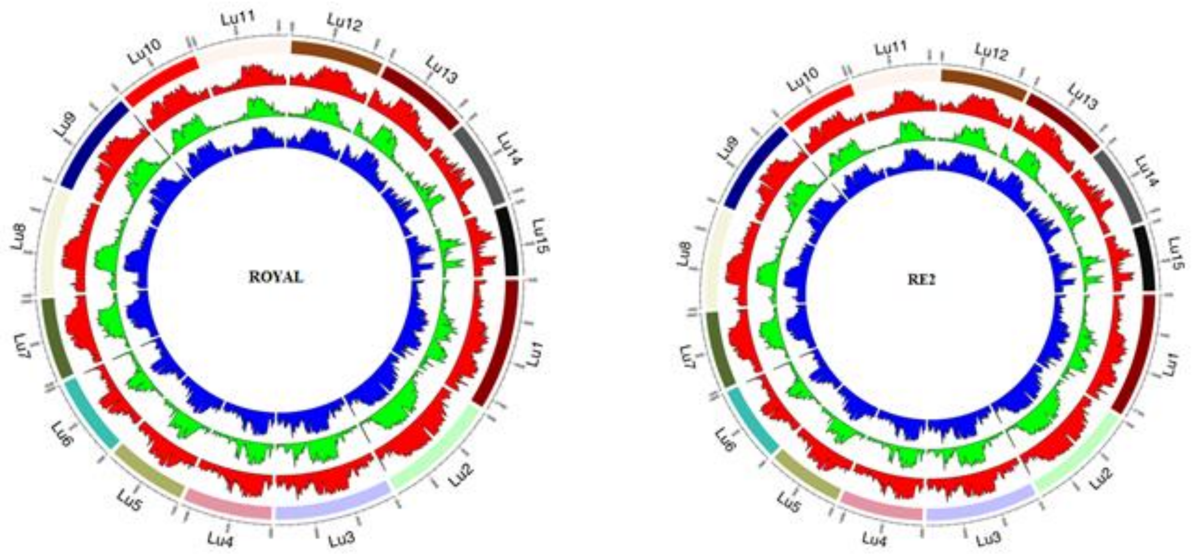

**Figure 6.1** The global methylation pattern of the cultivar ‘Royal’ and its epimutant ‘RE2’. The methylation in the CG context is depicted in red whereas, those in the CHG and CHH contexts are depicted in green and blue, respectively.

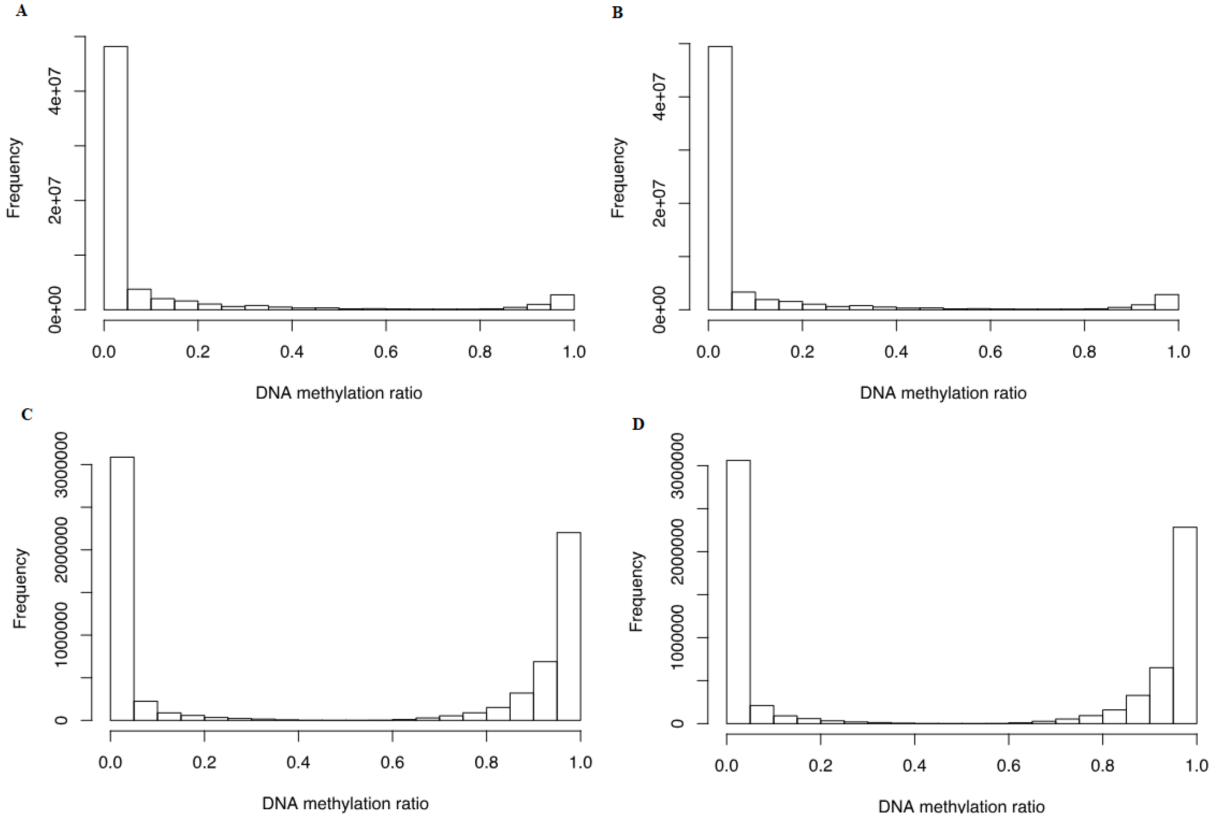

**Figure 6.2** Distribution of DNA methylation ratio in the early- (Panel A) and late (Panel B) flowering bulks. Distribution of DNA methylation ratio in the CG context in the early- (Panel C) and late (Panel D) flowering bulks. The number Cytosines to the total number of both Cytosines and Thymines in the consensus, at a given position is defined as methylation ratio.

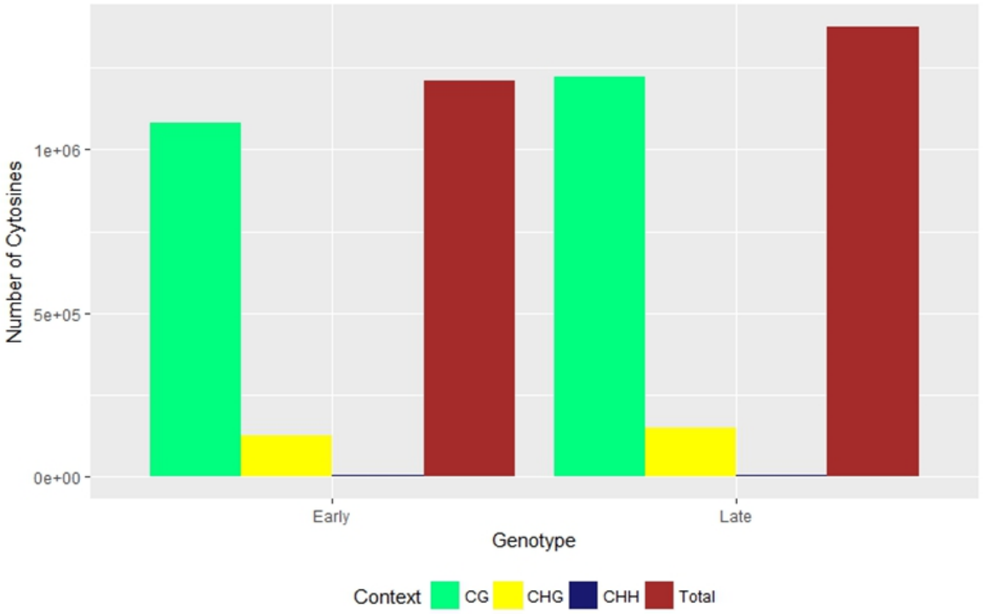

**Figure 6.3** The frequency of methylation in the CG, CHG and CHH contexts between the early- and late flowering bulks.

### 6.4.2 Differentially methylated regions between the bulks

Genome-wide distribution of DMRs between the bulks is depicted in the Figure 6.4. Interestingly, there was a cluster of highly significant DMRs on the chromosome 12. The potential association of DMRs with any of the genes was deciphered by parsing the GFF annotation file, specifically for extracting the coordinates of genes that encompasses these bins. A total of 127 DMRs were found to be significant (P<0.01). Among the significant DMRs, 59 overlapped with the flax genes, 35 DMRs were identified in the upstream region (5 kb interval) and the remaining were intergenic. The results from the homology search using BLASTP of the genes with overlapping DMRs and DMRs in the upstream region are listed in Table 6.1 and Table 6.2, respectively.

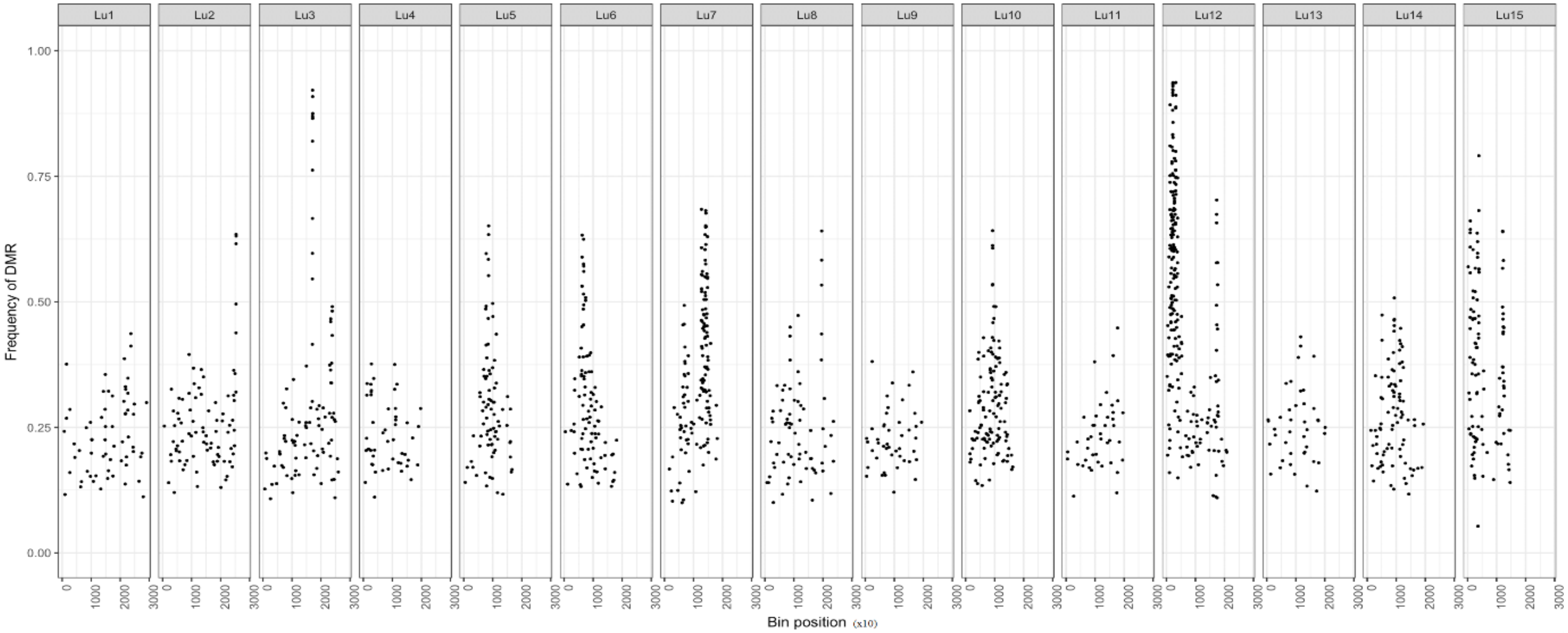

**Figure 6.4** The distribution of differentially methylated regions (DMRs) between the early- and late flowering bulks, across the 15 chromosomes (Lu 1-15) of flax. The X-axis represents the bin position, and along the Y-axis the frequency of DMRs can be observed.

**Table 6.1** Homology of flax genes overlapping significant DMRs

| Flax gene | Start position | End position | Length (bp) | Strand | Identity (%) | Homolog | E-value | Database for homology search |
|---|---|---|---|---|---|---|---|---|
| | | | | | **Chromosome 1** | | | |
| Lus10036114 | 386243 | 387058 | 815 | - | 56 | FLA11 FASCICLIN-like arabinogalactan-protein 11 in *Arabidopsis thaliana* | $2e^{-59}$ | Swiss-Prot |
| | | | | | **Chromosome 2** | | | |
| Lus10008664 | 2554109 | 2556618 | 2509 | + | 50 | Probable protein arginine N-methyltransferase 3 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10014341 | 3049938 | 3054477 | 4539 | - | 47 | Conserved hypothetical protein in *Ricinus communis* | $5e^{-141}$ | Non-redundant |
| Lus10038680 | 6857815 | 6865629 | 7814 | + | 43 | Protein RNA-directed DNA methylation 3 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10004633 | 24954252 | 24957818 | 3566 | - | 50 | Cytochrome P450 in *Panax ginseng* | 0.0 | Swiss-Prot |
| | | | | | **Chromosome 3** | | | |
| Lus10033491 | 17120019 | 17122613 | 2594 | - | 68 | E3 ubiquitin-protein ligase RGLG3 in *Arabidopsis thaliana* | $8e^{-97}$ | Swiss-Prot |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10033845 | 18851473 | 18858041 | 6568 | - | 66 | Protein HASTY 1 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10017072 | 24370784 | 24373820 | 3036 | - | 71 | Galactokinase in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| **Chromosome 4** | | | | | | | | |
| Lus10001594 | 8619737 | 8621386 | 1649 | + | 48 | Protein C2-DOMAIN ABA-RELATED 4 in *Arabidopsis thaliana* | $8e^{-60}$ | Swiss-Prot |
| **Chromosome 5** | | | | | | | | |
| Lus10032374 | 3348243 | 3350749 | 2506 | - | 48 | Probable polygalacturonase in *Vitis vinifera* | $2e^{-138}$ | Swiss-Prot |
| Lus10011759 | 8723553 | 8725664 | 2111 | - | 31 | Hypothetical protein A4A49_54874 *Nicotiana attenuate* | $2e^{-05}$ | Non-redundant |
| **Chromosome 6** | | | | | | | | |
| Lus10006391 | 12820091 | 12822797 | 2706 | + | 61 | Fasciclin-like arabinogalactan in *Arabidopsis thaliana* | $6e^{-152}$ | Swiss-Prot |
| Lus10013674 | 6472963 | 6474142 | 1179 | + | 77 | Vignain in *Phaseolus vulgaris* | 0.0 | Swiss-Prot |
| **Chromosome 7** | | | | | | | | |
| Lus10025435 | 14152109 | 14156328 | 4219 | - | 83 | 28 kDa heat- and acid-stable phosphoprotein in *Jatropha curcas* | $5e^{-82}$ | Non-redundant |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10023183 | 17420460 | 17421101 | 641 | - | 39 | Fasciclin-like arabinogalactan protein 19 in *Arabidopsis thaliana* | $4e^{-12}$ | Swiss-Prot |
| Lus10025427 | 14125327 | 14126710 | 1383 | + | 29 | Glutamate receptor 2.9 in *Arabidopsis thaliana* | $3e^{-15}$ | Swiss-Prot |
| | | | | | **Chromosome 8** | | | |
| Lus10014109 | 4205773 | 4206875 | 1102 | + | 63 | Glucan endo-1,3-beta-glucosidase, basic isoform in *Prunus persica* | $8e^{-144}$ | Swiss-Prot |
| Lus10023837 | 5310160 | 5311134 | 974 | + | 39 | Putative clathrin assembly protein in *Arabidopsis thaliana* | $2e^{-57}$ | Swiss-Prot |
| Lus10012659 | 12052165 | 12052629 | 464 | + | 56 | Putative receptor protein kinase ZmPK1 in *Zea mays* | $4e^{-49}$ | Swiss-Prot |
| | | | | | **Chromosome 9** | | | |
| Lus10001664 | 493942 | 495527 | 1585 | - | 52 | NAC domain-containing protein 12 in *Arabidopsis thaliana* | $2e^{-115}$ | Swiss-Prot |
| Lus10008459 | 12593249 | 12596019 | 2770 | + | 63 | Protein Brevis radix-like 2 in *Arabidopsis thaliana* | $2e^{-154}$ | Swiss-Prot |
| | | | | | **Chromosome 10** | | | |
| Lus10009605 | 180320 | 183210 | 2890 | + | 43 | Polygalacturonase in *Gossypium hirsutum* | $7e^{-108}$ | Swiss-Prot |
| | | | | | **Chromosome 11** | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10036501 | 7555296 | 7556977 | 1681 | + | 66 | Ecotropic viral integration site 5 protein homolog *Hevea brasiliensis* | $8e^{-54}$ | Non-redundant |
| Lus10001121 | 9824517 | 9834015 | 9498 | + | 61 | Uncharacterized protein LOC110650594 isoform X3 *Hevea brasiliensis* | 0.0 | Swiss-Prot |
| | | | | | | **Chromosome 12** | | |
| Lus10005864 | 3846506 | 3849179 | 2673 | + | 39 | Transcription factor MYB119 in *Arabidopsis thaliana* | $5e^{-62}$ | Swiss-Prot |
| Lus10023272 | 2131061 | 2135409 | 4348 | - | 29 | Disease resistance protein TAO1 in *Arabidopsis thaliana* | $5e^{-95}$ | Swiss-Prot |
| Lus10023269 | 2156358 | 2159141 | 2783 | + | 68 | Nucleotide exchange factor SIL1 *Hevea brasiliensis* | $1e^{-179}$ | Swiss-Prot |
| Lus10036222 | 1205275 | 1208332 | 3057 | + | 53 | Probable potassium transporter 11 in *Oryza sativa* | 0.0 | Swiss-Prot |
| Lus10015270 | 3172159 | 3173214 | 1055 | - | 37 | Pathogenesis-related protein PR-4B in *Nicotiana tabacum* | $8e^{-07}$ | Swiss-Prot |
| Lus10036228 | 1253756 | 1254400 | 644 | + | 74 | Uncharacterized protein LOC110610418 in *Manihot esculenta* | $9e^{-102}$ | Non-redundant |
| Lus10015309 | 2991501 | 2994690 | 3189 | - | 67 | Protein trichome birefringence-like 37 in *Arabidopsis thaliana* | $1e^{-76}$ | Swiss-Prot |
| Lus10006732 | 890903 | 894258 | 3355 | + | 31 | Putative disease resistance protein in *Arabidopsis thaliana* | $1e^{-98}$ | Swiss-Prot |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10023270 | 2141942 | 2153365 | 11423 | - | 52 | 187-kDa microtubule-associated protein AIR9 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10006972 | 2687278 | 2691032 | 3754 | - | 69 | NADPH--cytochrome P450 reductase 2 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10006971 | 2697545 | 2701065 | 3520 | + | 60 | Polyol transporter 5 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10015268 | 3179644 | 3185302 | 5658 | - | 38 | Cation/H (+) antiporter 15 in *Arabidopsis thaliana* | $8e^{-166}$ | Swiss-Prot |
| Lus10006964 | 2735694 | 2737801 | 2107 | + | 31 | PREDICTED: cysteine-rich receptor-like protein kinase 10 *Gossypium hirsutum* | $4e^{-06}$ | Non-redundant |
| Lus10015302 | 3025921 | 3027927 | 2006 | + | 83 | Magnesium-chelatase subunit ChlI, chloroplastic in *Glycine max* | 0.0 | Swiss-Prot |
| Lus10023214 | 2368195 | 2370477 | 2282 | - | 29 | BAHD acyltransferase in *Arabidopsis thaliana* | $2e^{-45}$ | Swiss-Prot |
| Lus10015256 | 3252003 | 3254033 | 2030 | - | 61 | Pentatricopeptide repeat-containing protein in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10001631 | 3717722 | 3722742 | 5020 | - | 89 | Plasma membrane ATPase in *Oryza sativa japonica* | 0.0 | Swiss-Prot |
| Lus10005865 | 3850636 | 3852542 | 1906 | - | 69 | GDP-mannose transporter GONST2 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10015323 | 2920117 | 2925889 | 5772 | - | 65 | Mediator of RNA polymerase II transcription subunit 15a in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10033083 | 14498445 | 14501938 | 3493 | - | 60 | tRNA (guanine(37)-N1)-methyltransferase 1 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10015257 | 3243228 | 3248023 | 4795 | + | 49 | uncharacterized protein LOC110604770 isoform X2 in *Manihot esculenta* | 0.0 | Non-redundant |
| Lus10015303 | 3022724 | 3024682 | 1958 | + | 30 | Spermidine sinapoyl-CoA acyltransferase in *Arabidopsis thaliana* | $2e^{-47}$ | Swiss-Prot |
| | | | | | | **Chromosome 13** | | |
| Lus10001038 | 2911140 | 2912021 | 881 | - | 28 | Disease resistance protein TAO1 in *Arabidopsis thaliana* | $3e^{-19}$ | Swiss-Prot |
| Lus10001039 | 2912176 | 2914466 | 2290 | - | 27 | Disease resistance protein TAO1 in *Arabidopsis thaliana* | $3e^{-42}$ | Swiss-Prot |
| Lus10013521 | 7748666 | 7791579 | 42913 | - | 83 | ATP-dependent zinc metalloprotease FTSH 4 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10010827 | 12642960 | 12643325 | 365 | + | | NA | | |
| Lus10032079 | 16147622 | 16153687 | 6065 | - | 85 | Glucose-6-phosphate 1-dehydrogenase in *Solanum tuberosum* | 0.0 | Swiss-Prot |
| | | | | | | **Chromosome 14** | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10014214 | 5038380 | 5040113 | 1733 | + | 60 | PREDICTED: uncharacterized protein LOC105110757 in *Populus euphratica* | $5e^{-169}$ | Non-redundant |
| Lus10025499 | 1965383 | 1968674 | 3291 | + | 66 | PREDICTED: putative CCA tRNA nucleotidyltransferase 2 in *Populus euphratica* | 0.0 | Non-redundant |
| Lus10014992 | 18815618 | 18816421 | 803 | - | 52 | Protein phosphatase inhibitor 2 in *Arabidopsis thaliana* | $9e^{-34}$ | Swiss-Prot |
| Lus10006918 | 7610879 | 7612459 | 1580 | + | | NA | | |
| **Chromosome 15** | | | | | | | | |
| Lus10012685 | 3937136 | 3938389 | 1253 | + | 73 | Putative elongation of fatty acids protein DDB_G0272012 in *Hevea brasiliensis* | $5e^{-108}$ | Non-redundant |
| Lus10041102 | 9326319 | 9330425 | 4106 | - | 76 | Uncharacterized WD repeat-containing protein C2A9.03 isoform X1 in *Jatropha curcas* | 0.0 | Non-redundant |
| Lus10012714 | 3788512 | 3788962 | 450 | - | 74 | 60S acidic ribosomal protein P3-2 in *Arabidopsis thaliana* | $5e^{-28}$ | Swiss-Prot |
| Lus10007635 | 152797 | 159648 | 6851 | - | 39 | Cysteine-rich receptor-like protein kinase 29 in *Arabidopsis thaliana* | $6e^{-121}$ | Swiss-Prot |

**Swiss-Prot is the primary database for homology search for queries. Non-redundant database was used when the query had no hits in the manually curated Swiss-Prot.*

**Table 6.2** Homology of flax genes containing significant DMRs in their upstream region

| Flax gene | Start position | End position | Length (bp) | Strand | Identity (%) | Homolog | E-value | Database for homology search |
|---|---|---|---|---|---|---|---|---|
| **Chromosome 1** | | | | | | | | |
| Lus10020478 | 21453813 | 21455174 | 1362 | - | 46 | Basic helix-loop-helix (bHLH) DNA-binding superfamily protein in *Arabidopsis thaliana* | $2e^{-86}$ | Swiss-Prot |
| **Chromosome 2** | | | | | | | | |
| Lus10009220 | 6440923 | 6446522 | 5600 | + | 74 | Adenosine kinase 2 in *Arabidopsis thaliana* | $6e^{-175}$ | Swiss-Prot |
| **Chromosome 3** | | | | | | | | |
| Lus10037689 | 24903671 | 24905223 | 1553 | + | 61 | Uncharacterized protein LOC7475680 in *Populus trichocarpa* | $2e^{-101}$ | Non-redundant |
| Lus10017172 | 23808599 | 23812123 | 3525 | - | 89 | MOB kinase activator-like 1A in *Arabidopsis thaliana* | $7e^{-141}$ | Swiss-Prot |
| Lus10033551 | 17035798 | 17039559 | 3762 | - | 43 | Endonuclease or glycosyl hydrolase, putative isoform 1 in *Theobroma cacao* | 0.0 | Non-redundant |
| Lus10029226 | 8160367 | 8160723 | 357 | + | 50 | Non-specific lipid-transfer protein in *Gossypium hirsutum* | $3e^{-32}$ | Swiss-Prot |
| **Chromosome 6** | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10043394 | 2452401 | 2454874 | 2474 | + | 86 | Vesicle-associated membrane protein 714 in *Arabidopsis thaliana* | 6e$^{-143}$ | Swiss-Prot |
| **Chromosome 7** | | | | | | | | |
| Lus10007408 | 12976817 | 12978130 | 1314 | - | | NA | | |
| Lus10025391 | 13889242 | 13890968 | 1727 | + | 52 | High mobility group B protein 14 in *Arabidopsis thaliana* | 8e$^{-41}$ | Swiss-Prot |
| Lus10028791 | 1682626 | 1683381 | 756 | + | 53 | B-box zinc finger protein 21 in *Arabidopsis thaliana* | 1e$^{-64}$ | Swiss-Prot |
| Lus10011634 | 12732518 | 12739509 | 6992 | + | 52 | E3 SUMO-protein ligase SIZ1 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| **Chromosome 8** | | | | | | | | |
| Lus10021856 | 7630107 | 7632239 | 2133 | - | 55 | Serine carboxypeptidase-like 45 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10014090 | 4133426 | 4134103 | 678 | + | 38 | Uncharacterized protein LOC110428252 isoform X2 in *Herrania umbratica* | 3e$^{-13}$ | Non-redundant |
| **Chromosome 9** | | | | | | | | |
| Lus10031062 | 6408336 | 6411234 | 2899 | + | 78 | Ubiquitin carboxyl-terminal hydrolase 2 in *Arabidopsis thaliana* | 7e$^{-180}$ | Swiss-Prot |
| Lus10007668 | 11532626 | 11532946 | 321 | - | 51 | PREDICTED: uncharacterized protein LOC105122337 in *Populus euphratica* | 1e$^{-19}$ | Non-redundant |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10024894 | 19387836 | 19391088 | 3253 | + | 65 | Protein EARLY-RESPONSIVE TO DEHYDRATION STRESS 4 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| | | | | | **Chromosome 10** | | | |
| Lus10022998 | 6877076 | 6877348 | 273 | - | 53 | Metallothionein-like protein 4A in *Arabidopsis thaliana* | $7e^{-17}$ | Swiss-Prot |
| Lus10039990 | 9199789 | 9203106 | 3318 | + | 35 | Uncharacterized protein LOC111890747 in *Lactuca sativa* | $3e^{-07}$ | Non-redundant |
| | | | | | **Chromosome 11** | | | |
| Lus10012668 | 11569436 | 11570482 | 1047 | - | 41 | Mitogen-activated protein kinase kinase kinase 17 in *Arabidopsis thaliana* | $1e^{-60}$ | |
| | | | | | **Chromosome 12** | | | |
| Lus10023307 | 1979199 | 1979432 | 234 | + | 55 | Outer envelope membrane protein 7 in *Arabidopsis thaliana* | $7e^{-09}$ | Swiss-Prot |
| Lus10023215 | 2367039 | 2367584 | 546 | + | 32 | Cell wall / vacuolar inhibitor of fructosidase 2 in *Arabidopsis thaliana* | $8e^{-06}$ | Swiss-Prot |
| Lus10019967 | 266946 | 268130 | 1185 | + | 86 | UDP-glucuronate 4-epimerase 1 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10015284 | 3104572 | 3107592 | 3021 | - | 57 | Protochlorophyllide-dependent translocon component 52, chloroplastic in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Lus10015239 | 3312038 | 3317370 | 5333 | - | 61 | Probable LRR receptor-like serine/threonine-protein kinase in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10023323 | 1893942 | 1898077 | 4136 | + | 50 | LRR receptor-like serine/threonine-protein kinase FLS2 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10036234 | 1287415 | 1290657 | 3243 | + | 62 | Protein SUPPRESSOR OF FRI 4 in *Arabidopsis thaliana* | $1e^{-149}$ | Swiss-Prot |
| Lus10015319 | 2944696 | 2947594 | 2899 | + | 33 | Protein FRIGIDA-ESSENTIAL 1 in *Arabidopsis thaliana* | $5e^{-29}$ | Swiss-Prot |
| Lus10024240 | 10093696 | 10097647 | 3952 | + | 57 | RINT1-like protein MAG2 in *Arabidopsis thaliana* | 0.0 | Swiss-Prot |
| Lus10018322 | 5271096 | 5272091 | 996 | + | 72 | Uncharacterized protein LOC105641058 in *Jatropha curcas* | $7e^{-162}$ | Non-redundant |
| Lus10004942 | 3447012 | 3448943 | 1932 | + | 90 | 60S ribosomal protein L23a-2 in *Arabidopsis thaliana* | $8e^{-69}$ | Swiss-Prot |
| **Chromosome 14** | | | | | | | | |
| Lus10003753 | 10413002 | 10413400 | 399 | + | | NA | | |
| Lus10018087 | 16292336 | 16294771 | 2436 | + | 56 | BTB/POZ and TAZ domain-containing protein 2 in *Arabidopsis thaliana* | $1e^{-133}$ | Swiss-Prot |
| **Chromosome 15** | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| Lus10041344 | 10397686 | 10400952 | 3267 | - | 38 | Cation/H (+) antiporter 3 in *Arabidopsis thaliana* | $3e^{-169}$ | Swiss-Prot |
| Luss10022545 | 12283535 | 12286539 | 3005 | - | 63 | Farnesyl pyrophosphate synthase 2 in *Lupinus albus* | $3e^{-163}$ | Swiss-Prot |

**Swiss-Prot is the primary database for homology search for queries. Non-redundant database was used when the query had no hits in the manually curated Swiss-Prot.*

### 6.5 Discussion

The current study for methylation analysis is first of its kind in flax, employing whole genome bisulfite sequencing methodology. A higher proportion of methylated cytosines were observed in the CG context and the methylation was minimum in the CHH context which could be observed as a shift in the distribution (Figure 6.2), although most of the cytosines in the genome were present in the CHH sequence-context. This trend of higher CG methylation compared to CHH is similar to that observed in Arabidopsis (Lister et al. 2008), soybean (*Glycine max* (L.) Merr.; Schmitz et al. 2013) and rice (*Oryza sativa* L. ssp. *japonica*; Li et al. 2012) and is contrary to that observed in mungbean (*Vigna radiata* (L.) Wilczek; Kang et al. 2017), common bean (*Phaseolus vulgaris* L.; Kim et al. 2015) and cassava (*Manihot esculenta* Crantz; Wang et al. 2015). The Circos plot (Figure 6.1) depicting the distribution of whole genome DNA methylation pattern of ‘Royal’ (late flowering) and ‘RE2’ (early flowering) were found to be similar. This might be because the Circos plot was developed by binning the genome with a size of 250,000 nucleotides which might not have sufficient resolution to distinguish small but significant locus-specific methylation differences between ‘Royal’ and ‘RE2’.

In the early- and late flowering bulks, at the single nucleotide level, there were 494,263 cytosines differentially methylated, which were distributed across the genome. This implied the lack of any bias in the influence of 5-AzaC all through the genome (Griffin et al. 2016). Among the total differentially methylated cytosines, 260,193 were in an unmethylated state in the early flowering bulk in contrast to the late flowering bulk. This is potentially due to the incorporation of 5-AzaC, a hypomethylating chemical in place of cytosine (as an analogue) during DNA-replication making it impossible for the transfer of methyl group to the fifth carbon, from the metabolite pool, and in addition, 5-AzaC forms a permanent covalent bond with DNA methyltransferase, diminishing the availability of this enzyme also leading to reduced methylation level across the genome (Pecinka and Liu 2014).

In plants, DNA methylation differences at specific loci, called epialleles, are found to be inherited without changes across generations (Johannes et al. 2009; Hofmeister et al. 2017). Hence, the epigenetic basis of phenotypic diversity can be exploited as a source of variation in crop improvement. In the transition of cotton from the photoperiod sensitive state observed in wild accession to current day photoperiod insensitive cultivars of *Gossypium hirsutum* L. and *Gossypium barbadense* L., the basis has been identified to be the hypomethylation of *CONSTANS-LIKE 2D* (*COL2D*; Song et al. 2017), a key player in promoting flowering. An epiallele of rice *ADENYLATE KINASE* (*OsAK1*) generated by the presence of hypermethylation in the promoter, altered the photosynthetic efficiency by differentially regulating *OsAK1* and other photosynthesis related genes such as *PHOTOCHLOROPHYLLIDE OXIDOREDUCTASE* and *β-CAROTENE HYDROXYLASE DSM2* (Wei et al. 2017).

Out of the total 127 significant DMRs detected in this study, 59 were found to be overlapping with genes. A total of 35 DMRs were within the upstream region (5kb interval) of genes and the remaining 33 DMRs were present in the intergenic region. Gene body methylation influences gene expression both quantitatively and qualitatively by controlling level of transcription (Zilberman et al. 2007) and nature of transcripts through modifying the splice acceptor sites as observed in oil palm (Ong-Abdullah et al. 2015), respectively. Homology searches (BLASTP) of proteins encoded by genes overlapping DMRs identified three flax genes, *Lus10036114* (on linkage group 1), *Lus10006391* (on linkage group 6) and *Lus10023183* (on linkage group 7) to be homologues of FASCICLIN-LIKE ARABINOGALACTAN (FLA) group of proteins in Arabidopsis. In flax, FLA proteins have been reported to control fibre development (Roach & Deyholes 2007). These FLA group of proteins were found to contribute to secondary cell wall synthesis and hence, influencing the biomechanics of the plant shoot development (MacMillan et al. 2010). Interestingly, plant height was also observed to be segregating in the ‘RC’ x ‘RE2’ RIL population, evaluated over three years. In textile hemp, the promoter region of FASCICLIN-LIKE ARABINOGALACTAN encoding gene are found to harbour motif associated with photoperiod recognition and regulation of flowering time (Guerriero et al. 2017). Also, the over expression of ARABINOGALACTAN encoding gene in cucumber (*CsAGP1*) has been documented to result in stem elongation and earlier flowering phenotypes (Park et al. 2003).

Apart from gene body methylation, DNA methylation in the upstream region plays a significant role in the regulation of gene expression. In tomato (*Solanum lycopersicum* L.), non-ripening phenotype has been reported to be due to hypermethylation of a region 2000 bp upstream to the colourless non-ripening *(CNR)* locus (Manning et al. 2006). In rice (*Oryza sativa* L.), the hypermethylation state of the promoter of *DWARF1* and its consequent silencing led to a highly stable dwarf phenotype maintained for more than 90 years (Miura et al. 2009). The methylation level in the promoter region of floral development controlling gene *(MeGI)* in hexaploid persimmon (*Diospyros kaki* Thunb.) is associated with sex determination (Akagi et al. 2016).

In flax, significant DMRs were observed to be present in the upstream region of flax genes on chromosome 12, whose Arabidopsis orthologues were involved in flowering. Interestingly, in our QTLseqr analysis also, genomic region on the same linkage group (12) was found to have significant association with flowering time. The protein encoded by *Lus10036234* was homologous to SUPPRESSOR OF FRI 4 (SUF4; percent identity=62; E-value=1e$^{-149}$) and that of *Lus10015319* was homologous to FRIGIDA-ESSENTIAL 1 (FES1; percent identity=33; E-value=5e$^{-29}$). SUF4 is a zinc finger transcription factor which delays flowering through FRIGIDA (FRI) by up regulating *FLOWERING LOCUS C* (*FLC*; Kim and Michaels 2006). Also, it has been reported that SUF4 is bound by LUMINIDEPENDENS (LD) of the autonomous pathway, in the absence of FRI, leading to its suppression (Kim et al. 2006). In the present study, QTL-seq analysis using the pipeline developed by Takagi et al. (2013), identified a SNP upstream of LD. *FES1* also produces a zinc finger protein which aids in the transcriptional activation of *FLC*, in the presence of *FRI*, leading to delayed flowering (Schmitz et al. 2005). Since, both FES1 and SUF4 are constituents of the FRIGIDA-COMPLEX, playing a key role in transcriptional activation of *FLC* (Choi et al. 2011) subsequently resulting in delayed flowering, further functional characterization of these genes using current tools such as clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system will give better insights into the mechanics of flowering time in flax. Further investigation into the expression data from RNA-sequencing of ‘Royal’ and ‘RE2’ would give more insights about the genes differentially expressed between the lines, the influence of their methylation state and their association with flowering time. The observed early-flowering phenotype was found to be inherited over multiple generations, indicating the stable transmission of the underlying genomic variation. However, differences in DNA methylation states between the various cell-types in plants have been reported (Widman et al. 2013; Kawakatsu et al. 2016b). Hence, the analysis of differences in methylation status between the shoot apical meristem and the leaf tissues can unearth the presence of additional mechanism operating at a tissue-specific manner in the early flowering trait.

## 6.6 Conclusion

Generating heritable variability for agronomically desirable trait is a continuous process in crop improvement. When the reservoir of variability in the germplasm did not harbour the desired allele, plant breeders induce variation through hybridization and mutagenesis. In flax, ‘2126’ was the earliest maturing cultivar with a duration of ~95 days (Dribnenki et al. 2005). However, to expand the cultivable area to the northern prairies, early flowering and consequently early maturing cultivars (duration of ~90 days) are needed. Characterization of early flowering mutant ‘RE2’ derived from 5-AzaC treated ‘Royal’ and its *epi*RIL derivatives were used in this study employing classical bulked segregant analysis methodology. Bisulfite sequencing of few of the lines constituting the bulks (five lines each) and novel statistical analysis identified the DMRs potentially underlying the early flowering phenotype on the linkage group 12. Further validation can be carried out with functional genomic analysis to identify the associated candidate genes.

# Chapter 7 General Discussion and Conclusion

Gregor Johann Mendel presented his pioneering work on the principles of inheritance in garden pea (*Pisum sativum* L.) in the year 1865 (Mendel 1866). The Mendelian principles, after their rediscovery in 1900s, have remained as the foundation for genetic studies (Smýkal et al. 2016). Linkage is one of the exceptions to Mendelian genetics, and was proposed by Bateson et al. (1909), when studying flower colour and pollen shape in garden pea. However, the phenomenon of linkage was first empirically demonstrated by Thomas Hunt Morgan with the sex-linked inheritance of eye colour in Drosophila (Morgan 1910). Association of a qualitative trait (pigmentation of seed-coat in *Phaseolus vulgaris* L.) as a marker for the selection of a quantitative trait (seed weight) was proposed by Karl Sax (1923). John Marion Thoday (1961) suggested the use of the genetic markers to position polygenic traits on chromosomes.

The limited availability of multi-marker lines with morphological markers used in these formative linkage-mapping studies, were overcome with the advent of the era of DNA markers (Tanksley 1993). In their classic study to develop a human linkage map, Botstein et al. (1980) used the variation between the genomes for the location of restriction sites to develop the first-generation DNA marker named Restriction Fragment Length Polymorphism (RFLP). Generation of RFLP-based linkage map and positioning of Quantitative Trait Loci (QTLs) underlying the polygenic traits including fruit weight was carried out in tomato, a first of its kind study in plants (Paterson et al. 1988). Exploitation of polymorphism for primer binding sites led to the development of novel class of markers called Random Amplified Polymorphic DNA (RAPD), and combining both the variations (restriction sites and primer binding sites) yielded Amplified Fragment Length Polymorphism (AFLP). Using these marker types several QTLs have been mapped, including a QTL underlying fruit size, an important domestication trait in tomato, was fine mapped and the candidate gene was identified (Frary et al. 2000). First-generation sequencing (Sanger) based identification of Single Nucleotide Polymorphisms (SNP) resulted in development of SNP-based linkage maps and positioning of genes controlling quantitative traits (Rafalski 2002). To date, several genes underlying QTLs have been cloned across crop species using map-based gene cloning methodology (Kumar et al. 2017a).

The advent of next generation sequencing (NGS) technology unearthed the power of a large number of SNP present in different accessions of a crop species in studies of genetic diversity and mapping (Barabaschi et al. 2016). Illumina is most commonly used among the next-generation sequencing platforms and is capable of generating a huge volume of data in a shorter time for a lower cost in comparison to Sanger sequencing (Shendure et al. 2017). These advancements in the sequencing technologies led to the rediscovery of power of Bulked Segregant Analysis (BSA; Michelmore et al. 1991) with thousands of more SNP markers increasing precision, with the mapping-by-sequencing strategies like QTL-seq (Takagi et al. 2013).

An early flowering mutant 'RE2' was derived from the treatment of 'Royal' using the 5-Azacytidine (5-AzaC; Fieldes et al. 1994). The modified phenotype was observed to be stable across nine generations (Sun 2015, M. Sc., Thesis, University of Saskatchewan) and hence, a recombinant inbred line population (RIL) from 'Royal' x 'RE2' was developed. The RIL population was characterized phenotypically in three field seasons (2015, 2016 and 2017) at the Kernen crop research farm at the University of Saskatchewan, Saskatoon. Additionally, the data generated in the growth cabinet complemented the field data. The RILs were ranked based on the days to start of flowering in 2015-, 2016 field season and the growth cabinet data, using which the stable distributional extremes were identified. The estimation of genetic parameters is also described in Chapter 3. The bulking of sequencing data generated from individuals sharing a common phenotype, increases the signal intensity for detecting genomic regions associated with the quantitative traits (Pires and Grossniklaus 2018). Nature and number of individuals constituting the bulks determine the potential to detect both QTLs with major- and minor effects. A total of eleven individuals each in early- and late flowering bulks were used in the present analysis. Phenotypic evaluation with more replicates under controlled environments (Zou et al. 2016) and increasing the number of individuals constituting the bulks (Sun et al. 2010) would positively influence the power of BSA. However, a different expression pattern of key genes like *FLOWERING LOCUS T* (*FT*) in controlled and natural environmental conditions have been demonstrated in Arabidopsis (Song et al. 2018), suggesting the need for field-based phenotyping under natural environments. The initial analysis of the DNA sequencing data from the bulks using the QTL-seq pipeline (Takagi et al. 2013), followed by characterization of impact of nucleotide polymorphisms using SnpEff (Cingolani et al. 2012), identified a SNP upstream of a flax gene homologous to Arabidopsis *LUMINIDIPENDENS* (*LD*; Chapter 4). Early flowering phenotype observed among mutants derived from a population of 5-AzaC-treated 'Royal', was suggested to be controlled by minimum three independent loci (Fieldes and Amyot

1999). Hence, to explore the possibility of multiple loci governing the complex flowering time trait, the DNA sequencing data was further analysed using the recently published package-QTLseqr (Mansfeld and Grumet 2018). The reanalysis identified significant regions on chromosomes 9 and 12, conditioning the phenotype (Chapter 5). The genomic variants in these regions were found to be associated with genes encoding LATE EMBRYOGENESIS ABUNDANT (LEA) HYDROXYPROLINE-RICH GLYCOPROTEIN FAMILY, MAINTENANCE OF MERISTEMS-LIKE, CYTOCHROME P450 87A3 and PHLOEM PROTEIN 2-A12, based on homology analysis. These genes, though not directly involved in flowering time in Arabidopsis, might have evolved to acquire new functions (neo-functionalization; Flagel and Wendel 2009) to be associated with flowering time in flax. In addition, these detected polymorphisms are markers and might be in linkage disequilibrium with the gene(s) in the flanking region responsible for the modified phenotype, which can be localized in the region adjacent to the significant peak. Increasing the number of individuals constituting the bulks and generating more data might further enable this analysis to identify the candidate genes on the chromosomes 9 and 12 and elsewhere.

*FLOWERING LOCUS T* (*FT*) is a prime gene responsible for transition to floral meristem which in turn is regulated by multiple pathways responding to varied environmental stimuli (Andrés and Coupland 2012). The flax gene homologous to Arabidopsis *FT* is localized on chromosome 13, however, no significant associations were observed in this study. The possible reason could be the presence of a different regulatory gene substituting the function of *FT* because of the long evolutionary divergence between Arabidopsis and flax (106 million years ago - http://www.timetree.org/; Hedges et al. 2006).

As 'RE2' accession was derived from the treatment of 'Royal' using the hypomethylating chemical 5-AzaC, Whole Genome Bisulfite Sequencing (WGBS) data was generated to identify variation in methylation patterns and its potential association with early flowering phenotype in 'RE2' was examined. The WGBS data of five lines from each of early flowering and late flowering distributional extremes were pooled into two bulks and the methylation status of the cytosines in the two bulks were identified. The significant differentially methylated regions were unraveled using a Fisher's exact test. A total of 127 significant DMRs distributed across the genome were identified, and interestingly, 35 DMRs were present on chromosome 12 (Chapter 6). The DNA methylation differences between accessions can have significant impact on gene expression. Variation in methylation levels in the regulatory regions of the genome modify gene expression qualitatively while the impact of gene body methylation is still being explored (reviewed in Niederhuth and Schmitz 2017). Though the significant genomic regions as identified by QTLseqr, and the cluster of DMRs detected by the novel methylation analysis do not overlap, the co-localization of potential controlling regions on same chromosome suggesting the possible interaction of these two non-overlapping regions at the chromatin level by three-dimensional looping. The distal transcriptional regulators have been reported to be brought into contact with the genes by chromatin looping, influencing the expression (Liu and Weigel 2015; Grob and Grossniklaus 2017).

The conclusions from the study were:

i. The phenotypic evaluation of the 656 ‘Royal’ x ‘RE2’ recombinant inbred lines indicated that days to- start of flowering and full flowering were traits with moderate heritability.

ii. In the literature, there are no reports of point mutations induced by 5-AzaC. However, we cannot rule out the potential of 5-AzaC to cause point mutations without extensive investigation. The SNP observed could have arisen either from action of 5-AzaC or natural background mutations accumulated over generations. The presence of majority of SNP specific to the early flowering bulk in the upstream, downstream and intergenic region suggested the need for further detailed investigation on effect of variants in regulatory region.

iii. The genes in the significant region identified by NGS based BSA analysis with QTLseqr on chromosomes 9 and 12 were not directly associated with flowering time genes, based on homology studies. Hence, flax specific annotation of these genes, and that in the adjacent region and their validation will give further insights.

iv. Differentially methylated regions observed between early- and late flowering bulks owing to 5-AzaC treatment signify that there can be a possible epigenetic variation underlying the trait and hence, the role of epialleles in governing the early flowering phenotype needs to be further explored.

Hence, this study has laid the foundation for mining the epiallelic variation in ‘Royal’ x ‘RE2’ *epi*RIL population, with potential application in genomics-assisted breeding of flax.

The null hypothesis of genetic variation conditioning the early flowering phenotype in ‘RE2’ could not be rejected because of the lack of evidence. However, between genomic regions on chromosomes 9 and 12 having significant association with flowering time, the latter is found to co-localize with epigenetic variation, suggesting the potential link between genetic and epigenetic variants, which need further investigation, still leaving the basis of the early flowering trait unclear.

Further analysis to understand the mechanism and gene regulatory network (GRN; Gupta and Tsiantis 2018) underlying the early flowering phenotype would involve RNA sequencing (RNAseq) of the bulks in addition to the parents ‘Royal’ and ‘RE2’. The differential expression of the potential candidate genes in varied tissue-samples collected from RILs across time points will uncover the over- or under- expression of the genes which result in earlier transition to the reproductive phase. Intricate modification to a gene in its coding- or regulatory region, might alter both the qualitative- (splice variants) and quantitative (transcript abundance) aspects of gene expression, and can provide greater opportunities for generating variability in polygenic traits including flowering time (Rodríguez-Leal et al. 2017; Scheben and Edwards 2018). Once the major candidate genes underlying the polygenic traits were identified, precise editing can be made using the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 technology (reviewed in Knott and Doudna 2018) and hence, can be deployed for validating gene function (Fernandez i Marti and Dodd 2018).

Multiple levels of gene regulatory mechanisms have been studied in plants, and one among which is the orientation of the gene in the three-dimensional domain in the nucleus and the resulting interactions (Sotelo-Silveria et al. 2018). For instance, *FLOWERING LOCUS C* (*FLC*), a repressor of the key flowering time gene *FLOWERING LOCUS T* (*FT*), is under epigenetic regulation, in which, the three-dimensional looping in *FLC* is lost during vernalization leading to its silencing and subsequent upregulation of *FT* (Whittaker and Dean 2017). The chromatin loop formation as the mechanistic basis of repression of the *WUSCHEL* (*WUS*) gene, responsible for the determinate habit in Arabidopsis, through the interaction of *AGAMOUS* (*AG*) transcription factor with *TERMINAL FLOWER 2* (*TFL2*) was reported recently (Guo et al. 2018). In future experiments, it would be worthwhile to investigate the three-dimensional interactions of the genomic regions underlying the early flowering trait by using novel methodologies including chromatin conformation capture (3C; Dekker et al. 2002) and Hi-C (Lieberman-Aiden et al. 2009; reviewed in Doğan and Liu 2018), perhaps associating the variation in methylation (DMRs) with cis-regulated gene showing expression differences.

In several instances, the phenotypic changes induced by epigenetic variation is the result of modified methylation state of the genome and consequent activation of transposable elements which are silenced by hypermethylation, leading to new insertion sites (Seymour and Becker 2017). The contribution of cis-regulatory elements from the transposon fragments were found to cause varied expression of neighbouring genes (Hirsch and Springer 2017; Dubin et al. 2018). Interestingly, a retroelement insertion in the promoter region of winter wheat *VERNALIZATION3* (*VRN3*) lead to early flowering (Yan et al. 2006), and the insertion of a retrotransposon ~600 bp upstream of the photoperiod sensitive *HEADING DATE 1* (*Hd1*) caused delayed flowering in rice (Hori et al. 2016). Suggesting that further detailed assessment is required for presence/absence variation of insertions and deletions (InDels) applying strategies like Indel-seq (Singh et al.2017) on the RIL population.

The epigenetic marks including DNA methylation and histone modifications, mainly involved in maintaining genome stability (Ito and Kakutani 2014; Underwood and Martienssen 2015), can be inherited meiotically and mitotically in plant genomes (Niederhuth and Schmitz 2014). The alteration in methylation patterns and the resulting changes in gene expression often exhibit transgenerational inheritance (Quadrana and Colot 2016; Hosaka and Kakutani 2018) and hence, generation of epialleles controlling novel traits and *epi*RIL population became a reality (Johannes et al.2009; Brocklehurst et al. 2018). Beyond DNA nucleotide polymorphism-based SNP markers, DMR-dependent markers have been already developed in both mammalian- (Kim et al.2018) and plant (Ong-Abdullah et al. 2015) systems, which can also be developed in flax to tag the epigenetic differences underlying phenotypic variation. Flax is a good model organism to explore these variations because of the small genome size (~373 Mb - You et al. 2018) and the currently available genomic resources. Furthermore, inferences from this study can be tested in the other early flowering lines (‘RE1’ and ‘RE3’) of ‘Royal’ flax.

# References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamouna, S., and Terauchi, R. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnol* 30: 174-178.

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., Ichinoki, H., Notaguchi, M., Goto, K., and Araki, T. (2005). FD, a bZIP protein mediating signals from the floral pathway integrator *FT* at the shoot apex. *Science* 309: 1052-1056.

Akagi, T., Henry, I. M., Kawai, T., Comai, L., and Tao, R. (2016). Epigenetic regulation of the sex determination gene *MeGI* in polyploid persimmon. *Plant Cell* 28: 2905-2915.

Akimoto, K., Katakami, H., Kim, H. J., Ogawa, E., Sano, C. M., Wada, Y., and Sano, H. (2007). Epigenetic inheritance in rice plants. *Ann Bot* 100: 205-217.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.

Amado, L., Abranches, R., Neves, N., and Viegas, W. (1997). Development-dependent inheritance of 5-azacytidine-induced epimutations in triticale: analysis of rDNA expression patterns. *Chromosome Res* 5: 445-450.

Amoah, S., Kurup, S., Lopez, C. M. R., Welham, S. J., Powers, S. J., Hopkins, C. J., Wilkinson, M. J., and King, G. J. (2012). A hypomethylated population of *Brassica rapa* for forward and reverse epi-genetics. *BMC Plant Biol* 12: 1.

Andres, F. and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nature Rev Genet* 13: 627-639.

Atton, M. (1989). Flax culture from flower to fabric. The Ginger Press, Owen Sound, Ontario.

Aukerman, M. J., Lee, I., Weigel, D., and Amasino, R. M. (1999). The Arabidopsis flowering-time gene *LUMINIDEPENDENS* is expressed primarily in regions of cell proliferation and encodes a nuclear protein that regulates *LEAFY* expression. *Plant J* 18(2): 195-203.

Auld, D. L., Bettis, B. L., Crock, J. E., & Kephart, K. D. (1988). Planting date and temperature effects on germination, emergence, and seed yield of chickpea. *Agron J* 80(6): 909-914.

Barabaschi, D., Tondelli, A., Desiderio, F., Volante, A., Vaccino, P., Valè, G., and Cattivelli, L. (2016). Next generation breeding. *Plant Sci* 242: 3-13.

Bateson, W., Saunders, E. R., and Punnett, R. C. (1909). Experimental studies in the physiology of heredity. *Zeitschrift für Induktive Abstammungs-und Vererbungslehre*, 2(1): 17-19.

Bazakos, C., Hanemian, M., Trontin, C., Jiménez-Gómez, J. M., and Loudet, O. (2017). New strategies and tools in quantitative genetics: How to go from the phenotype to the genotype. *Annu Rev Plant Biol* 68:435-455.

Belchev, I., Tchorbadjieva, M., and Pantchev, I. (2004). Effect of 5-azacytidine on callus induction and plant regeneration potential in anther culture of wheat (*Triticum aestivum* L.). *Bulg J Plant Physiol* 30: 45-50.

Berry, S., and Dean, C. (2015). Environmental perception and epigenetic memory: mechanistic insight through *FLC*. *Plant J* 83(1): 133-148.

Bevan, M. W., Uauy, C., Wulff, B. B., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature* 543(7645): 346.s

Bewick, A. J., and Schmitz, R. J. (2017). Gene body DNA methylation in plants. *Curr Opin Plant Biol* 36, 103-110.

Bloomer, R. H., and Dean, C. (2017). Fine-tuning timing: natural variation informs the mechanistic basis of the switch to flowering in *Arabidopsis thaliana*. *J Exp Bot* 68(20): 5439-5452.

Blümel, M., Dally, N., and Jung, C. (2015). Flowering time regulation in crops - what did we learn from Arabidopsis? *Curr Opin Biotechnol* 32: 121-129.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114-2120.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32(3): 314-331.

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2015). FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res* 44: D1167-1171.

Bouché, F., Woods, D. P., and Amasino, R. M. (2017). Winter memory throughout the plant kingdom: different paths to flowering. *Plant Physiol* 173(1): 27-35.

Bowman, J. L., Smyth, D. R., and Meyerowitz, E. M. (1991). Genetic interactions among floral homeotic genes of Arabidopsis. *Development* 112: 1-20.

Boyko, A., and Kovalchuk, I. (2013). Epigenetic modifications in plants under adverse conditions: agricultural applications. In *Plant Acclimation to Environmental Stress* (pp. 233-267). Springer, New York.

Breuil-Broyer, S., Trehin, C., Morel, P., Boltz, V., Sun, B., Chambrier, P., Ito, T., and Negrutiu, I. (2016). Analysis of the Arabidopsis superman allelic series and the interactions with other genes demonstrate developmental robustness and joint specification of male–female boundary, flower meristem termination and carpel compartmentalization. *Ann Bot* 117: 905-923.

Brocklehurst, S., Watson, M., Carr, I. M., Out, S., Heidmann, I., and Meyer, P. (2018). Induction of epigenetic variation in Arabidopsis by over-expression of DNA *METHYLTRANSFERASE1* (*MET1*). *PloS One* 13(2): e0192170.

Bu, F., Chen, H., Shi, Q., Zhou, Q., Gao, D., Zhang, Z., and Huang, S. (2016). A major quantitative trait locus conferring subgynoecy in cucumber. *Theor Appl Genet* 129: 97-104.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill1, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T, R., Romay, M. C., Romero, S., Salvo1, S., Villeda, H. S., da Silva, H. S., Sun, Q., Tian, F., Upadyayula, N., Ware1, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., McMullen M. D. (2009). The genetic architecture of maize flowering time. *Science* 325(5941): 714-718.

Bueckert, R. A., and Clarke, J. M. (2013). Annual crop adaptation to abiotic stress on the Canadian prairies: Six case studies. *Can J Plant Sci* 93(3): 375-385.

Candela H, Casanova-Sáez R, Micol JL (2014) Getting started in mapping-by-sequencing. *J Integr Plant Biol 57*: 606–612. 10.1111/jipb.12305

Chan, S. W. L., Zilberman, D., Xie, Z., Johansen, L. K., Carrington, J. C., and Jacobsen, S. E. (2004). RNA silencing genes control de novo DNA methylation. *Science* 303(5662): 1336-1336.

Chandler, C. H., Chari, S., and Dworkin, I. (2013). Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends Genet* 29(6): 358-366.

Chanvivattana, Y., Bishopp, A., Schubert, D., Stock, C., Moon, Y. H., Sung, Z. R., and Goodrich, J. (2004). Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. *Development* 131: 5263-5276.

Choi, K., and Henderson, I. R. (2015). Meiotic recombination hotspots - a comparative view. *Plant J* 83(1): 52-61.

Choi, K., Kim, J., Hwang, H. J., Kim, S., Park, C., Kim, S. Y., and Lee, I. (2011). The FRIGIDA complex activates transcription of *FLC,* a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors. *Plant Cell* doi: https://doi.org/10.1105/tpc.110.075911.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2): 80-92.

Claus, R., and Lübbert, M. (2003). Epigenetic targets in hematopoietic malignancies. Oncogene 22(42): 6489-6496.

Cockram, J., and Mackay, I. (2018). Genetic mapping populations for conducting high-resolution trait mapping in plants. In *Advances in Biochemical Engineering/Biotechnology vol 164* (pp. 109–138). Springer, Cham.

Colombo, L., Franken, J., Koetje, E., van Went, J., Dons, H. J., Angenent, G. C., and van Tunen, A. J. (1995). The petunia *MADS* box gene *FBP11* determines ovule identity. *Plant Cell* 7: 1859-1868.

Colomé-Tatché, M., Cortijo, S., Wardenaar, R., Morgado, L., Lahouze, B., Sarazin, A., Etcheverry, M., Martin, A., Feng, S., Duvernois-Berthet, E., Labadie, K., Wincker, P., Jacobsen, S. E., Jansen, R. C., Colot, V., and Johannes, F. (2012). Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci U S A* 109(40): 16240-16245.

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., Giakountis, A., Farrona, S., Gissot, L., Turnbull, C., and Coupland, G. (2007). FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. *Science* 316: 1030-1033.

Corrales, A. R., Carrillo, L., Lasierra, P., Nebauer, S. G., Dominguez-Figueroa, J., Renau-Morata, B., Pollmann, S., Granell, A., Molina, R., V., Vicente Carbajosa, J., and Medina, J. (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(5): 748-764.

Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J., Wincker, P., Roudier, F., Jansen, R. C., Colot, V., and Johannes, F., (2014). Mapping the epigenetic basis of complex traits. *Science* 343(6175): 1145-1148.

Cullis, C. A. (2007). Flax. In: Kole, C. (editor) Oilseeds. Springer Berlin Heidelberg. pp. 275-295.

Daba, K., Deokar, A., Banniza, S., Warkentin, T. D., and Tar'an, B. (2016). QTL mapping of early flowering and resistance to ascochyta blight in chickpea. *Genome* 59(6): 413-425.

Darapuneni, M. K., Morgan, G. D., Ibrahim, A. M., and Duncan, R. W. (2014). Effect of vernalization and photoperiod on flax flowering time. *Euphytica* 195(2): 279-285.

Das, S., Singh, M., Srivastava, R., Bajaj, D., Saxena, M. S., Rana, J. C., Bansal, K. C., Tyagi, A. K., and Parida, S. K. (2016). mQTL-seq delineates functionally relevant candidate gene harbouring a major QTL regulating pod number in chickpea. *DNA Res* 23: 53-65.

Das, S., Upadhyaya, H.D., Bajaj, D., Kujur, A., Badoni, S., Kumar, V., Tripathi, S., Gowda, C.L., Sharma, S., Singh, S. and Tyagi, A.K., and Parida, S. K. (2015). Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Res* 22: 193-203.

Dash, P. K., Cao, Y., Jailani, A. K., Gupta, P., Venglat, P., Xiang, D., Rai, R., Sharma, R., Thirunavukkarasu, N., Abdin, M. Z., Yadava, D. K., Singh N. K., Singh, J., Selvaraj, G., Deyholos, M., Kumar, P. A., and Datla, R. (2014). Genome-wide analysis of drought induced gene expression changes in flax (*Linum usitatissimum*). *GM crops Food* 5(2): 106-119.

de Massy, B. (2013). Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annu Rev Genet* 47: 563-599.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295(5558): 1306-1311.

Dennis, E. S., and Peacock, W. J. (2007). Epigenetic regulation of flowering. *Curr Opin Plant Biol* 10(5): 520-527.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet* 43(5): 491.

Diederichsen, A., and Richards, K. (2003). Cultivated flax and the genus *Linum* L. In: Muir, A. D and Westcott, N. D. (editors) *Flax: the genus Linum*. Taylor and Francis. pp-22-54.

Dillman, A. C., and Hopper, T. H. (1943). Effect of climate on the yield and oil content of flaxseed and on the iodine number of linseed oil. US Department of Agriculture. Technical bulletin. 844.

Dinant, S., Clark, A. M., Zhu, Y., Vilaine, F., Palauqui, J. C., Kusiak, C., and Thompson, G. A. (2003). Diversity of the superfamily of phloem lectins (Phloem protein 2) in angiosperms. *Plant Physiol* 131(1): 114-128.

Doğan, E. S., and Liu, C. (2018). Three-dimensional chromatin packing and positioning of plant genomes. *Nature Plants* 4: 521–529.

Doitsidou, M., Jarriault, S., and Poole, R. J. (2016). Next-generation sequencing-based approaches for mutation mapping and identification in Caenorhabditis elegans. *Genetics* 204(2): 451-474.

Dribnenki, J. C. P., McEachern, S. F., Chen, Y., Green, A. G., and Rashid, K. Y. (2005). 2126 low linolenic flax. *Can J Plant Sci* 85: 155-157.

Dubin, M. J., Scheid, O. M., and Becker, C. (2018). Transposons: a blessing curse. *Curr Opin Plant Biol:* 42: 23-29.

Duchêne, E., Butterlin, G., Dumas, V., and Merdinoglu, D. (2012) Towards the adaptation of grapevine varieties to climate change: QTLs and candidate genes for developmental stages. *Theor Appl Genet* 124: 623-635.

Duguid, S. D. (2009). Flax. In: Vollmann, J. and Rajcan, I. (editors). Oil Crops. Springer New York. pp. 233-255.

Eggermont, L., Verstraeten, B., and Van Damme, E. J. (2017). Genome-wide screening for lectin motifs in *Arabidopsis thaliana*. *Plant Genome* 10(2) doi: 10.3835/plantgenome2017.02.0010

Ehrenreich, I. M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J. A., Gresham, D., Caudy, A. A., and Kruglyak, L. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* 464(7291): 1039-1042.

Fechter, I., Hausmann, L., Zyprian, E., Daum, M., Holtgräwe, D., Weisshaar, B., and Töpfer, R. (2014). QTL analysis of flowering time and ripening traits suggests an impact of a genomic region on linkage group 1 in Vitis. *Theor Appl Genet* 127: 1857-1872.

Feng, S., and Jacobsen, S. E. (2011). Epigenetic modifications in plants: an evolutionary perspective. *Curr Opin Plant Biol* 14(2): 179-186.

Fernandez i Marti, A., and Dodd, R. S. (2018). Using CRISPR as a gene editing tool for validating adaptive gene function in tree landscape genomics. *Front Ecol Evol 6*: 76.

Fieldes, M. A. (1994). Heritable effects of 5-azacytidine treatments on the growth and development of flax (*Linum usitatissimum*) genotrophs and genotypes. *Genome* 37: 1-11.

Fieldes, M. A., and Amyot, L. M. (1999). Epigenetic control of early flowering in flax lines induced by 5-azacytidine applied to germinating seed. *J Heredity* 90: 199-206.

Fieldes, M. A., and Harvey, C. G. (2004). Differences in developmental programming and node number at flowering in the 5-Azacytidine-induced early flowering flax lines and their controls. *Int J Plant Sci* 165: 695-706.

Finnegan, E. J., Peacock, W. J., and Dennis, E. S. (1996). Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. *Proc Natl Acad Sci U S A* 93(16): 8449-8454.

Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytol* 183(3): 557-564.

Fornara, F., and Coupland, G. (2009). Plant phase transitions make a *SPL*ash. *Cell* 138: 625-627.

Fornara, F., Panigrahi, K. C., Gissot, L., Sauerbrunn, N., Rühl, M., Jarillo, J. A., and Coupland, G. (2009). Arabidopsis DOF transcription factors act redundantly to reduce *CONSTANS* expression and are essential for a photoperiodic flowering response. *Dev Cell* 17: 75-86.

Fowler, S., Lee, K., Onouchi, H., Samach, A., Richardson, K., Morris, B., Coupland, G., and Putterill, J. (1999). *GIGANTEA*: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains. *EMBO J* 18: 4679-4688.

Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., Van Der Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K. B., and Tanksley, S. D. (2000). *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289(5476): 85-88.

Gan, Y. T., Miller, P. R., Liu, P. H., Stevenson, F. C., & McDonald, C. L. (2002). Seedling emergence, pod development, and seed yields of chickpea and dry pea in a semiarid environment. *Can J Plant Sci* 82(3): 531-537.

Garner, W. W., and Allard, H. A. (1920). Effect of the relative length of day and night and other factors of the environment on growth and reproduction in plants 1. *Monthly Weather Review* 48: 415.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6): 333.

Grandbastien, M. A. (2015). LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochimica et Biophysica Acta (BBA)-Gene Regulat Mech* 1849: 403-416.

Greenham, K., and McClung, C. R. (2015). Integrating circadian dynamics with physiological processes in plants. *Nature Rev Genet* 16: 598-610.

Griffin, P. T., Niederhuth, C. E., and Schmitz, R. J. (2016). A comparative analysis of 5-azacytidine and zebularine induced DNA demethylation. *G3-Genes Genom Genet* doi: https://doi.org/10.1534/g3.116.030262.

Grob, S., and Grossniklaus, U. (2017). Chromosome conformation capture-based studies reveal novel features of plant nuclear architecture. *Curr Opin Plant Biol* 36: 149-157.

Gubbels, G. H., Bonner, D. M., and Kenaschuk, E. O. (1994). Effect of frost injury on quality of flax seed. *Can J Plant Sci* 74: 331-333.

Guerriero, G., Mangeot-Peter, L., Legay, S., Behr, M., Lutts, S., Siddiqui, K. S., and Hausman, J. F. (2017). Identification of fasciclin-like arabinogalactan proteins in textile hemp (*Cannabis sativa* L.): In silico analyses and gene expression patterns in different tissues. *BMC Genomics* 18(1): 741.

Guo, B., Wei, Y., Xu, R., Lin, S., Luan, H., Lv, C., Zhang, X., Song, X., and Xu, R. (2016). Genome-wide analysis of APETALA2/ethylene-responsive factor (AP2/ERF) gene family in barley (*Hordeum vulgare* L.). *PloS one* 11(9): e0161322.

Guo, L., Cao, X., Liu, Y., Li, J., Li, Y., Li, D., Zhang, K., Gao, C., Dong, A. and Liu, X. (2018). A chromatin loop represses *WUSCHEL* expression in Arabidopsis. *Plant J* 94(6): 1083-1097.

Gupta, M. D., and Tsiantis, M. (2018). Gene networks and the evolution of plant morphology. *Curr Opin Plant Biol* 45: 82-87.

Gusta, L. V., Johnson, E. N., Nesbitt, N. T., and Kirkland, K. J. (2004). Effect of seeding date on canola seed quality and seed vigour. *Can J Plant Sci* 84(2): 463-471.

Hall, L. M., Booker, H., Siloto, R. M. P., Jhala, A. J., and Weselake, R. J. (2016). Flax (*Linum usitatissimum* L.). In: McKeon, T.et al. (editors). Industrial Oil Crops. Elsevier. pp-157-194.

Han, Y., Lv, P., Hou, S., Li, S., Ji, G., Ma, X., Du, R., and Liu, G. (2015). Combining next generation sequencing with bulked segregant analysis to fine map a stem moisture locus in sorghum (*Sorghum bicolor* L. Moench). *PloS One* 10: e0127065.

He, Y., and Li, Z. (2018). Epigenetic Environmental Memories in Plants: Establishment, Maintenance, and Reprogramming. *Trends Genet* doi: https://doi.org/10.1016/j.tig.2018.07.006

Healey, A., Furtado, A., Cooper, T., and Henry, R. J. (2014). Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* 10(1): 21.

Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22: 2971-2972.

Hepworth, J., and Dean, C. (2015). FLOWERING LOCUS C's lessons: conserved chromatin switches underpinning developmental timing and adaptation. *Plant Physiol* 168(4): 1237-1245.

Heslop-Harrison, J. S. (1990). Gene expression and parental dominance in hybrid plants. *Dev (Suppl):* 21-28.

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim Biophys Acta-Gene Regulatory Mechanisms* 1860(1): 157-165.

Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W., and Schmitz, R. J. (2017). Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol* 18(1): 155.

Holland, J. B. (2007). Genetic architecture of complex traits in plants. *Current Opin Plant Biol* 10: 156-161.

Hori, K., Matsubara, K., Uga, Y., and Yano, M. (2016). A novel Tos17 insertion upstream of Hd1 alters flowering time in rice. *Plant Breeding* 135(5): 588-592.

Horváth, E., Szalai, G., Janda, T., Páldi, E., Rácz, I., and Lásztity, D. (2003). Effect of vernalisation and 5-azacytidine on the methylation level of DNA in wheat (*Triticum aestivum L., cv. Martonvásár* 15). *Plant Sci* 165: 689-692.

Hosaka, A., and Kakutani, T. (2018). Transposable elements, genome evolution and transgenerational epigenetic variation. *Curr Opin Genet Dev* 49: 43-48.

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., Leung, H., Varshney, R. K., and Cavanagh, C. R. (2015). MAGIC populations in crops: current status and future prospects. *Theor Appl Genet* 128(6): 999-1017.

Illa-Berenguer, E., Van Houten, J., Huang, Z., and van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theor Appl Genet* 128: 1329-1342.

Ito, H., and Kakutani, T. (2014). Control of transposable elements in *Arabidopsis thaliana*. *Chromosome Res* 22(2): 217-223.

Janoušek, B., Široký, J., and Vyskot, B. (1996). Epigenetic control of sexual phenotype in a dioecious plant, *Melandrium album*. *Mol Gen Genet* 250: 483-490.

Jeong, H. J., Yang, J., Yi, J., and An, G. (2015). Controlling flowering time by histone methylation and acetylation in arabidopsis and rice. *J Plant Biol* 58(4): 203-210.

Jin, L. G., Li, H., and Liu, J. Y. (2010). Molecular characterization of three ethylene responsive element binding factor genes from cotton. *J Integr Plant Biol* 52(5): 485-495.

Johannes, F., Colot, V., and Jansen, R. C. (2008). Epigenome dynamics: a quantitative genetics perspective. *Nature Rev Genet* 9: 883-890.

Johannes, F., Porcher, E., Teixeira, F. K., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., Audigier, P., Bouchez, D., Dillmann, C., Guerche, P., Hospital, F., and Colot, V. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* 5(6): e1000530.

Johansson, M., and Staiger, D. (2014). *SRR1* is essential to repress flowering in non-inductive conditions in *Arabidopsis thaliana*. *J Exp Bot* 65(20): 5811-5822.

Jun, X. U., Wang, X. Y., and Guo, W. Z. (2015). The cytochrome P450 superfamily: key players in plant development and defense. *Journal Integr Agr* 14(9): 1673-1686.

Kadambari, G., Vemireddy, L. R., Srividhya, A., Nagireddy, R., Jena, S. S., Gandikota, M., Patil, S., Veeraghattapu, R., Deborah, D. A. K., Reddy, G. E., Shake, M., Dasari, A., Ramanarao, P. V., Durgarani, Ch. V., Neeraja, C. N., Siddiq, E. A., Sheshumadhav, M. (2018). QTL-Seq-based genetic analysis identifies a major genomic region governing dwarfness in rice (*Oryza sativa* L.). *Plant Cell Rep* 37(4): 677-687.

Kang, Y. J., Bae, A., Shim, S., Lee, T., Lee, J., Satyawan, D., Kim, M., Y., and Lee, S. H. (2017). Genome-wide DNA methylation profile in mungbean. *Sci Rep 7*: 40503

Kawakatsu, T., Huang, S. S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery, J. R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C., Wang, C., Bemm, F., Becker, C., O'Niel, R., O'Malley, R. C., and Ecker, J.R. (2016a). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 166(2): 492-505.

Kawakatsu, T., Stuart, T., Valdes, M., Breakfield, N., Schmitz, R. J., Nery, J. R., Urich, M. A., Han, X., Lister, R., Benfey, P. N., and Ecker, J. R. (2016b). Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nature Plants* 2(5): 16058.

Keurentjes, J. J., Bentsink, L., Alonso-Blanco, C., Hanhart, C. J., Blankestijn-De Vries, H., Effgen, S., Vreugdenhil, D., and Koornneef, M. (2007). Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics* 175(2): 891-905.

Kim, H. R., Wang, X., and Jin, P. (2018). Developing DNA methylation-based diagnostic biomarkers. *J Genet Genomics* 45: 87-97

Kim, K. D., El Baidouri, M., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grinwood, J., and Scott, A. J. (2015). A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant Physiol* 168: 1433-1447.

Kim, K. D., El Baidouri, M., and Jackson, S. A. (2014). Accessing epigenetic variation in the plant methylome. *Briefings in Funct Genomics* 13(4): 318-327.

Kim, S. Y., and Michaels, S. D. (2006). *SUPPRESSOR OF FRI 4* encodes a nuclear localized protein that is required for delayed flowering in winter-annual Arabidopsis. *Development* 133(23): 4699-4707.

Kim, S., Choi, K., Park, C., Hwang, H. J., and Lee, I. (2006). *SUPPRESSOR OF FRIGIDA4*, encoding a C2H2-Type zinc finger protein, represses flowering by transcriptional activation of Arabidopsis *FLOWERING LOCUS C*. *Plant Cell* 18(11): 2985-2998.

King, G. J., Amoah, S., and Kurup, S. (2010). Exploring and exploiting epigenetic variation in crops. *Genome* 53(11): 856-868.

King, M., Altdorff, D., Li, P., Galagedara, L., Holden, J., and Unc, A. (2018). Northward shift of the agricultural climate zone under 21st-century global climate change. *Sci Rep* 8(1): 7904.

Kirk, J. T. (1994). Light and photosynthesis in aquatic ecosystems. Cambridge university press.

Knott, G. J., and Doudna, J. A. (2018). CRISPR-Cas guides the future of genetic engineering. *Science* 361(6405): 866-869.

Kondo, H., Shiraya, T., Wada, K. C., and Takeno, K. (2010). Induction of flowering by DNA demethylation in *Perilla frutescens* and *Silene armeria*: Heritability of 5-azacytidine-induced effects and alteration of the DNA methylation state by photoperiodic conditions. *Plant Sci* 178: 321-326.

Kong, L., Lu, S., Wang, Y., Fang, C., Wang, F., Nan, H., Su, T., Li, S., Zhang, F., Li, X., Zhao, X., Yuan, X., Liu, B., and Kong, F. (2018). Quantitative trait locus mapping of flowering time and maturity in soybean using next-generation sequencing-based analysis. *Front Plant Sci* doi: 10.3389/fpls.2018.00995

Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. *Science* 312(5778): 1392-1396.

Koornneef, M., Hanhart, C. J., and Van der Veen, J. H. (1991). A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. *Mol Gen Genet* 229: 57-66.

Kosugi, S., Natsume, S., Yoshida, K., MacLean, D., Cano, L., Kamoun, S., and Terauchi, R. (2013). Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. *PLoS One* 8(10): e75402.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9): 1639-1645.

Kujur, A., Upadhyaya, H. D., Bajaj, D., Gowda, C. L. L., Sharma, S., Tyagi, A. K., and Parida, S. K. (2016). Identification of candidate genes and natural allelic variants for QTLs governing plant height in chickpea. *Sci Rep* 6: 27968.

Kumar, J., Gupta, D. S., Gupta, S., Dubey, S., Gupta, P., and Kumar, S. (2017a). Quantitative trait loci from identification to exploitation for crop improvement. *Plant Cell Rep* 36(8): 1187-1213.

Kumar, S. V., Lucyshyn, D., Jaeger, K. E., Alós, E., Alvey, E., Harberd, N. P., and Wigge, P. A. (2012). Transcription factor *PIF4* controls the thermosensory activation of flowering. *Nature* 484: 242-245.

Kumar, S., Hash, C. T., Nepolean, T., Satyavathi, T. S., Singh, G., Mahendrakar, M. D., Yadav, R. S., and Srivastava, R. K. (2017b). Mapping QTLs Controlling Flowering Time and Important Agronomic Traits in Pearl Millet [*Pennisetum glaucum* (L.) R. Br.]. *Front Plant Sci* 8: 1731.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat methods* 9(4): 357.

Lata, C., and Prasad, M. (2011). Role of DREBs in regulation of abiotic stress responses in plants. *J Exp Bot 62*(14): 4731-4748.

Law, J. A., and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev Genet* 11: 204-220.

Lee, I., Aukerman, M. J., Gore, S. L., Lohman, K. N., Michaels, S. D., Weaver, L. M., John, M. C., Feldmann, K. A., and Amasino, R. M. (1994). Isolation of LUMINIDEPENDENS: a gene involved in the control of flowering time in Arabidopsis. *The Plant Cell*, *6*(1), 75-83.

Lenhard, M., Bohnert, A., Jürgens, G., and Laux, T. (2001). Termination of stem cell maintenance in Arabidopsis floral meristems by interactions between WUSCHEL and AGAMOUS. *Cell* 105: 805-814.

Lev-Yadun, S., Gopher, A., and Abbo, S. (2000). The cradle of agriculture. *Science* 288: 1602-1603.

Li, D., Wang, X., Zhang, X., Chen, Q., Xu, G., Xu, D., Wang, C., Liang, Y., Wu, L., Huang, C., Tian, J., Wu, Y., and Tian, F. (2016). The genetic architecture of leaf number and its genetic relationship to flowering time in maize. *New Phytol* 210(1): 256-268.

Li, H., and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25: 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics 25*(16): 2078-2079.

Li, H., Wang, Y., Wu, M., Li, L., Li, C., Han, Z., Yuan, J., Chen, C., Song, W., and Wang, C. (2017b). Genome-wide identification of AP2/ERF transcription factors in cauliflower and expression profiling of the ERF family under salt and drought stresses. *Front Plant Sci* 8: 946.

Li, S. F., Zhang, G. J., Yuan, J. H., Deng, C. L., Lu, L. D., and Gao, W. J. (2015). Effect of 5-azaC on the growth, flowering time and sexual phenotype of spinach. *Russian J Plant Physiol*. 62: 670-675.

Li, W, Zhu Z, Chern M, Yin, J., Yang, C., Ran, L., Cheng, M., He, M., Wang, K., Wang, J., Zhou, X., Zhu, X., Chen, Z., Wang, J., Zhao, W., Ma, B., Qin, P., Chen, W., Wang, Y., Liu, J., Wang, W., Wu, X., Li, P., Wang, J., Zhu, L., Li, S., and Chen, X. (2017a) A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell* 170:114–126.e15

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., Zhang, G., Zheng, X., Zhang, H., Zhang, S., Li, Q., Luo, R., Yu, C., Yu, J., Sun, J., Zou, X., Cao, X., Xie, X., Wang, J., and Wang, W. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 13(1): 300.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstien, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E.S., Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950): 289-293.

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

Lindroth, A. M., Cao, X., Jackson, J. P., Zilberman, D., McCallum, C. M., Henikoff, S., and Jacobsen, S. E. (2001). Requirement of *CHROMOMETHYLASE3* for maintenance of CpXpG methylation. *Science* 292(5524): 2077-2080.

Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J. C., Doerge, R. W., Colot, V., and Martienssen, R. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430(6998): 471-476.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133(3): 523-536.

Liu, B., Watanabe, S., Uchiyama, T., Kong, F., Kanazawa, A., Xia, Z., Nagamatsu, A., Arai, M., Yamada, T., Kitamura, K., Masuta, C., Harada, K., and Abe, J. (2010). The soybean stem growth habit gene *dt1* is an ortholog of Arabidopsis *TERMINAL FLOWER1*. *Plant Physiol* 153(1): 198-210.

Liu, C., and Weigel, D. (2015). Chromatin in 3D: progress and prospects for plants. *Genome Biol* 16(1): 170.

Liu, H., Du, D., Guo, S., Xiao, L., Zhao, Z., Zhao, Z., Xing, X., Tang, G., Xu, L., Fu, Z., Yao, Y., and Duncan, W. (2016) QTL analysis and the development of closely linked makers for days to flowering in spring oilseed rape (*Brassica napus* L.). *Mol Breed* Doi: 10.1007/s11032-016-0477-8.

Liu, S., Yeh, C. T., Tang, H. M., Nettleton, D., and Schnable, P. S. (2012). Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PloS one* 7(5): e36406.

Long, Y., Xia, W., Li, R., Wang, J., Shao, M., Feng, J., King, G. J., and Meng, J. (2011). Epigenetic QTL mapping in *Brassica napus*. *Genetics* https://doi.org/10.1534/genetics.111.131615.

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y., and Huang, S. (2014). QTL-seq identifies an early flowering QTL located near *FLOWERING LOCUS T* in cucumber. *Theor Appl Genet* 127: 1491-1499.

Luo, Y. X., Luo, C. Y., Du, D. Z., Fu, Z., Yao, Y. M., Xu, C. C., and Zhang, H. S (2014) Quantitative trait analysis of flowering time in spring rapeseed (*B. napus* L.). *Euphytica* 200: 321-335.

Lyons, D. M., and Lauring, A. S. (2017). Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. *Mol Biol Evol* 34(12): 3205-3215.

Ma, Y., Dai, X., Xu, Y., Luo, W., Zheng, X., Zeng, D., Pan, Y., Lin, X., Liu, H., Zhang, D., Xiao, J, Guo, X., Xu, S., Niu, Y., Jin, J., Zhang, H., Xu, X., Li, L., Wang, W., Qian, Q., Ge, S., and Chong, K. (2015). *COLD1* confers chilling tolerance in rice. *Cell* 160(6): 1209-1221.

Mackay, T. F. (2001). The genetic architecture of quantitative traits. *Annu Rev Genet* 35(1): 303-339.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*10(8): 565.

MacMillan, C. P., Mansfield, S. D., Stachurski, Z. H., Evans, R., and Southerton, S. G. (2010). Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. *Plant J* 62(4): 689-703.

Magome, H., Yamaguchi, S., Hanada, A., Kamiya, Y., and Oda, K. (2004). *dwarf and delayed-flowering 1*, a novel Arabidopsis mutant deficient in gibberellin biosynthesis because of overexpression of a putative AP2 transcription factor. *The Plant J* 37(5): 720-729.

Magwanga, R. O., Lu, P., Kirungu, J. N., Lu, H., Wang, X., Cai, X., Zhou, Z., Zhang, Z., Salih, H., Wang, K., and Liu, F. (2018). Characterization of the late embryogenesis abundant (LEA) proteins family and their role in drought stress tolerance in upland cotton. *BMC Genet* 19(1): 6.

Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., and Seymour, G. B. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature Genet* 38(8): 948.

Mansfeld, B. N., and Grumet, R. (2018). QTLseqr: An R package for bulk segregant analysis with next-generation sequencing. *Plant Genome* 11:180006.

Mansur, L. M., Orf, J., and Lark, K. G. (1993). Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (*Glycine max* L. Merr.). *Theor Appl Genet* 86: 914-918.

Marfil, C. F., Asurmendi, S., and Masuelli, R. W. (2012). Changes in micro RNA expression in a wild tuber-bearing *Solanum* species induced by 5-Azacytidine treatment. *Plant Cell Rep* 31: 1449-1461.

Masumoto, H., Takagi, H., Mukainari, Y., Terauchi, R., and Fukunaga, K. (2016). Genetic analysis of *NEKODE1* gene involved in panicle branching of foxtail millet, *Setaria italica* (L.) P. Beauv., and mapping by using QTL-seq. *Mol Breed* 36: 1-8.

Matsuo, N., Fukami, K., and Tsuchiya, S. (2016). Effects of early planting and cultivars on the yield and agronomic traits of soybeans grown in southwestern Japan. *Plant Prod Sci* 19(3): 370-380.

McGregor, C. E., Waters, V., Vashisth, T., and Abdel-Haleem, H. (2014) Flowering time in watermelon is associated with a major quantitative trait locus on chromosome 3. *J Amer Soc Hort Sci* 139: 48-53.

McGregor, W. G. (1953). Varieties of linseed flax. Canada Department of Agriculture. Publication number: 884.

Mendel, G. (1866). Experiments on plant hybrids. *Negotiations of the naturforschenden association in Brunn* 4: 3-44.

Michaels, S. D., and Amasino, R. M. (1999). *FLOWERING LOCUS C* encodes a novel *MADS* domain protein that acts as a repressor of flowering. *Plant Cell* 11:949-956.

Michelmore, R. W., Paran, I., and Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A* 88: 9828-9832.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T. (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* 411: 212-214.

Miura, K., Agetsuma, M., Kitano, H., Yoshimura, A., Matsuoka, M., Jacobsen, S. E., and Ashikari, M. (2009). A metastable *DWARF1* epigenetic mutant affecting plant stature in rice. *Proc Natl Acad Sci U S A* 106(27): 11218-11223.

Mohd-Yusoff, N. F., Ruperao, P., Tomoyoshi, N. E., Edwards, D., Gresshoff, P. M., Biswas, B., and Batley, J. (2015). Scanning the effects of ethyl methanesulfonate on the whole genome of Lotus japonicus using secondgeneration sequencing analysis. *G3 (Bethesda)* 5 (4): 559-567.

Morell, P.L., Buckler, E.S. and Ross-Ibarra, J. (2012) Crop genomics: advances and applications. *Nat Rev Genet* 13: 85–96.

Morgan, T. H. (1910). Sex-limited inheritance in Drosophila. *Science* 32: 120-122.

Munsamy, A., Rutherford, R. S., Snyman, S. J., and Watt, M. P. (2013). 5-Azacytidine as a tool to induce somaclonal variants with useful traits in sugarcane (*Saccharum* spp.). *Plant Biotechnol Rep* 7: 489-502.

Najafi, S., Sorkheh, K., and Nasernakhaei, F. (2018). Characterization of the APETALA2/Ethylene-responsive factor (AP2/ERF) transcription factor family in sunflower. *Sci rep 8*(1): 11576.

Nakano, H., Kobayashi, N., Takahata, K., Mine, Y., and Sugiyama, N. (2016). Quantitative trait loci analysis of the time of floral initiation in tomato. *Sci Hort* 201: 199-210.

Nelson, D. R., Schuler, M. A., Paquette, S. M., Werck-Reichhart, D., and Bak, S. (2004). Comparative genomics of Rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol* 135(2): 756-772.

Ng, M., and Yanofsky, M. F. (2001). Activation of the Arabidopsis B Class Homeotic Genes by *APETALA1*. *Plant Cell* 13: 739-753.

Niederhuth, C. E., and Schmitz, R. J. (2014). Covering your bases: inheritance of DNA methylation in plant genomes. *Mol Plant* 7(3), 472-480.

Niederhuth, C. E., and Schmitz, R. J. (2017). Putting DNA methylation in context: from genomes to gene expression in plants. *Biochim Biophys Acta-Gene Regulatory Mechanisms* 1860(1): 149-156.

Nuttonson, M. Y. (1948). Some preliminary observations of phenological data as a tool in the study of photoperiodic and thermal requirements of various plant material. Chronica Botanica Company 29-143.

Ong-Abdullah, M., Ordway, J. M., Jiang, N., Ooi, S. E., Kok, S. Y., Sarpan, N., Azimi, N., Hashim, A. T., Ishak, Z., Rosli, S., K., Malike, F. A., Abu Bakar, N. A., Marjuni, M., Abdullah, N., Yaakub, Z., Amiruddin, M. D., Nookiah, R., Singh, R., Low, E. L., Chan, K., Azizi, N., Smith, S. W., Bacher, B., Budiman, M. A., Brunt, A. V., Wischmeyer, C., Beil, M., Hogan, M., Lakey, N., Lim, C., Arulandoo, X., Wong, C., Choo, C., Wong, W., Kwan, Y., Alwee, S.S.R.S., Sambanthamurthi, R., and Martienssen, R. A. (2015). Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 525(7570): 533.

Pandey, M.K., Khan, A.W., Singh, V.K., Vishwakarma, M.K., Shasidhar, Y., Kumar, V., Garg, V., Bhat, R.S., Chitikineni, A., Janila, P. and Guo, B., (2017). QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in groundnut (*Arachis hypogaea* L.). *Plant Biotechnol J* 15(8): 927-941.

Park, M. H., Suzuki, Y., Chono, M., Knox, J. P., and Yamaguchi, I. (2003). *CsAGP1*, a gibberellin-responsive gene from cucumber hypocotyls, encodes a classical arabinogalactan protein and is involved in stem elongation. *Plant Physiol* 131(3): 1450-1459.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335(6192): 721.

Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Curr Protoc Bioinformatics* 42(1): 3.1.1-3.1.8.

Pecinka, A., and Liu, C. H. (2014). Drugs for plant chromosome and chromatin research. *Cytogenet Genome Res* 143: 51-59.

Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E., and Yanofsky, M. F. (2000). B and C floral organ identity functions require *SEPALLATA MADS-box* genes. *Nature* 405: 200-203.

Peng, J., Xia, B., and Yi, C. (2016). Single-base resolution analysis of DNA epigenome via high-throughput sequencing. *Science China Life Sci* 59: 219-226.

Pires, N. D., and Grossniklaus, U. (2018). Identification of parent-of-origin-dependent QTLs using bulk-segregant sequencing (Bulk-Seq). In *Plant Chromatin Dynamics* (pp. 361-371). Humana Press, New York, NY.

Porebski, S., Bailey, L. G., and Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Report* 15(1): 8-15.

Posé, D., Verhage, L., Ott, F., Yant, L., Mathieu, J., Angenent, G. C., Immink, R. G. and Schmid, M. (2013). Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* 503: 414-417.

Poorter, H., Fiorani, F., Pieruschka, R., Wojciechowski, T., van der Putten, W. H., Kleyer, M., Schurr, U., and Postma, J. (2016). Pampered inside, pestered outside? Differences and similarities between plants growing in controlled conditions and in the field. *New Phytol* 212: 838-855.

Price, A. H. (2006). Believe it or not, QTLs are accurate! *Trends Plant Sci* 11(5): 213-216.

Quadrana, L., and Colot, V. (2016). Plant transgenerational epigenetics. *Annu Rev Genet* 50: 467-491.

Qüesta, J. I., Song, J., Geraldo, N., An, H., and Dean, C. (2016). Arabidopsis transcriptional repressor VAL1 triggers Polycomb silencing at FLC during vernalization. *Science* 353: 485-488.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.

R Core Team. (2018). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Austria, 2015.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 5(2): 94-100.

Raman, H., Raman, R., Eckermann, P., Coombes, N., Manoli, S., Zou, X., Edwards, D., Meng, J., Prangnell, R., Stiller, J., Batley, J., Luckett, D., Wratten, N., and Dennis, E. (2013). Genetic and physical mapping of flowering time loci in canola (*Brassica napus* L.). *Theor Appl Genet* 126(1): 119-132.

Roach, M. J., and Deyholos, M. K. (2007). Microarray analysis of flax (*Linum usitatissimum* L.) stems identifies transcripts enriched in fibre-bearing phloem tissues. *Mol Genet Genomics* 278(2): 149-165.

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(2): 470-480.

Royo, C., Torres-Pérez, R., Mauri, N., Diestro, N., Cabezas, J. A., Marchal, C., Lacombe, T., Ibáñez, J., Tornel, M., Carreño, J., Martínez-Zapater, J. M., and Carbonell-Bejerano, P. (2018). The major origin of seedless grapes is associated with a missense mutation in the MADS-box gene VviAGL11. *Plant Physiol* doi: 10.1104/pp.18.00259.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K.A., Meeley, R., Ananiev, E.V., Svitashev, S., Bruggemann, E., Li, B., Hainey, C.F., Rodvic, S., Zaina, G., Rafalski, J.-A., Tingey, S.V., Miao, G., Phillips, R.L., and Tuberosa, R. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci U S A* 104(27): 11376-11381.

Sano, H., Kamada, I., Youssefian, S., Katsumi, M., and Wabiko, H. (1990). A single treatment of rice seedlings with 5-azacytidine induces heritable dwarfism and undermethylation of genomic DNA. *Mol Gen Genet* 220: 441-447.

Sasaki, E., Frommlet, F., and Nordborg, M. (2017). The genetic architecture of the network underlying flowering time variation in *Arabidopsis thaliana*. *bioRxiv* 175430.

Sawa, M., Nusinow, D. A., Kay, S. A., and Imaizumi, T. (2007). *FKF1* and *GIGANTEA* complex formation is required for day-length measurement in *Arabidopsis*. *Science* 318: 261-265.

Sax, K. (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8: 552-560.

Scarpeci, T. E., Frea, V. S., Zanor, M. I., and Valle, E. M. (2016). Overexpression of *AtERF019* delays plant growth and senescence and improves drought tolerance in Arabidopsis. *J Exp Bot* 68(3): 673-685.

Schatz, M. C., Witkowski, J., and McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome Biol* 13(4): 243.

Scheben, A., and Edwards, D. (2018). Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced allelic series to optimize quantitative and qualitative traits. *Curr Opin Plant Biol* doi: 10.1016/j.pbi.2018.04.013.

Scheben, A., Batley, J., and Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol J* 15(2):149-161.

Schmitz, R. J. (2014). The secret garden-epigenetic alleles underlie complex traits. *Science* 343(6175): 1082-1083.

Schmitz, R. J., He, Y., Valdés-López, O., Khan, S. M., Joshi, T., Urich, M. A., Nery, R. J., Diers, B., Xu, D., Stacey, G., and Ecker, J. R. (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* doi: 10.1101/gr.152538.112

Schmitz, R. J., Hong, L., Michaels, S., and Amasino, R. M. (2005). FRIGIDA-ESSENTIAL 1 interacts genetically with FRIGIDA and FRIGIDA-LIKE 1 to promote the winter-annual habit of Arabidopsis thaliana. *Development* 132(24): 5471-5478.

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci* 16(5): 282-288.

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat Rev Genet* 15(10): 662-676.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen J.-E., Weigel, D., and Andersen, S. U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* 6:550–551.

Seymour, D. K., and Becker, C. (2017). The causes and consequences of DNA methylome variation in plants. *Curr Opin Plant Biol* 36: 56-63.

Seymour, D. K., and Becker, C. (2017). The causes and consequences of DNA methylome variation in plants. *Curr Opin Plant Biol* 36: 56-63.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature* 550(7676): 345.

Shirasawa, K., Hirakawa, H., Nunome, T., Tabata, S., and Isobe, S. (2016). Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato. *Plant Biotechnol J* 14(1): 51-60.

Shu, J., Liu, Y., Zhang, L., Li, Z., Fang, Z., Yang, L., Zhuang, M., Zhang, Y., and Lv, H. (2018). QTL-seq for rapid identification of candidate genes for flowering time in broccoli × cabbage. *Theor Appl Genet* Doi: 10.1007/s00122-017-3047-5

Simon, M., Loudet, O., Durand, S., Bérard, A., Brunel, D., Sennesal, F. X., Durand-Tardif, M., Pelletier, G., and Camilleri, C. (2008). Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics* 178(4): 2253-2264.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 15(2): 121.

Singh, V. K., Khan, A. W., Jaganathan, D., Thudi, M., Roorkiwal, M., Takagi, H., Garg, V., Kumar, V., Chitikineni, A., Gaur, P.M. and Sutton, T., Terauchi, R., Varshney, R. K. (2016). QTL-seq for rapid identification of candidate genes for 100-seed weight and root/total plant dry weight ratio under rainfed conditions in chickpea. *Plant Biotechnol J* 14(11): 2110-2119.

Singh, V. K., Khan, A. W., Saxena, R. K., Kumar, V., Kale, S. M., Sinha, P., Chitikineni, A., Pazhamala, L. T., Garg, V., Sharma, M., Kumar, C. V. S., Parupalli, S., Vechalapu, S., Patil, S., Muniswamy, S., Ghanta, A., Yamini, K. N., Dharmaraj, P. S., and Varshney, R. K. (2015). Next-generation sequencing for identification of candidate genes for Fusarium wilt and sterility mosaic disease in pigeonpea (*Cajanus cajan*). *Plant Biotechnol J* 14: 1183-1194.

Singh, V. K., Khan, A. W., Saxena, R. K., Sinha, P., Kale, S. M., Parupalli, S., Kumar, V., Chitikineni, A., Vechalapu, S., Kumar, C. V. S., Sharma, M., Ghanta, A., Yamini, K. N., Muniswamy, S., and Varshney, R. K. (2017). Indel-seq: a fast-forward genetics approach for identification of trait-associated putative candidate genomic regions and its application in pigeonpea (Cajanus cajan). *Plant Biotechnol J* 15(7): 906-914.

Sintim, H. Y., Zheljazkov, V. D., Obour, A. K., Garcia y Garcia, A., and Foulke, T. K. (2016). Evaluating agronomic responses of camelina to seeding date under rain-fed conditions. *Agron J* 108(1): 349-357.

Smýkal, P., Varshney, R. K., Singh, V. K., Coyne, C. J., Domoney, C., Kejnovský, E., and Warkentin, T. (2016). From Mendel's discovery on pea to today's plant genetics and breeding. *Theor Appl Genet* 129(12): 2267-2280.

Song, Q., Zhang, T., Stelly, D. M., and Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol* 18(1): 99.

Song, Y. H., Kubota, A., Kwon, M. S., Covington, M. F., Lee, N., Taagen, E. R., Cintrón, D. L., Hwang, D. Y., Akiyama, R., Hodge, S. K., Huang, H., Nguyen, N. H., Nusinow, D. A., Millar, A. J., Shimizu, K. K., and Imaizumi, T. (2018). Molecular basis of flowering under natural long-day conditions in Arabidopsis. *Nature Plants* 4: 824–835.

Song, Y. H., Smith, R. W., To, B. J., Millar, A. J., and Imaizumi, T. (2012). *FKF1* conveys timing information for *CONSTANS* stabilization in photoperiodic flowering. *Science* 336: 1045-1049.

Soppe, W. J., Jacobsen, S. E., Alonso-Blanco, C., Jackson, J. P., Kakutani, T., Koornneef, M., and Peeters, A. J. (2000). The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol Cell* 6(4): 791-802.

Sosulski, F. W., and Gore, R. F. (1964). The effect of photoperiod and temperature on the characteristics of flaxseed oil. *Can J Plant Sci* 44: 381-382.

Sotelo-Silveira, M., Montes, R. A. C., Sotelo-Silveira, J. R., Marsch-Martínez, N., and de Folter, S. (2018). Entering the Next Dimension: Plant Genomes in 3D. *Trends Plant Sci* 23(7): 598-612.

Springer, N. M., and Schmitz, R. J. (2017). Exploiting induced and natural epigenetic variation for crop improvement. *Nat Rev Genet* 18(9): 563.

Sridhar, V. V., Surendrarao, A., Gonzalez, D., Conlan, R. S., and Liu, Z. (2004). Transcriptional repression of target genes by LEUNIG and SEUSS, two interacting regulatory proteins for Arabidopsis flower development. *Proc Nat Aca Sci* 101: 11494-11499.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33: W465-W467.

Suárez-López, P., Wheatley, K., Robson, F., Onouchi, H., Valverde, F., and Coupland, G. (2001). *CONSTANS* mediates between the circadian clock and the control of flowering in Arabidopsis. *Nature* 410: 1116-1120.

Sun, J. Flowering time studies in canadian cultivars and 5-Azacytidine mutants of oilseed flax (*Linum usitatissimum* L.). (2015). M.Sc. Thesis, Department of Plant Sciences, University of Saskatchewan.

Sun, Y., Wang, J., Crouch, J. H., and Xu, Y. (2010). Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Mol Breed* 26: 493-511.

Suzuki, M. M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Rev Genet* 9: 465-476.

Swinnen, G., Goossens, A., and Pauwels, L. (2016). Lessons from domestication: targeting cis-regulatory elements for crop improvement. *Trends Plant Sci* 21(6): 506-515.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S. and Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J* 74: 174-183.

Takuno, S., Ran, J. H., and Gaut, B. S. (2016). Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants* 2(2): 15222.

Tanksley, S. D. (1993). Mapping polygenes. *Annu Rev Genet* 27(1): 205-233.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. *Nature* 551(7678): 124.

Thoday, J. M. (1961). Location of polygenes. *Nature* 191(4786): 368.

Tiwari, S. B., Shen, Y., Chang, H. C., Hou, Y., Harris, A., Ma, S. F., McPartland, M., Hymus, G. J., Adam, L., Marion, C., and Belachew, A, Repetti, P. P., Reuber, T. L., Ratcliffe, O. J. (2010). The flowering time regulator *CONSTANS* is recruited to the FLOWERING LOCUS T promoter via a unique cis-element. *New Phytol* 187: 57-66.

Tiwari, S., Krishnamurthy, S. L., Kumar, V., Singh, B., Rao, A. R., Rai, V., Singh, A. K., and Singh, N. K. (2016). Mapping QTLs for Salt Tolerance in Rice (Oryza sativa L.) by Bulked Segregant Analysis of Recombinant Inbred Lines Using 50K SNP Chip. *PloS One* 11: e0153610.

Trick, M., Adamski, N. M., Mugford, S. G., Jiang, C. C., Febrer, M., andUauy, C. (2012). Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol* 12(1): 14.

Ühlken, C., Horvath, B., Stadler, R., Sauer, N., and Weingartner, M. (2014). *MAIN-LIKE 1* is a crucial factor for correct cell division and differentiation in *Arabidopsis thaliana*. *Plant J* 78(1): 107-120.

Underwood, C. J., and Martienssen, R. A. (2015). Argonautes team up to silence transposable elements in Arabidopsis. *EMBO J* 34: 579-580.

Venglat, P., Xiang, D., Qiu, S., Stone, S.L., Tibiche, C., Cram, D., Alting-Mees, M., Nowak, J., Cloutier, S., Deyholos, M. and Bekkaoui, F., (2011). Gene expression analysis of flax seed development. *BMC Plant Biol* 11(1): 74.

Veselý, J., and Čihák, A. (1978). 5-Azacytidine: mechanism of action and biological effects in mammalian cells. *Pharmac Ther* A Chemotherapy, Toxicology and Metabolic Inhibitors 2(4): 813-840.

Vyskot, B., Koukalova, B., Kovařík, A., Sachambula, L., Reynolds, D., and Bezděk, M. (1995). Meiotic transmission of a hypomethylated repetitive DNA family in tobacco. *Theor Appl Genet* 91: 659-664.

Wahl, V., Ponnu, J., Schlereth, A., Arrivault, S., Langenecker, T., Franke, A., Feil, R., Lunn, J. E., Sitt, M., and Schmid, M. (2013). Regulation of flowering by trehalose-6-phosphate signaling in Arabidopsis thaliana. *Science* 339: 704-707.

Wang, H., Beyene, G., Zhai, J., Feng, S., Fahlgren, N., Taylor, N. J., Bart, R., Carrington, J. C., Jacobsen, S. E., and Ausin, I. (2015). CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc Natl Acad Sci U S A* 112(44): 13729-13734.

Wang, H., Cheng, H., Wang, W., Liu, J., Hao, M., Mei, D., Zhou, R., Fu, L., and Hu, Q. (2016). Identification of *BnaYUCCA6* as a candidate gene for branch angle in *Brassica napus* by QTL-seq. *Sci Rep* 6: 38493.

Wang, X., Liu, S., Tian, H., Wang, S., and Chen, J. G. (2015). The small ethylene response factor ERF96 is involved in the regulation of the abscisic acid response in Arabidopsis. *Front Plant Sci* 6: 1064.

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., Takahashi, R., Anai, T., Tabata, S., Kitamura, K., and Harada, K. (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.

Wawrzynska, A., Christiansen, K. M., Lan, Y., Rodibaugh, N. L., and Innes, R. W. (2008). Powdery mildew resistance conferred by loss of the ENHANCED DISEASE RESISTANCE1 protein kinase is suppressed by a missense mutation in *KEEP ON GOING*, a regulator of abscisic acid signaling. *Plant Physiol* 148(3): 1510-1522.

Wei, Q. Z., Fu, W. Y., Wang, Y. Z., Qin, X. D., Wang, J., Li, J., Lou, Q., and Chen, J. F. (2016). Rapid identification of fruit length loci in cucumber (*Cucumis sativus* L.) using next-generation sequencing (NGS)-based QTL analysis. *Sci Rep* 6: 27496.

Wei, X., Song, X., Wei, L., Tang, S., Sun, J., Hu, P., and Cao, X. (2017). An epiallele of rice AK1 affects photosynthetic capacity. *J Integr Plant Biol* 59(3): 158-163.

Weigel, D., and Colot, V. (2012). Epialleles in plant evolution. *Genome Biol* 13(10): 249.

Wenig, U., Meyer, S., Stadler, R., Fischer, S., Werner, D., Lauter, A., Melzer, M., Hoth, S., Weingartner, M., and Sauer, N. (2013). Identification of MAIN, a factor involved in genome stability in the meristems of *Arabidopsis thaliana*. *Plant J* 75(3): 469-483.

Whittaker, C., and Dean, C. (2017). The *FLC* locus: a platform for discoveries in epigenetics and adaptation. *Annu Rev Cell Dev Biol* 33: 555-575.

Widman, N., Feng, S., Jacobsen, S. E., and Pellegrini, M. (2014). Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. *Epigenetics* 9(2): 236-242.

Win, K. T., Vegas, J., Zhang, C., Song, K., and Lee, S. (2016). QTL mapping for downy mildew resistance in cucumber via bulked segregant analysis using next-generation sequencing and conventional methods. *Theor Appl Genet* doi:10.1007/s00122-016-2806-z.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 10(1): 232.

Xu, J., Tanino, K. K., and Robinson, S. J. (2016a). Stable epigenetic variants selected from an induced hypomethylated *Fragaria vesca* population. *Front Plant Sci* 7: 1768.

Xu, J., Tanino, K. K., Horner, K. N., and Robinson, S. J. (2016b). Quantitative trait variation is revealed in a novel hypomethylated population of woodland strawberry (*Fragaria vesca*). *BMC Plant Biol* 16(1): 240.

Yamaguchi, A., Kobayashi, Y., Goto, K., Abe, M., and Araki, T. (2005). *TWIN SISTER OF FT (TSF)* acts as a floral pathway integrator redundantly with FT. *Plant Cell Physiol* 46(8): 1175-1189.

Yamanaka, N., Ninomiya, S., Hoshi, M., Tsubokura, Y., Yano, M., Nagamura, Y., Sasaki, T., and Harada, K. (2001). An informative linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. *DNA Res* 8(2): 61-72.

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., Sanchez, A., Valarik, M., Yasuda, S., and Dubcovsky, J. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci U S A* 103(51): 19581-19586.

Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., Li, K., Yang, N., Li, Y., Zhong, T., Li, J., Chen, Y., Yan, J., Yang, X., and Xu, M. (2013a). CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the post domestication spread of maize. *Proc Natl Acad Sci U S A* 110(42): 16969-16974.

Yang, Z., Huang, D., Tang, W., Zheng, Y., Liang, K., Cutler, A. J., and Wu, W. (2013b). Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. *Plos One* 8: e68433.

Yanovsky, M. J., and Kay, S. A. (2002). Molecular basis of seasonal time measurement in *Arabidopsis*. *Nature* 419: 308-312.

Yoo, S. K., Chung, K. S., Kim, J., Lee, J. H., Hong, S. M., Yoo, S. J., Yoo, S. Y., Lee, J. S., and Ahn, J. H. (2005). *CONSTANS* activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to promote flowering in Arabidopsis. *Plant Physiol*139(2): 770-778.

You, F. M., Duguid, S. D., Thambugala, D., and Cloutier, S. (2013). Statistical analysis and field evaluation of the type 2 modified augmented design (MAD) in phenotyping of flax (*Linum usitatissimum*) germplasms in multiple environments. *Australian J Crop Sci* 7: 1789-1800.

You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., Zhu, T., Luo, M. C., Wang, X., Deyholos, M. K., and Cloutier, S. (2018). Chromosome-scale pseudomolecules refined by optical, physical, and genetic maps in flax. *Plant J* doi: 10.1111/tpj.13944.

Yu, H., Ito, T., Wellmer, F., and Meyerowitz, E. M. (2004). Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. *Nature Genet* 36: 157-161.

Zhang, H., and Zhu, J. K. (2011). RNA-directed DNA methylation. *Curr Opin Plant Biol* 14(2): 142-147.

Zhang, H., Lang, Z., and Zhu, J. K. (2018). Dynamics and function of DNA methylation in plants. *Nat Rev Mol Cell Biol* 19: 489-506.

Zhang, H., Wang, B., Duan, C. G., and Zhu, J. K. (2013). Chemical probes in plant epigenetics studies. *Plant Signal Behav* 8: e25364.

Zhang, J. Z., Mei, L., Liu, R., Khan, M. R. G., and Hu, C. G. (2014). Possible involvement of locus-specific methylation on expression regulation of *LEAFY* homologous gene (*CiLFY*) during precocious trifoliate orange phase change process. *PloS One* 9: e88558.

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., and Jiang, G. L. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16(1): 217.

Zhang, S., Meng, L., Wang, J., & Zhang, L. (2017). Background controlled QTL mapping in pure-line genetic populations derived from four-way crosses. *Heredity* 119(4): 256-264.

Zhong, C., Sun, S., Li, Y., Duan, C., and Zhu, Z. (2018). Next-generation sequencing to identify candidate genes and develop diagnostic markers for a novel Phytophthora resistance gene, *RpsHC18*, in soybean. *Theor Appl Genet* 131(3): 525-538.

Zhu, D., Rosa, S., and Dean, C. (2015). Nuclear organization changes and the epigenetic silencing of FLC during vernalization. *J Mol Biol* 427: 659-669.

Zhu, Z., Tan, L., Fu, Y., Liu, F., Cai, H., Xie, D., Wu, F., Wu, J., Matsumoto, T., and Sun, C. (2013). Genetic control of inflorescence architecture during rice domestication. *Nat Commun* 4: 2200.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet* 39(1): 61.

Zou, C., Wang, P., and Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol J* 14(10): 1941-1955.

Zuo, Z., Liu, H., Liu, B., Liu, X., and Lin, C. (2011). Blue light-dependent interaction of *CRY2* with *SPA1* regulates *COP1* activity and floral initiation in Arabidopsis. *Curr Biol* 21: 841-847.

**Appendix A** Protocol followed for CTAB buffer preparation

CTAB buffer of volume 1 l was prepared using the following components

1M Tris-HCL (pH 8.0) – 100 ml, for a final concentration of 100 mM Tris-HCl

0.5M EDTA – 50 ml, for a final concentration of 25 mM EDTA

NaCl – 87.66 g, for a final concentration of 1.5 M NaCl

CTAB – 20.00 g, for a final concentration of 2%

The final volume was made-up to 1 l with distilled water, and the buffer was autoclaved and stored at room temperature. β–mercaptoethanol - 0.3%, was added to the required volume of buffer freshly before use, every time.

**Appendix B** Quantitation of DNA using Qubit assay

The working solution was prepared by adding 1 µl Qubit dsDNA BR reagent to 199 µl of Qubit dsDNA buffer, per sample. For the two standards provided, 190 µl of working solution was added to two tubes, followed by addition of 10 µl of each standard to the respective tube. For the samples, 198 µl of working solution was taken in each tube and the added with 2 µl of sample to the appropriate tube based on the label. The solution was vortexed for 2 – 3 seconds and incubated at room temperature for two minutes. In the Qubit 2.0 fluorometer, first the standards were read for calibration, followed by the samples.

**Appendix C** Analysis of fragmented DNA using agarose gel electrophoresis

A volume of 150 ml of 1x TAE buffer was taken in conical flask and added with 2.25 g of agar and heated for two minutes and 30 seconds in the microwave oven. The 1.5% agar solution was placed on a magnetic stirrer and allowed to cool to 60$^{o}$C. The agar solution was added with 7 µl of Envirosafe dye (An ethidium bromide equivalent; Helixtec.com) and stirred. The gel was then cast in to a casting tray set with a comb (15 wells). The time taken for solidification was half an hour. The samples were prepared for loading by mixing 3 µl of sheared DNA, 1 µl of the loading dye xylene cyanol and nuclease-free water to make-up to a final volume of 10 µl.

The solidified gel was placed in Bio-Rad wide mini-sub cell GT (Bio-Rad, CA, USA). filled with 600 ml of 1x TAE buffer and the comb was carefully removed. The samples were loaded in individual wells. For the ladder lane, 10 µl of 1kb+ ladder from a stock of 100ng/µl was added to the lanes at both the ends of the gel. The electrodes were connected to the corresponding points on Bio-Rad Power Pac 200 (Bio-Rad, CA, USA). The setting was 80 volts for 80 minutes. After the run was complete, the gel was documented on Bio-Rad Geldoc XR+ (Bio-Rad, CA, USA).

**Appendix D** Protocol followed for analysis of sequencing-libraries using Agilent Bioanalyzer HS DNA assay

Briefly, a gel-dye mix was prepared by adding the HS DNA dye concentrate (15 µl) to the HS DNA gel matrix. Then the gel-dye mix was vortexed for 10 seconds and spun down a spin filter at 2240 x g for 10 minutes. The gel-dye mix (9 µl) was loaded to the third well from top, marked 'G' on the right end of the chip, placed on the priming station. With the plunger at 1ml position, the priming station was locked, and the plunger was pushed down to be held by the clip. After 60 seconds the plunger was released and brought back to its original position. The three other wells marked 'G' were added with 9 µl of gel-dye mix, each. The marker provided with the kit was added to 11 sample wells and the one ladder well, at the rate of 5 µl per well. To the ladder well, 1 µl of ladder was added. The sample wells were loaded with 1 µl of sample. The chip was vortexed on IKA vortexer at 2200 rpm for one minute. The chip was placed into the Agilent bioanalyzer beneath the electrode cartridge appropriately, and the assay was run using the HS DNA assay option.

**Appendix E** Seed inventory

- The seeds from the 735 hills in the 2015 field season are available in three boxes (Agriculture building room number 3C13).
- The seeds from 288 RILs derived from the cross in which ‘RE2’ was the pollen donor are present in a single box (Agriculture building room number 3C13).
- The seeds from the 735 hills grown in 2016 field season are available in two boxes (Kernen Crop Research Farm).
- The seeds from the RCBD trial of the lines chosen as the distributional extremes for flowering time, from ‘Royal’ x ‘RE2’ derived RILs are available in three boxes (180 hills total; each box contains seeds from one replication) (Agriculture building room number 3C13).

**Appendix F** List of DNA sequencing libraries prepared

| Library number | Sample name | Parent/Early/ Late |
|---|---|---|
| 1 | Royal | Parent |
| 2 | RE2 | Parent |
| 5 | Plant 6-E1 | Early |
| 6 | Plant 6-E2 | Early |
| 7 | Plant 6-E3 | Early |
| 8 | Plant 6-E4 | Early |
| 9 | Plant 6-E5 | Early |
| 10 | Plant 6-E6 | Early |
| 11 | Plant 6-E7 | Early |
| 12 | Plant 7-E1 | Early |
| 13 | Plant 7-E2 | Early |
| 14 | Plant 7-E3 | Early |
| 15 | Plant 7-E4 | Early |
| 16 | Plant 7-E5 | Early |
| 17 | Plant 7-E6 | Early |
| 18 | Plant 6-1L | Late |
| 19 | Plant 6-2L | Late |
| 20 | Plant 6-3L | Late |
| 21 | Plant 6-4L | Late |
| 22 | Plant 7-1L | Late |
| 23 | Plant 7-2L | Late |
| 24 | Plant 7-3L | Late |
| 25 | Plant 7-4L | Late |
| 26 | Plant 7-5L | Late |
| 27 | Plant 7-6L | Late |
| 28 | Plant 7-7L | Late |

*Early – early flowering bulk*

*Late – late flowering bulk*