**Literature Review**

The molecular characterisation of vernalisation in safflower via the development of genomic and transcriptomic resources

Darren Cullerne

c3194970

University of Newcastle

# Abstract

Safflower (*Carthamus tinctorius* L.) is an ancient *Astersceae* that is native to Northern Africa, the Middle East and India. In modern times, safflower seed oil is used as an industrial lubricant and as a base for paints and varnishes. Australian grown Safflower accounts for only 0.5% of global safflower seed production, and is primarily used as a break crop. Safflower is best suited for growing in northern areas of Australia where the climate is hot with little chance of prolonged cold weather. Safflower is a diploid that contains 12 chromosomes, with a haploid genome size of approximately 1.4 Gbp. Currently there is no publicly available safflower genome.

Vernalisation is characterised by an extended period of non-freezing cold (greater than 3 weeks). Many plant species, such as *Arabidopsis*, wheat, barley and sugar beet, have varieties that respond to a vernalisation environmental cue that triggers flowering. This can either be an absolute necessity for floral induction (i.e. *absolute* vernalisation requirement) or simply as a response that leads to early flowering (i.e. *facultative* vernalisation requirement). Different plant species have evolved different molecular pathways that respond to vernalisation, but in every plant species investigated, the vernalisation response pathway is epigenetic in nature.

There are a number of different methods of characterising molecular pathways: Traditional genetics cross breed different varieties of a crop with specific traits, forward genetics induces random genetic mutations and reverse genetics interferes with specific genes. Resulting phenotypes are selected for and individual progeny pooled together and sequenced.

While Sanger Sequencing is the highest quality sequencing technology available, Next Generation Sequencing produces the most data for the most affordable price and can be used for generating *de novo* assemblies. While tools exist for assessing the quality of a *de novo* assembly, aligning high quality known transcripts or genomic regions to the assembly is the best way of determining its quality.

This project will examine the vernalisation response pathway in safflower. Expressed phenotypes produced by vernalisation will be measured and the genetic mechanisms that underpin it investigated. A draft *de novo* genome and transcriptome for safflower will also be created to support the characterisation of genes and elucidate the molecular pathways that are triggered in response to vernalisation environmental cues. By examining how safflower responds to vernalisation conditions, it is hoped that the regions in Australia where safflower can be grown will be extended into more southern regions where extended periods of cold over winter are common.

Table of Contents:

Abstract 2

Abbreviations 4

Gene abbreviations 6

1. Safflower – history and characteristics 8

1.1. Genetic and molecular characteristics of safflower 8

1.2. Flowering time in safflower 9

2. The Vernalisation Response 10

2.1. History of vernalisation 10

2.2. The biology of vernalisation 11

2.3. Vernalisation response in *Arabidopsis thaliana* 12

2.4. Vernalisation response in cereals 14

2.5. Vernalisation response in sugar beet 15

2.6. Vernalisation response in other *Asteraceae* 16

2.7. Determining factors affecting vernalisation 16

3. Mapping of pathways and genome characterisation 17

3.1. Traditional crossing methods 17

3.2. Characterisation using forward genetics 17

3.3. Characterisation using reverse genetics 19

4. Sequencing Methodologies 20

4.1. Sanger sequencing 20

4.2. Expressed sequence tags 21

4.3. Microarrays 22

4.4. Next generation sequencing 22

4.5. Creating a *de novo* assembly 24

4.6. Assessing the quality of a *de novo* assembly 26

5. Aims of this research project: 27

References 28

# Abbreviations

*Arabidopsis – Arabidopsis thaliana* (thale cress)

*B. vulgaris – Beta vulgaris* spp. *vulgaris* (sugar beet)

bp – base pair

*C. elegans – Caenorhabditis elegans*

CEGMA – Core Eukaryotic Genes Mapping Approach

cDNA – clonal DNA

DAG – diacyl glyceride

ddNTP – di-deoxynucleotide triphosphate

dicot – dicotyledonous species

DNA – deoxyribonucleic acid

dNTP – deoxynucleotide triphosphate

EST – Expressed Sequence Tag

F# – the generation of progeny (#) after a population has been created by the crossing of two plant varieties

Gbp – gigabase pair (1 Gbp = 1,000,000,000 bp)

kbp – kilobase pair (1 kbp = 1,000 bp)

M# – the generation of mutated progeny (#) after genomic mutations have been introduced to a population

Mbp – megabase pair (1 Mbp = 1,000,000 bp)

Monocot – monocotyledous species

mRNA – messenger RNA

NCBI – National Center for Biotechnology Information

nt – nucleotide

PacBio – Pacific Biosciences

PC – phosphatidylcholine

PCR – polymerase chain reaction

PHD – Plant Homeo Domain

PRC2 – Polycomb Repression Complex 2

RNA – ribonucleic acid

RNAi – RNA interference

t – tonne (metric)

# Gene abbreviations

*AP1* – APETALA 1

*BTC1* – BOLTING TIME CONTOL 1

*CiFL* – FLC-LIKE 1(*Cichorium intybus*)

*CLF –* CURLY LEAF

*CO* – CONSTANS

*FAD2* – FATTY ACID DESATURASE 2

*FCA –* FLOWERING TIME CONTROL PROTEIN

*FIE* – FERTILIZATION-INDEPENDENT ENDOSPERM

*FLC* – FLOWERING LOCUS C

*FLD* – FLOWERING LOCUS D

*FLK* – plant specific K-homology (KH)-domain RNA-binding protein

*FPA* – FLOWERING TIME CONTROL PROTEIN

*FRI* – FRIGIDA

*FRL(1|2)* – FRIGIDA-LIKE (1|2)

*FT(1|2)* – FLOWERING TIME (1|2)

*FVE* – WD-40 REPEAT CONTAINING PROTEIN MSI4

*FY – A. thaliana* homologue of yeast RNA 3' processing factor (Pfs2p)

*LFY* – LEAFY

*LOA* – LOSS OF APOMIXIS

*LOP* – LOSS OF PARTHENOGENESIS

*OS2* – ODDSOC2

*PPD* – PHOTOPERIOD

*SWN* – SWINGER

*SOC1* – SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1

*TSF –* TWIN SISTER OF *FT*

*VEL1 –* VERNALISATION 5/VIN3-LIKE 1

*VRN1|2*|5 – VERNALISATION (NB: *VRN1* and *VRN2* in cereals are different to *VRN1* and *VRN2* in *A. thaliana*)

*VIN3* – VERNALISATION INSENSITIVE 3

*Gene Abbreviations Note 1* – An abbreviation in *CAPITAL ITALICS* represents a gene or mRNA. An abbreviation in CAPITALS represents a translated protein i.e. ‘*FLC*’is the gene or mRNA, ‘FLC’ is the protein

*Gene Abbreviations Note 2* – Where a gene exists as a variant in a specific species, the gene will be prefixed with a two letter abbreviation of the species i.e. *Arabidopsis* (*At*), *Beta vulgaris* (*Bv*)

*Gene Abbreviations Note 3* – When a gene has a recessive allele, it will be indicated as lower case i.e. *FLOWERING TIME – FT* (dominant, normal) verses. *ft* (recessive)

# Safflower – history and characteristics

Safflower (*Carthamus tinctorius* L.), a member of the *Asteraceae* family of flowering plants, is native to the regions of the eastern and southern Mediterranean, the Middle East and India (Knowles 1960). It has been cultivated in these regions for thousands of years, with safflower arrangements and safflower based dyes found in many archaeological sites in Egypt and Mesopotamia, including Pharaoh Tutankhamen’s tomb (Zohary & Hopf 1993). Originally, safflower was cultivated for its edible seeds and the dyes produced from its vibrant yellow and orange flowers, Because of its high oxidative stability, in modern times, safflower oil is more commonly used as a base for varnishes and oil based paints as well as being an effective industrial lubricant (Işigigür et al. 1995; Gecgel et al. 2007). After the oil has been extracted from safflower seeds, the residual meal can also be used as animal feed (Knowles 1949). In 2012, the global harvest of safflower seed was 833,793 tonnes (t) (**Fig 1**), with 65% (536,651 t) originating from Mexico (30.9%, 257,451 t), 18.2% from India (152,000 t) and 15.2% from Kazakhstan (127,200 t) (United Nations 2014). Australian produced safflower seed accounted for just over 0.5% (4,800 t) of the 2012 global production, as it is primarily grown as a ‘break crop’, to break up hard clay pans or to remove excess water from soils before the cultivation of traditional crops, such as wheat and barley (Knights 2010).

## Genetic and molecular characteristics of safflower

Safflower is a dicotyledonous plant (dicot) with a diploid genome of 12 chromosomal pairs. The approximate haploid genome size (i.e. the estimated size of 12 unpaired chromosomes) in four safflower cultivars, namely ‘Ljubljana’, ‘Uzbekistan’, ‘S-2190’ and ‘Huesca’, have been calculated as 1.34 Gbp (1 Gbp = one Gigabase pair = 1,000,000,000 base pairs), 1.38 Gbp, 1.39 Gbp and 1.40 Gbp respectively (Garnatje et al. 2006). Despite publicly available online transcriptomic and expressed sequence tag resources for safflower (Li et al. 2011; Li et al. 2012; Lulin et al. 2012; National Centre for Biotechnology Information 2014), currently, there is no publicly accessible genome for safflower, or indeed, any other *Asteraceae* species. However, there has been research undertaken to characterise specific metabolic pathways in specific *Asteraceae*. For example, oilseed research in safflower has allowed the characterisation of eleven members of the *FATTY ACID DESATURASE 2* (FAD2) gene family, providing insight on the metabolic synthesis of fatty acids and its effect on the composition of safflower oil (Cao et al. 2013). Similarly, research to characterise the *LOSS OF PARTHENOGENESIS* (LOP) and *LOSS OF APOMEOSIS* (LOA) loci in *Hieracium* is unlocking the genetic mechanisms and regulatory pathways that underpin apomictic behaviour (Koltunow et al. 2011). Research into these pathways is quite comprehensive within the context of these specific species. However, despite being the largest and most successful of the

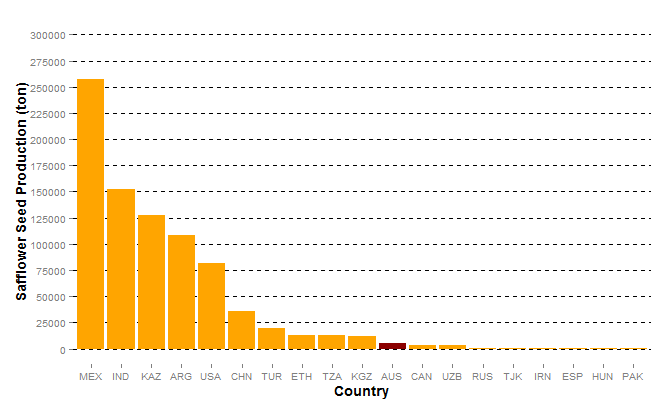


Figure 1 – Global safflower production in 2012 ordered by quantity. Australia (AUS, indicated) is ranked 11th in the world for safflower seed production at approximately 4 800 t (United Nations 2014).

flowering plant families, there is very little research characterising the molecular pathways that exist in safflower and the *Asteraceae* family as a whole (Cronquist 1980).

## Flowering time in safflower

Of particular interest is the coordination of flowering time in safflower. This trait, if modifiable, is likely to have enormous impact on safflower’s ability to adapt to climate change, expand the range of environmental conditions where safflower can be cultivated and increase total yield. Research in the late 1970s characterised both winter and spring varieties of safflower, specifically describing a low survival rate for spring safflower when planted in winter and, conversely, poor performance of winter safflower planted in the spring (Yazdi-Samadi & Zali 1979). Early research in *Arabidopsis thaliana* (*Arabidopsis*) (Johanson et al. 2000) and cereals (Yan et al. 2004) indicated that the expression of a vernalisation response phenotype is underpinned by molecular modifications in the vernalisation response pathway. It is this response that characterises whether a variety in a species is ‘winter’ or ‘spring’. Based on previous research (Johnson et al. 2006), it is hypothesised that the ‘winter hardy’ varieties of safflower will express similar molecular and anatomical phenotypes and involve molecular pathways comparable to other winter and spring varieties in other dicots, such as *Arabidopsis*. However, it is important to note that while some safflower varieties progress rapidly to flowering following vernalisation, it is not a necessary environmental cue for safflower to flower, indicating that in safflower, the vernalisation response is *facultative* (Salisbury & Ross 1992).

It has been documented in model plant systems, such as *Arabidopsis*, as well as agronomically important species such as wheat and sugar beet, that vernalisation is related to crop yield. However, the genetic mechanism(s) of vernalisation in safflower and the effect of an extended cold treatment have not been molecularly characterised. This lack of molecular characterisation is confounded by the absence of publicly available draft genomic resources on which to draw comparisons to existing vernalisation systems.

# The Vernalisation Response

Vernalisation, and its effect on harvest time and crop yield, has been a central research focus of the plant biology community for over 150 years. Vernalisation is characterised by a prolonged (i.e. greater than three weeks) exposure to low but non-freezing temperature (Burn et al. 1993). Furthermore, in *Arabidopsis*, it has been demonstrated that the resultant time to flowering is directly proportional to the length of cold exposure (Sheldon et al. 2000). In numerous plant species, planting time has been optimised to increase the length of time the plant remains in the vegetative growth stage. For example, carrots (*Daucus carota*) store carbohydrates in the root organ during the vegetative growing period. Then, following a ‘wintering’ period where the plant is exposed to vernalisation conditions, transitions from a vegetative growth state to a reproductive state, utilising this stored energy reserve for flowering and seed production (Ingram et al. 2008). While flowering is often attributed to the increased day length of spring, it is not the only environmental factor to invoke the flowering pathway. Exposing a seedling to an extended period of cold will also trigger an early transition from the vegetative growth the flowering, largely as a result of activating the molecular vernalisation pathway. Vernalisation is, therefore, an important determinant of flowering time, and flowering time is crucial to overall crop yield. By restricting development until after winter has passed, the vernalisation response also protects delicate reproductive organs, created during flowering, from cold damage.

## History of vernalisation

The necessity of a prolonged period of cold exposure to promote flowering and crop yield performance of ‘winter’ wheat cultivars was documented as early as 1857 (Klippart 1857). This initial finding stimulated further investigated in rye and other crop species by Gassner in 1918. The term ‘vernalisation’ (from the Latin *vernum*, meaning *spring*) was coined in 1928 by Lysenko, who conducted a vast quantity of agricultural research for the Union of Soviet Socialist Republics (USSR, Soviet Union). These early studies by Lysenko and others showed that the performance and resulting yield of a number of cereal cultivars increased substantially after undergoing a period of wintering when compared to their performance when planted in spring. However, Lysenko incorrectly hypothesised that the progeny of a ‘vernalised’ cereal maintained the attributes of their vernalised parents, and therefore did not need the same exposure to wintering conditions to produce the same yields (reviewed in (Chouard 1960)). In addition, in early Australian colonies, wheat varieties transported from Europe struggled under the warmer, dryer conditions and mild winters of the Australian environment, leading to widespread hunger for the early settlers when crops failed (Macindoe & Brown 1968). It took over 100 years of repeated crop failures before research by Farrar developed many ‘Australian’ wheat varieties. Using several different European wheats as parental lines, Farrer employed traditional cross breeding techniques breeding approaches, selecting the progeny that performed best under the harsh Australian agricultural conditions (Evans 1980).

## The biology of vernalisation

Flowering in plants is an intricate process requiring the coordination of gene expression in response to a number of environmental cues. In *Arabidopsis*, both the shoot apical meristem and floral meristem are located in the shoot apex. During vegetative growth, cells in the shoot apical meristem actively divide and differentiate into vegetative growth tissues while floral meristem tissues remain dormant (Fletcher 2002). Once *FLOWERING TIME* (*FT*) is expressed in leaf tissue in response to vernalisation (or other factors), the FT protein travels through the phloem and into meristem tissues (Corbesier et al. 2007). Once there, it triggers floral induction and floral meristem cells commence dividing and differentiating, causing the plant to bolt to flowering. It has been demonstrated that after exposure to vernalisation conditions, the vernalised shoot tissue can be grafted to another plant. When a vernalised shoot tip was grafted onto non-vernalised root stock, the plant flowered as if the entire plant had been exposed to the vernalisation treatment. Conversely, when a non-vernalised shoot tip was grafted onto vernalised root stock, the opposite was observed. This observation has been corroborated in both *facultative* (where vernalisation decreases time to flowering, but is not essential for flowering), and *absolute* (where vernalisation is required to progress from vegetative growth to flowering) vernalisation sensitive species.

Contemporary research has shown that the regulatory pathways and genetic mechanisms of vernalisation are species-specific (Reeves et al. 2007). For example, vernalisation in *Arabidopsis*,and many other dicots, is centrally regulated via expression of the floral repressor *FLOWERING LOCUS C* (*FLC*), whereas cereals such as barley are regulated via expression of *VERNALISATION 2* (*VRN2*). In beets (*Beta vulgaris* ssp. *Vulgaris*) however, *FLOWERING TIME 1* (*FT1*) is responsible for regulating *FLOWERING TIME 2* (*FT2*),(Pin et al. 2010). Being epigenetic in nature, the response pathway of all vernalisation responsive plant species is mechanistically related, i.e. environmental cues can modify gene expression between varieties of the same species, which in turn can result in the expression of a different phenotype, without any alteration to the underlying DNA sequence.

## Vernalisation response in *Arabidopsis thaliana*

The MADS-box transcription factor FLC mediates the transition of *Arabidopsis,* and many other dicots, from vegetative to reproductive growth (**Fig 2**). Research has shown that both genetic and epigenetic mechanisms contribute to repressing the expression of the floral repressor *FLC* during this transition (Boss et al. 2004; Finnegan et al. 2005). When *FLC* is expressed at high levels, promoted by *FRIGIDA* (*FRI*), *FRIGIDA-LIKE 1* (*FRL1*) and *FRIGIDA-LIKE 2* (*FRL2*) (**Fig 2.1**), FLC represses the expression of *FLOWERING TIME* (*FT*), the FT homolog *TWIN SISTER OF FT* (*TSF*), and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*). This increased FT, TSF and SOC1 levels which, in turn, suppresses expression of *LEAFY* (*LFY*) and *APETALA1* (*AP1*), the two primary promoters of floral apical meristem growth (Amasino 2004).

Upon induction of FLC expression, the VERNALISATION 2 (*VRN2*)/Plant Homeo domain Polycomb Repression Complex 2 (PHD-PRC2) consisting of *VRN2* and PHD-PRC2 proteins, *CURLY LEAF* (*CLF*), *SWINGER* (*SWN*) and *FERTILIZATION-INDEPENDENT ENDOSPERM* (*FIE*) (Köhler & Villar 2008) is constitutively bound to the *FLC* locus. The binding of the VRN2/PHD-PRC2 complex to *FLC* maintains the locus in an open conformational shape, allowing transcriptional machinery to access the *FLC* regulatory sequence and promote *FLC* expression via H3 acetylation (De Lucia et al. 2008). During vernalisation, *VERNALISATION INSENSITIVE 3 (VIN3)* levels increase. VIN3, along with *VERNALISATION 5/VIN3-LIKE* (*VEL1*) and *VERNALISATION 5* (*VRN5*), bind to the PHD-PCR2 complex to promote histone H3 deacetylation and *VRN2*-directed methylation of H3K9 and H3K27. Methylation of the *FLC* locus closes the open conformation of *FLC*, blocking the transcription machinery from accessing *FLC*, thereby repressing *FLC* expression (Finnegan & Dennis 2007). This epigenetic repression of *FLC* is stable and irreversible, ensuring the transition to a flowering state is permanent (Levy et al. 2002; Sung & Amasino 2004). Reduced levels of FLC leads to increased expression of *SOC1* and *FT*. Increased SOC1 and FT levels in turn enhance the expression of the floral promoters *LFY* and *AP1* (**Fig 2.2**).

➋

➊

➍

➌

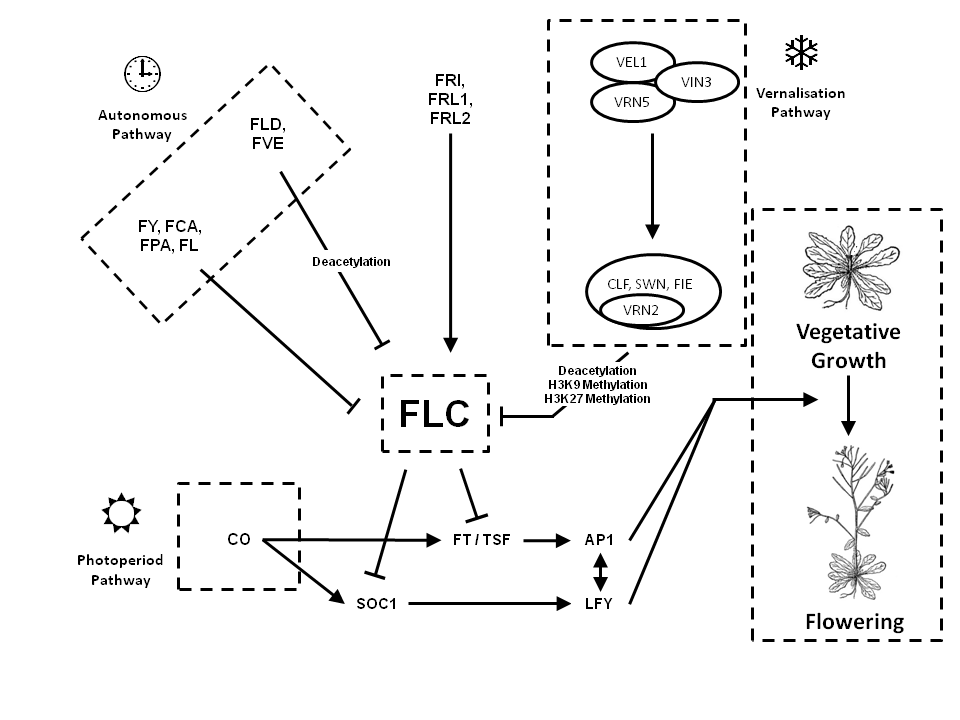


Figure 2 – The FLC regulatory pathways involved in the triggering of flowering in *Arabidopsis thaliana*:   
**1)** FRI, FRL1 and FRL2 promotes *FLC* expression (Michaels et al. 2004).  
**2)** Exposure to extended periods of cold triggers represses FLC expression via a PHD-PRC2 complex, H3K9me and H3K27me in FLC, repressing it’s expression (De Lucia et al. 2008).  
**3)** As day length increases, the photoperiod pathway is triggered, which increases CO expression and bypasses the repressive effects of FLC (Hayama & Coupland 2004; Golembeski et al. 2014).  
**4)** Flowering will eventually be initiated by the autonomous pathway. FLD and FVE deacetylate FLC and FY, FCA, FPA, and FL repress FLC expression. (Simpson 2004).

The Photoperiod Pathway, also referred to as the External Coincidence Model, mediates the transition to flowering by exposure to increasing day length without a vernalisation period (**Fig 2.3**) (Hayama & Coupland 2004). Increased periods of day light promotes the expression of *CONSTANS* (*CO*), which in turn overrides the repressive effects of *FLC* via CO-mediated activation of *FT* and *SOC1* expression (Golembeski et al. 2014).

If the vernalisation or photoperiod pathways are not activated, eventually the Autonomous Pathway will trigger the transition of *Arabidopsis* to flowering*.* Two groups of transcription factors work independently to down regulate *FLC* expression. The first group, consisting of *FLOWERING LOCUS D* (*FLD*) and *WD-40 repeat-containing protein MSI4* (*FVE*), work epigenetically to deacetylate histones at the *FLC* locus. The second group, comprising of *FLOWERING TIME CONTROL PROTEIN* (*FCA*), A. thaliana *homologue of yeast RNA 3' processing factor* (*FY*), *FLOWERING TIME CONTROL PROTEIN* (FPA) and *plant specific K-homology* (*KH*)*-domain RNA-binding* (*FLK*) prematurely polyadenelate *FLC* mRNA, resulting in an inactive version of FLC (Simpson 2004).

While the molecular pathways for vernalisation and flowering in *Arabidopsis* are well characterised, these pathways differ substantially with the pathways in other plant species.

## Vernalisation response in cereals

In monocotyledous species (monocots), such as wheat (*Triticum aestivum*), barley (*Hordeum vulgare*) and *Brachypodium distachyon* (a model monocot), the vernalisation pathway differs greatly from *Arabidopsis*. While there are many mechanisms conserved between monocots and *Arabidopsis,* such as the Photoperiod Pathway (Griffiths et al. 2003) and CO as a regulator of *FT* the most notable difference in cereals is the absence of a known homolog for AtFLC(Figure 6 below). Instead, HvVRN1 (a MADS-box transcription factor) and HvVRN2 (a zinc-finger motif with CCT domain, which in cereals is distinct from AtVRN2) (Yan et al. 2004), along with HvFT, are responsible for regulating flowering time in cereals. Before wintering, the cereal remains in a state of vegetative growth due to the presence of HvVRN2, which represses *HvFT* expression (Ream et al. 2014). The floral repressor, ODDSOC2 (OS2), also present at high levels in winter cereals prior to vernalisation, maintains the cereal’s vegetative growth state (Greenup et al. 2010). Once exposed to vernalisation conditions, winter barleys increase expression of *HvVRN1* via lower H3K27me3 and higher H3K4me3 levels in the *HvVRN1* locus. This opens the histone conformational shape, allowing *VRN1* expression (Oliver et al. 2009). Similar to *FLC* in *Arabidopsis*, this change is stable, but in winter barley, the shape change allows access (rather than restricting access) to the *HvVRN1* gene. Increased levels of HvVRN1 in turn repress expression of *HvVRN2*. Low HvVRN2 levels promote *HvFT* expression, triggering the transition of the winter barley to flowering (Trevaskis et al. 2006). Exposure to vernalisation conditions and the presence of HvVRN1 also stably inhibits the expression of *OS2*. Lower levels of OS2 promote the expression of *FPF1*, which in turn promotes the transition to a flowering state.

Due to its acclimatisation to tropical environments, the vernalisation response in rice (*Oryza sativa*) differs substantially to other cereals. Exposing rice to cold conditions during development, significantly warmer and for shorter periods than seen in vernalisation, damages the developing flower and seed heads, resulting in sterility and significant crop losses (Oliver et al. 2007).

***PPD1***

***CO***

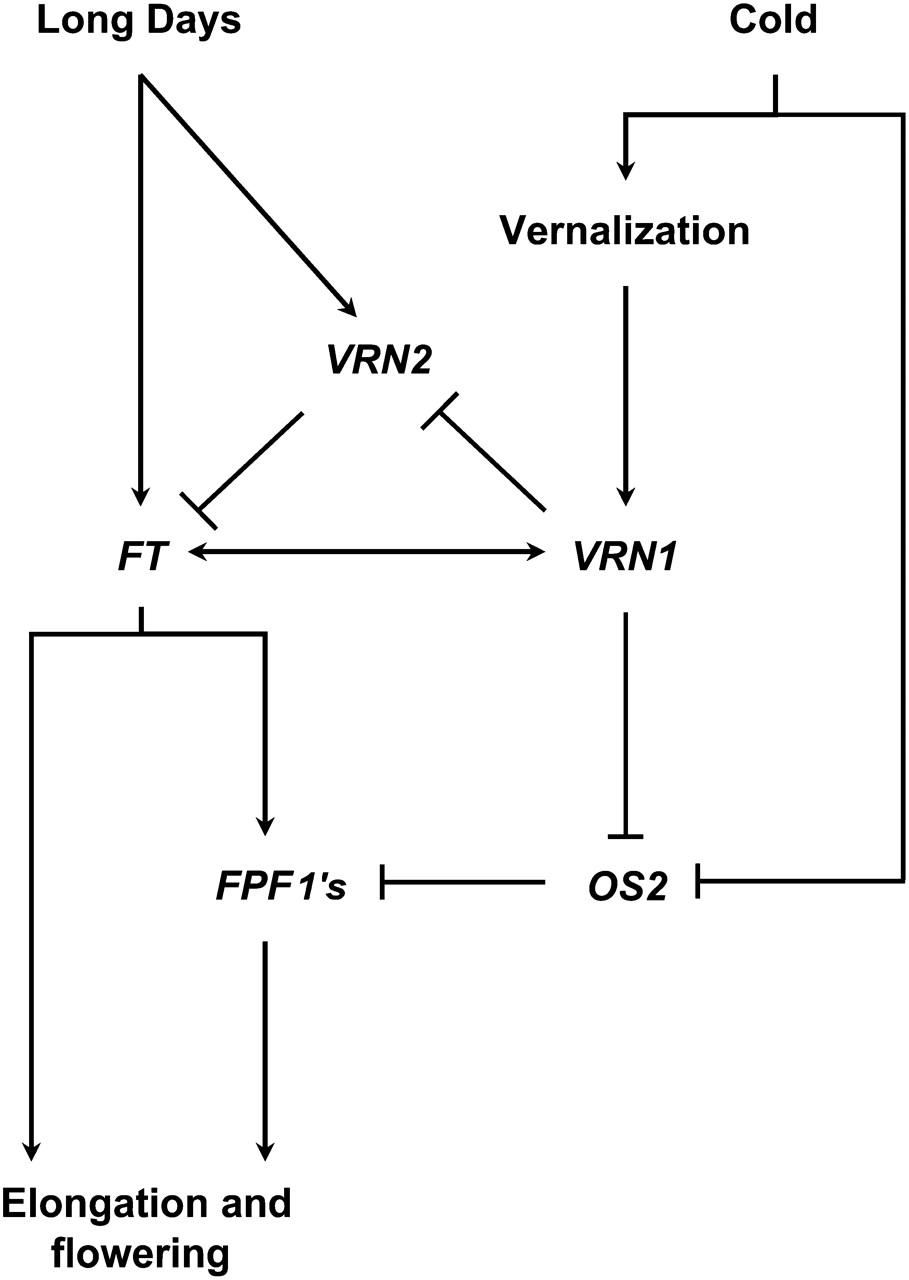


Figure 3 – Vernalisation gene regulatory pathway involved in triggering flowering in winter cereals. While the vernalisation response is different, response to day length cues contains similarities to *A. thaliana* (Modified from (Greenup et al. 2010) and (Trevaskis et al. 2007))

## Vernalisation response in sugar beet

Sugar Beet (*Beta vulgaris*) has also evolved an alternate method of regulating flowering time based on day length and wintering. Photothermal induction (i.e. exposure to vernalisation conditions followed by increased day length) was shown to be required by many varieties of sugar beet to induce flowering (Owen et al. 1940). Therefore, breeders have selected for phenotypes that have an absolute vernalisation responsive as a way of maintaining the vegetative growth stage to maximise yield. In a study of wild ‘Sea Beet’ (*B vulgaris* ssp. *maritama*) populations around the coast of France, it was found that in each population, the necessity for vernalisation increased proportionally to the northern latitude of the population (Boudry et al. 2002). Recent studies have characterised the molecular mechanisms underpining the vernalisation response in beet as distinct from both *Arabidopsis* and cereals (Pin et al. 2012). Rather than a single *FT* gene responsible for triggering flowering time controlled by FLC, as seen in *Arabidopsis*, two paralogous *FT* genes, *BvFT1* and *BvFT2*, central to the regulation of flowering, are controlled by *BOLTING TIME CONTROL 1* (*BvBTC1*). In biannual sugar beet, it is the recessive allele of *Bvbtc1* that produces a phenotype with an absolute vernalisation requirement. While *B. vulgaris* contains a number of pathways containing substantial homology with the flowering pathways found in *Arabidopsis*, including the photoperiod pathway (Chia et al. 2008) and the autonomous pathway (Abou-Elwafa et al. 2011), these pathways are yet to be fully characterised in sugar beet.

## Vernalisation response in other *Asteraceae*

While much research has been undertaken with regards to the vernalisation response in *Arabidopsis*, cereals and sugar beet, current research regarding the vernalisation response in safflower, and indeed, other members of the *Asteraceae* family is relatively scarce. Early research in lettuce (*Lactuca sativa*) observed that germinated seeds that were vernalised prior to planting responded by bolting up to four weeks earlier than unvernalised seeds (Gray 1942; Warne 1947; Rappapport et al. 1956). Later studies indicated that, similar to *Arabidopsis*, given the correct daylight exposure, lettuce bolts to flowering without the need for vernalisation, indicating a *facultative* vernalisation response in vernalisation sensitive lettuce species (Waycott 1995). However, all of this research focuses on the expressed phenotype and did not examine anything at the genetic level.

Chicory (*Cichorium intybus*), is an *Asteraceae* with an absolute vernalisation requirement. It has been demonstrated that when exposed to vernalisation conditions, chicory expresses CiFL1, a MADS-box transcription factor with significant sequence homology to AtFLC (Périlleux et al. 2013). To confirm this similarity, when *CiFL1* was transformed into *Arabidopsis* and over expressed, the mutants showed a significant delay in onset of flowering, regardless of vernalisation exposure, indicating a similar molecular pathway effect between CiFL1 and AtFLC.

Early studies in safflower indicate a relationship between photoperiod, vernalisation and a decrease in the time to flowering in some safflower varieties (Zimmerman 1973). However, as mentioned above, the limited availability of genetic resources for safflower and the *Asteraceae* make characterisation of molecular pathways and mechanisms in these species challenging. As further resources become available and are annotated, the mechanisms by which vernalisation, and other factors, affect flowering time will be better understood.

## Determining factors affecting vernalisation

There are a number of differing and redundant pathways in *Arabidopsis*, cereals and *B. vulgaris* that can trigger flowering. While a mutation in the vernalisation pathway in a *facultative* vernalisation sensitive species may not necessarily stop the transition from vegetative growth to flowering, it will certainly prevent flowering in an *absolute* vernalisation sensitive species. For example, a mutation that removes the *CO* regulators may cause the plant to lose the requirement for lengthening light as a flowering trigger. This results in a fast transition to flowering regardless of the plant’s existing requirement for vernalisation. Recent *Arabidopsis* research indicates that higher levels of a diacyl glyceride (DAG) comprising of phosphatidylcholine (PC) and oleic acid (18:1) in the shoot apical meristem accelerates the transition to flowering, and vice versa (Nakamura et al. 2014). This research also shows other DAGs containing PC bind to *FT*. This indicates that the levels of certain specific lipids in plant tissues may be another factor that contributes to the triggering of flowering in *Arabidopsis*. However, further research is required to determine exactly how lipid levels in plant tissues affect the vernalisation response in *Arabidopsis* and other plant species.

While safflower, *Arabidopsis* and *B. vulgaris* are all dicots, there are so many divergences in the molecular pathways between *Arabidopsis* and *B. vulgaris* in their vernalisation response, it is plausible that there are a number of factors in safflower, possibly even across the *Asteraceae* as a family, that set them apart from other dicots in the context of vernalisation. These confounding factors will impede the identification and elucidation of genes critical in the vernalisation response, especially when comparing vernalisation sensitive and vernalisation resistant safflower varieties.

# Mapping of pathways and genome characterisation

## Traditional crossing methods

One of the fundamental methods of characterising a genome is through cross breeding. A plant with specific expressed phenotype or marker is crossed with another variety and the progeny selected for this trait. The resulting progeny can then be sequenced, either individually or as a pool. These reads can then be mapped to a reference and identified as expressing or lacking the observable trait. A specific example of a crossing methodology used for genetic mapping is a Multiparent Advanced Generation Inter-Cross (MAGIC) population. Markers are identified in an organism and a number of parents are crossed. The resulting progeny are crossed within the population over a number of generations until a very specific set of traits are expressed. These progeny are sequenced along with the parents and the traits traced back to parental varieties. This method was successfully used for the genetic mapping of markers in wheat (Huang et al. 2012)

## Characterisation using forward genetics

Forward genetics induces random mutations to the genome, as either single point or clustered, via chemical or radiation exposure to express phenotypes (**Fig 4**)

➍

➌

➋

➊

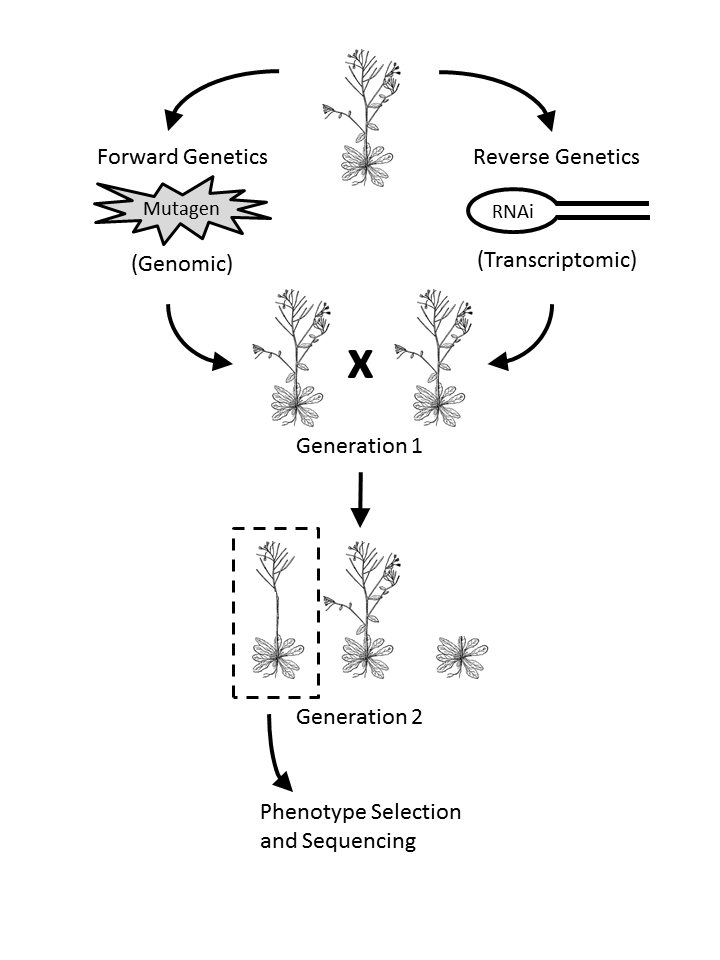


Figure 4 – Forward and reverse genetics. Forward genetics (1) involves random mutation of the genome with a chemical or radiation mutagen. Reverse genetics (2) involves interference of a particular transcript, for example, with an RNAi construct. In both, after the first cross, phenotyoes caused by mutation or RNAi are generally not seen in the first generation (3). In subsequent generations, phenotypes are selected, sequenced and reads mapped against a reference (4).

(Page & Grossniklaus 2002). In diploid plants, when a germinated seed is mutated (the M1 parent), it contains random bud differing mutations on each genome in each cell. Therefore, these heterozygous mutations will not express any phenotype resulting from the mutation because there is an unmutated locus present that functions normally. However, should these mutations be passed to the gametes and the plant produces seed via crossing, they will be replicated in both chromosomal copies of the genome. If the mutation is located in a transcribed or regulatory locus, it may interfere with normal metabolic function and express a phenotype. In each generation of mutated plants (M2, M3, M4 etc), assuming no new genetic material is introduced, these mutations remain and phenotypes continue to be expressed (Sikora et al. 2011). Plants expressing a specific phenotype are selected and pooled together for DNA extraction and sequencing, with resultant reads mapped against a normal reference. Mutated regions will be indicated, whereas unrelated mutations will appear as random “noise” (**Fig 5**). The greater the number of different mutant phenotypes in the pool, the clearer and stronger the signal is to indicate regions controlling expression of that phenotype. Because reads and mutations are mapped to specific regions of the genome, forward genetics allows mapping of traits and the characterisation of a genome without necessarily needing to have *a priori* knowledge of the function of that region. However, a high quality *de novo* genome, or at least, a quality reference for the region of interest, is required for forward genetics to be a useful and applicable approach to characterising a species.

## Characterisation using reverse genetics

Another approach for examining the underlying molecular mechanisms of a species is reverse genetics. This process examines the phenotypical effect when specific genes are manipulated to either modify them to make the resulting gene unusable in the system or silence them completely (**Fig 4.2**). RNAi constructs are a method to produce such manipulations. In *Nicotiana benthamiana*, the composition of the leaf oil was altered by silencing *FATTY ACID DESATURASE 2* (NbFAD2) and introducing two *Arabidopsis* genes, *acyl-CoA:diacylglycerol acyltransferase 1* (AtDGAT1) and *FATTY ACID ELONGASE 1* (AtFAE1) (Naim et al. 2012). The drawback of using an RNAi construct for reverse genetics is the requirement for *a priori* knowledge of expressed transcripts in the system, as well as the need to create knockout constructs that successfully silence the gene being investigated. However, with the cost of sequencing having decreased substantially in the last decade, creating useful *de novo* transcriptomes is now much easier and cheaper, providing a wealth of candidate genes that can be examined with reverse genetics.

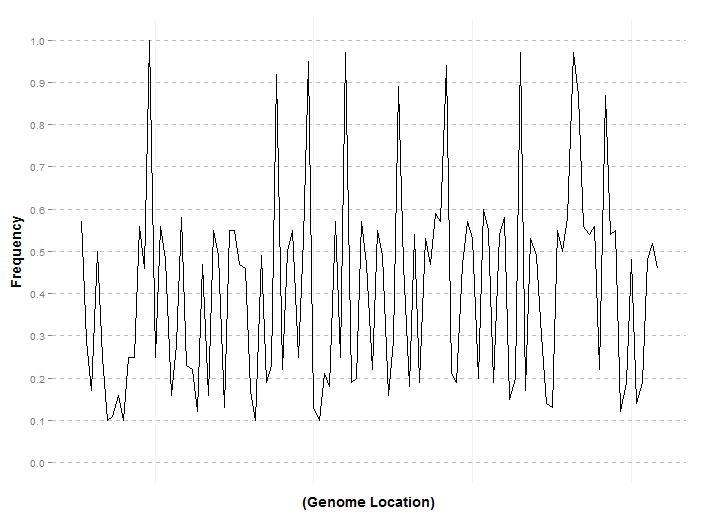


Figure 5 – An example of detection of locations of mutations in a segregated population. Mutated locations unrelated to the segregated phenotype are represented by noise and are randomly distributed throught the population. Those mutations that relate to the phenotype are found in nearly every member of the population and are indicated by the high peaks.

# Sequencing Methodologies

The availability of genomic and transcriptomic resources has advanced biological research in leaps and bounds. Since the 1970s, genetic resources have allowed the elucidation of genes and transcription factors involved in not only vernalisation, but almost every major metabolic pathway in plants, animals and microbes. These advancements in sequencing technology have exponentially increased the quantity and quality of data acquired from a sample. This is made even more accessible inthat sequencing data can be generated in a fraction of the time and at a far lower cost than ever before.

## Sanger sequencing

In 1975, Sanger and Coulson developed one of the first sequencing techniques, the ‘chain termination’ method, colloquially called ‘Sanger Sequencing’ (Sanger & Coulson 1975). It involves cloning the sequence of interest, genomic or transcriptomic, and performing four separate Polymerase Chain Reactions (PCRs), each containing a single type of di-deoxynucleotide triphosphate (ddNTP). During elongation, if the polymerase binds a ddNTP to the nucleotide chain instead of a deoxynucleotide triphosphate (dNTP), the absence of a 3'-hydroxyl group on the ddNTP prevents another dNTP or ddNTP from bonding. When electrophoresed over polyacrylamide gel and fluoresced, the result is a number of varying length bands in each column of the gel corresponding to the order of nucleotides. Many significant genomes have been constructed using Sanger Sequencing, including *Caenorhabditis elegans* (*C. elegans*), *Arabidopsis* (Arabidopsis Genome Initiative 2000)and the human genome (The C. Elegans Sequencing Consortium 1998; Arabidopsis Genome Initiative 2000; International Human Genome Sequencing Consortium 2004). The disadvantages of Sanger sequencing is the time and cost involved. The human genome (approximately 3.4 Gbp, diploid) cost approximately USD $13 billion dollars and took hundreds of labs across the world thirteen years to complete, at a cost of around $4 per nucleotide (International Human Genome Sequencing Consortium 2004). In comparison, today the cost of data generated by a single lane of an Illumina HiSeq 2500 sequencer is approximately $0.05 per million nucleotides (Wetterstrand 2014). While Sanger Sequencing is the oldest, most time consuming and expensive method of sequencing per nucleotide, it is still considered the gold standard for sequencing quality.

## Expressed sequence tags

Expressed sequence tags (ESTs) are short cDNA fragments between 200 bp and 800 bp long representing an mRNA transcript. They are created by randomly cloning and sequencing transcripts in a sample using PCR. The result is an ‘expression profile’, a unique set of genes expressed in a tissue type or environmental condition. Each sequence within the set uniquely identifies the gene expressed. These expression profiles can be compared to identify differentially expressed genes e.g. diseased and healthy tissue (Rezvani et al. 2000). ESTs have a number of notable limitations; the first is the requirement for *a priori* knowledge of homologues transcripts to characterise novel genes. A novel EST is difficult to identify if there is no sequence homology against a known transcript, i.e. a transcript will be known to be expressed without necessarily knowing its function. Likewise, identification of rare transcripts, such as those found in the human brain, can be difficult without the use of unique primer sets. Rare transcripts require either *a priori* knowledge of the transcript (Adams et al. 1991) or additional molecular techniques to successfully detect their presence. If a rare transcript and an abundant transcript both contain the same PCR primers, the probability of the primers binding to the rare transcript is low. The high signal produced by the abundant transcript makes it difficult to detect expression of the rare transcript.

Newer sequencing technologies surpass ESTs with regard to the quantity of data produced and the accuracy of the resulting expression analysis. However, because ESTs are directly cloned from extracted mRNA and do not require assembly of short fragments via an algorithm, EST libraries often complement *de novo* assemblies, verifying the quality and accuracy of an assembly.

## Microarrays

Microarrays are constructed by binding millions of oligonucleotide fragments to a glass slide either by photochemistry (Fodor et al. 1993) or technology similar to an inkjet printer (Allain et al. 2001). The type of array is determined by its use: RNA fragments represent key genes, tissues or sets of genes for medical diagnostics; genomic locations to observe nucleotide polymorphisms; methylation arrays to examine gene regulation and methylation; or sets of tiling arrays for sequencing whole genomes. Each microarray is limited to 6.5 million oligonucleotides on its surface, restricting their application to specific medical diagnostics, use in characterising small genomes such as bacteria, smaller transcriptomes or sets of common transcripts. Sets of microarrays can be used to sequence smaller eukaryotic genomes such as *Arabidopsis*. Like ESTs, the requirement for *a priori* knowledge of oligonucleotides for microarray construction and analysis decreases their efficacy for discovering *de novo* or low count transcripts. Similar related species have been used as an adjunct, with mixed results observed (Buckley 2007).

## Next generation sequencing

In the last 15 years, Next Generation Sequencing(NGS) has expanded the quantity and quality of information available to levels never before seen, with data available in less time and at an ever decreasing cost (**Fig 6**). Illumina Paired End (PE) sequencing is one of the most widespread sequencing technologies used today. It takes a DNA sample (or cDNA sample, synthesised from RNA) and randomly fragments it. These fragments are filtered by length, the oligonucleotide adaptors are ligated to each end of the fragment and the entire fragment is amplified. These lengths of double stranded cDNA are denatured and hybridised to complementary adaptor oligonucleotides and the complementary sequence synthesised. These are then denatured again, but during the annealing stage, the other end of the fragment hybridises to the surface oligonucleotide, forming a bridge. Once this bridge is formed, multiple cycles of denaturing and elongation form free strands that can then be sequenced (Bentley et al. 2008). A single lane on an Illumina HiSeq 2500 sequencer produces between 240 and 300 million reads of 100 bp length (approximately 24 Gbp to 30 Gbp) in under 48 hours . Even on the lower end Illumina MiSeq which costs around a third of the initial infrastructure cost of a HiSeq,

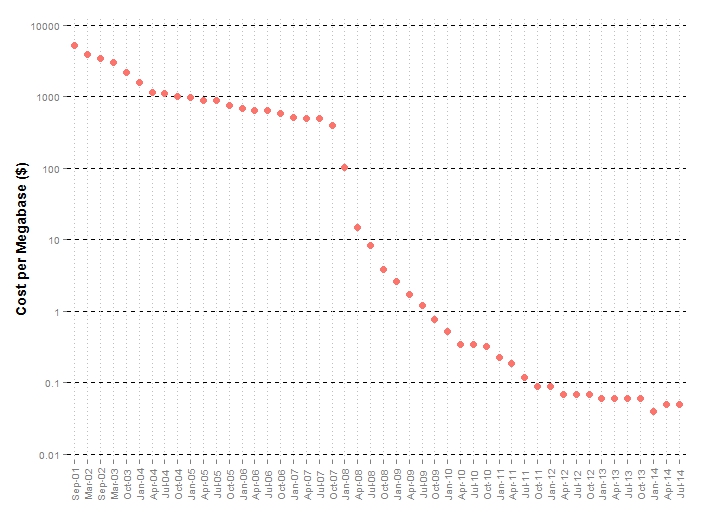


Figure 6 – Sequencing cost per megabase (1 Mbp = 1,000,000 bp) between September 2001 and July 2014 (Wetterstrand 2014)

20 million reads of 75 bp length (approximately 1.5 Gbp) can be produced in around the same time (Quail et al. 2012).

Mate Pair sequencing is an Illumina protocol which is used to improve a genomic DNA assembly, but not by producing long reads. Instead, it produces pairs of reads with an insert length of several thousand nucleotides (Illumina 2012). While this sequencing methodology is unsuitable for transcriptomic samples, it is well suited for scaffolding contigs from a *de novo* genomic assembly together and orienting them in relation to one another (**Section 4.5 - ‘Creating a *de novo* assembly’**). In general, the error rates for PE Illumina sequencing is quite low, with a reported error rate of 0.26% for the HiSeq (approximately 1 error in every 385 nt) and 0.8% for the MiSeq (approximately 1 error in every 125 nt) (Quail et al. 2012).

Pacific Biosciences (PacBio) ‘Single Polymerase Molecule’ sequencing is an alternative to Illumina’s PE sequencing. Rather than hybridising short fragments of DNA samples with adaptors, PacBio technology sequences long strands of single stranded DNA. A DNA polymerase is bound to a well where a fluorescent detector is located. As the DNA polymerase incorporates tagged dNTPs during elongation, the well fluoresces at a frequency corresponding to the bound nucleotide. The fluorescence frequency is read by the detector and the nucleotide recorded (Eid et al. 2009). This allows far longer reads to be sequenced, with average PacBio reads being around 8 kbp in length, with some extending to over 40 kbp. However, PacBio sequencing is far more error prone than Illumina sequencing. PacBio reads have an approximate error rate of 12.86% (approximately 1 in every 7 nt), far in excess of the error rates seen in Illumina sequencing (Quail et al. 2012)

NGS technology is constantly being updated and revolutionised. The IonTorrent system, however, is different from the previously mentioned sequencing technology as it is able to detect the minute electrical change associated with the release of hydrogen ions when each dNTP is bonded to an elongating DNA strand (Purushothaman et al. 2006). The Moleculo sequencing system, which also boasts the ability to sequence a single molecule of DNA, is of benefit in instances when the quantity of sample to be sequenced is very low (Kuleshov et al. 2014).

## Creating a *de novo* assembly

While the cost of creating next generation sequencing information is significantly less than it was 15 years ago, only a fraction of the organisms on Earth have had their genome sequenced. Those sequenced include model organisms such as *C. elegans*, *Drosophila melanogaster* (fruit fly) (Adams et al. 2000), *Arabidopsis*, and other important and endangered species, such as the Giant Panda (Li et al. 2010) and the platypus (Warren et al. 2008). There are two major methods employed for creating a *de novo* assembly (**Fig 7**). The oldest and most common method is overlapping short reads to create contigs of consensus nucleotides (Peltola et al. 1984; Huang & Madan 1999; Li et al. 2008). This method is effective, because each read is compared with every other to find the best possible match. While subsequence ‘seeds’ can be samples from a short read to find initial candidate overlaps to extend, these algorithms tend to scale quadratically. Therefore, without some way of filtering out unsuitable or duplicate reads, running the algorithm to create a *de novo* assembly soon becomes prohibitively time consuming.

An alternate method proposed for *de novo* assembly is the creation of d*e Brujin* graphs from short reads. Each short read is broken down into a set of ‘k-mers’, smaller subsequences of the short reads, and *de Brijun* graph created, which describes how to proceed from one k-mer to the next. Graphs from each read are collapsed into larger graphs, combining identical subgraph regions, thereby extending the length of the reconstructed sequence until eventually, complete contigs are built. (Pevzner et al. 2001). This methodology is far more computationally efficient when compared to the traditional overlapping method. It also allows better assembly of repeat regions, a task that most overlapping *de novo* assemblers are

\*

\*\*

➏

➎

➍

➌

➋

➊

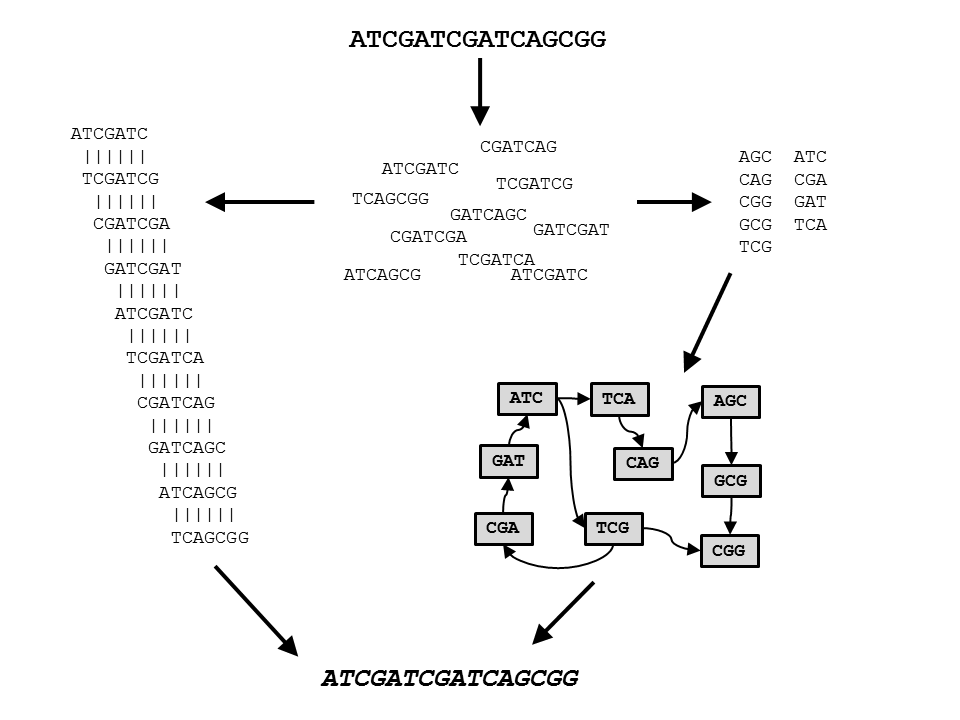


Figure 7 – Two approaches for *de novo* sequence assembly. In both cases, the sample sequence (1) is fragmented and sequenced as a number of shorter reads (2). Traditional approaches for assembly involve aligning every sequence against each other and reading the consensus (3). In the *De Brujin* approach, reads are first broken down into a number of k-mers (subsequences), sorted and duplicates removed (4). A *De Brujin* graph is then created that connects each k-mer (node) to the next via paths (edges). This graph is then traversed, starting at the input node (the node containing more edges exiting the node than entering it, (\*)) and finishing at the output node (the node containing more edges entering the node than exiting it (\*\*)), ensuring every node is visited at least once (5). While both approaches can lead to the creation of a *de novo* assembly (6), using *De Brujin* graphs is far more computationally efficient than overlapping consensus methods.

poor at. Algorithms based on the d*e Brujin* method include Velvet (Zerbino & Birney 2008), ABySS (Simpson et al. 2009), SOAPDeNovo2 (Luo et al. 2012) and BioKanga (unpublished, http://sourceforge.net/projects/biokanga/).

The calibre of the source data used for a *de novo* assembly affects the quality of the resulting assembly. In any sequencing experiment, not just *de novo* assembly, filtering out reads that contain low quality scores, lengths of ambiguous nucleotides (Ns) or substantial contamination from adaptors will decrease the time taken for an algorithm to run, as fewer reads are processed. A common approach for producing a useful assembly, especially when assembling a *de novo* genome, or any assembly where long contigs are important, is to combine a number of different technologies in a multi-step approach. Short paired end reads (e.g. Illumina 100 bp paired end reads) are first assembled into contigs, using either *de Brujin* or an overlap approach. Next, paired end mate pair reads are used to ‘scaffold’ contigs together, i.e. bringing contigs together that are separated by a distance greater than the short read insert length. This orients contigs in relation to one another across a known distance, allowing gaps of a known length to be produced. These ambiguous regions can then be explored using complementary primers at the known region, oriented to extend the cloned sequence across into the unknown region.

Recently, long, error prone PacBio reads have complemented *de novo* assemblies constructed from Illumina short paired end and mate pair reads to improve the length and quality of the resulting assembly. The PacBio reads bring together assembled short read contigs that cannot be brought together with mate pair scaffolding. They also can eliminate misassembled and chimeric contigs resulting from the assembly algorithm incorrectly merging two disparate sequence segments into a single consensus contig. In turn, high quality short read assemblies are used to correct errors in the PacBio reads. (Bashir et al. 2012; Utturkar et al. 2014)

The problem of *de novo* assembly is so computationally complex and biologically important that ‘Assemblathons’ are now organised. These enable bioinformaticians and software developers to bring together assemblers and other tools to compare algorithm efficiency and *de novo* assembly performance on reads generated from a known reference (Earl et al. 2011; Bradnam et al. 2013).

## Assessing the quality of a *de novo* assembly

Regardless of whether a *de novo* assembly is genomic or transcriptomic, having no reference for comparison makes assessing the quality of a *de novo* assembly challenging. A number of techniques have had varying degrees of success in addressing this problem. The ‘N50’ value for an assembly is calculated by sorting assembly contigs in descending order and identifying the length of the contig containing 50% of the nucleotides of the entire sorted assembly (Paszkiewicz & Studholme 2010). This provides a metric where the quality of a *de novo* assembly can be assessed at a glance. However, the N50 does not provide insight regarding potential misassembly or chimeric contigs produced by the assembly algorithm (Salzberg & Yorke 2005). As the N50 is represented as a single number, assemblies that contain a few large contigs and numerous small ones are frequently misrepresented as being of good quality. Therefore, the N50 must not be taken as a single point of truth of an assembly’s quality, but simply as an attribute of the assembly to be used alongside others.

The most effective way to assess a transcriptome is to align good quality *a priori* sequences, such as transcripts that have been sequenced by Sanger techniques or ESTs, against the *de novo* assembly. The longer and more numerous the alignments, the better the quality of the assembly. Where reference sequences do not exist, the Core Eukaryotic Genes Mapping Approach (CEGMA) may be used instead. Aligning these 248 most common eukaryotic transcripts against the *de novo* genome can indicate assembly quality. The higher the number and score of alignments of *de novo* contigs against the conserved CEGMA sequences, the better the quality of the assembly (Parra et al. 2007).

# Aims of this research project:

This project will examine the vernalisation response in safflower. The overall goal is to observe how safflower responds to vernalisation and determine the underlying genetic pathways and molecular mechanisms responsible for this phenotypic expression. Three different approaches will be undertaken to achieve this. The first will be to examine and document the phenotypic expression of two different varieties of safflower, the first being a vernalisation responsive variety and the other being a vernalisation unresponsive variety. The second approach will sequence genomic and transcriptiomic material to create a *de novo* safflower genome and annotate it using the transcriptomic data. These *in silico* transcripts will also be used to conduct differential expression analysis on safflower to determine the regulation of transcripts in response to vernalisation. Single nucleotide polymorphisms will also be investigated to characterise difference between vernalisation responsive and unresponsive safflower. The third technique will use forward and reverse genetics techniques as well as traditional cross breeding of vernalisation responsive and unresponsive e safflower varieties to characterise transcripts and genomic regions critical to the vernalisation response in safflower. The results of this research will influence traditional breeding techniques and the genetic manipulation of safflower cultivars to allow development of safflower that can be grown in cooler climates and produces comparable yield when compared to current safflower varieties suited to warmer climates, with the end goal to diversify the regions of Australia where safflower can be successfully grown as a viable crop.

# References

Abou-Elwafa, S.F. et al., 2011. Conservation and divergence of autonomous pathway genes in the flowering regulatory network of *Beta vulgaris*. *Journal of Experimental Botany*, 62(10), pp.3359–3374.

Adams, M.D. et al., 1991. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, 252(5013), pp.1651–1656.

Adams, M.D. et al., 2000. The Genome Sequence of *Drosophila melanogaster*. *Science*, 287 (5461 ), pp.2185–2195.

Allain, L. et al., 2001. Microarray sampling-platform fabrication using bubble-jet technology for a biochip system. *Fresenius’ Journal of Analytical Chemistry*, 371(2), pp.146–150.

Amasino, R., 2004. Vernalization, Competence, and the Epigenetic Memory of Winter. *The Plant Cell Online*, 16(10), pp.2553–2559.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814), pp.796–815.

Bashir, A. et al., 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotech*, 30(7), pp.701–707.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59.

Boss, P.K. et al., 2004. Multiple Pathways in the Decision to Flower: Enabling, Promoting, and Resetting. *The Plant Cell Online*, 16(suppl 1), pp.S18–S31.

Boudry, P., McCombie, H. & Dijk, H. van, 2002. Vernalization Requirement of Wild Beet *Beta vulgaris* ssp. *maritima*: Among Population Variation and Its Adaptive Significance. *Journal of Ecology*, 90(4), pp.693–703.

Bradnam, K. et al., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1), p.10.

Buckley, B.A., 2007. Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays. *Journal of Experimental Biology*, 210(9), pp.1602–1606.

Burn, J.E. et al., 1993. DNA methylation, vernalization, and the initiation of flowering. *Proceedings of the National Academy of Sciences*, 90(1), pp.287–291.

Cao, S. et al., 2013. A large and functionally diverse family of Fad2 genes in safflower (*Carthamus tinctorius* L.). *BMC Plant Biology*, 13(1), p.5.

Chia, T.Y.P. et al., 2008. Sugar beet contains a large CONSTANS-LIKE gene family including a CO homologue that is independent of the early-bolting (B) gene locus. *Journal of Experimental Botany*, 59(10), pp.2735–2748.

Chouard, P., 1960. Vernalization and its Relations to Dormancy. *Annual Review of Plant Physiology*, 11(1), pp.191–238.

Corbesier, L. et al., 2007. FT Protein Movement Contributes to Long-Distance Signaling in Floral Induction of *Arabidopsis.* *Science*, 316(5827), pp.1030–1033.

Cronquist, A., 1980. *Asteraceae*. In *Vascular Flora of the Southeastern United States*. University of North Carolina Press.

Earl, D.A. et al., 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21, pp.2224-2241.

Eid, J. et al., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910), pp.133–138.

Evans, L.T., 1980. Response to Challenge: WIlliam Farrer and the Making of Wheats. *Journal of the Australian Institute of Agricultural Science*, 46(1), pp.3–13.

Finnegan, E.J. et al., 2005. The downregulation of FLOWERING LOCUS C (FLC) expression in plants with low levels of DNA methylation and by vernalization occurs by distinct mechanisms. *The Plant Journal*, 44(3), pp.420–432.

Finnegan, E.J. & Dennis, E.S., 2007. Vernalization-Induced Trimethylation of Histone H3 Lysine 27 at FLC Is Not Maintained in Mitotically Quiescent Cells. *Current Biology*, 17(22), pp.1978–1983.

Fletcher, J.C., 2002. Shoot and Floral Meristem Maintenance in *Arabidopsis*. *Annual Review of Plant Biology*, 53(1), pp.45–66.

Fodor, S.P.A. et al., 1993. Multiplexed biochemical assays with biological chips. *Nature*, 364, pp.555–556.

Garnatje, T. et al., 2006. Genome Size Variation in the Genus Carthamus (*Asteraceae, Cardueae*): Systematic Implications and Additive Changes During Allopolyploidization. *Annals of Botany*, 97(3), pp.461–467.

Gecgel, U. et al., 2007. Fatty Acid Composition of the Oil from Developing Seeds of Different Varieties of Safflower (*Carthamus tinctorius* L.). *Journal of the American Oil Chemists’ Society*, 84(1), pp.47–54.

Golembeski, G.S. et al., 2014. Chapter One - Photoperiodic Flowering Regulation in *Arabidopsis thaliana*. In F. F. B. T.-A. in B. Research, ed. *The Molecular Genetics of Floral Transition and Flower Development*. Academic Press, pp. 1–28.

Gray, S.G., 1942. Increased Earlieness of Flowering in Lettuce Through Vernalisation. *Journal of the Council for Scientific and Industrial Research, Australia*, 15(3), pp.211–212.

Greenup, A.G. et al., 2010. ODDSOC2 Is a MADS Box Floral Repressor That Is Down-Regulated by Vernalization in Temperate Cereals. *Plant Physiology*, 153(3), pp.1062–1073.

Griffiths, S. et al., 2003. The Evolution of CONSTANS-Like Gene Families in Barley, Rice, and *Arabidopsis.* *Plant Physiology*, 131(4), pp.1855–1867.

Hayama, R. & Coupland, G., 2004. The Molecular Basis of Diversity in the Photoperiodic Flowering Responses of *Arabidopsis* and Rice. *Plant Physiology*, 135(2), pp.677–684.

Huang, B.E. et al., 2012. A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnology Journal*, 10(7), pp.826–839.

Huang, X. & Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9), pp.868–877.

Illumina, 2012. Data Processing of Nextera Mate Pair Reads on Illumina Sequening Platforms. Available at: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote\_nextera\_matepair\_data\_processing.pdf [Accessed January 4, 2015].

Ingram, D.S., Vince-Prue, D. & Gregory, P.J., 2008. *Science and the Garden: The Scientific Basis of Horticultural Practice*, Wiley-Blackwell, West Sussex.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

Işigigür, A., Karaosmanoglu, F. & Aksoy, H.A., 1995. Characteristics of safflower seed oils of turkish origin. *Journal of the American Oil Chemists’ Society*, 72(10), pp.1223–1225.

Johanson, U. et al., 2000. Molecular Analysis of FRIGIDA, a Major Determinant of Natural Variation in *Arabidopsis* Flowering Time. *Science*, 290(5490), pp.344–347.

Johnson, R.C., Dajue, L. & Bradley, V., 2006. Autumn growth and its relationship to winter survival in diverse safflower germplasm. *Canadian Journal of Plant Science*, 86(3), pp.701–709.

Klippart, J.H., 1857. An essay on the origin, growth, diseases, varieties etc of the wheat plant. *Annual Report of the Ohio State Board of Agriculture*, pp.562–720.

Knights, S., 2010. *Raising the bar with better safflower agronomy*, Available at: http://www.australianoilseeds.com/\_\_data/assets/pdf\_file/0010/6949/GDRC\_Raising\_the\_Bar\_-\_Better\_Safflower\_Agronomy.pdf.

Knowles, P.F., 1949. Dual purpose safflower seed: Produces drying oil for the paint industry and seed meal for livestock and poultry feed. *California Agriculture*, 3(2), pp.11–16.

Knowles, P.F., 1960. Safflower’s Native Home. *Crops and Soils*, 12(6), p.17.

Köhler, C. & Villar, C.B.R., 2008. Programming of gene expression by Polycomb group proteins. *Trends in Cell Biology*, 18(5), pp.236–243.

Koltunow, A.M.G., Johnson, S.D. & Okada, T., 2011. Apomixis in hawkweed: Mendel’s experimental nemesis. *Journal of Experimental Botany*, 62(5), pp.1699–1707.

Kuleshov, V. et al., 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotech*, 32(3), pp.261–266.

Levy, Y.Y. et al., 2002. Multiple Roles of Arabidopsis VRN1 in Vernalization and Flowering Time Control. *Science*, 297(5579), pp.243–246.

Li, H. et al., 2012. De Novo Transcriptome of Safflower and the Identification of Putative Genes for Oleosin and the Biosynthesis of Flavonoids. *PLoS ONE*, 7(2), p.e30987.

Li, H. et al., 2011. Investigation of the microRNAs in safflower seed, leaf, and petal by high-throughput sequencing. *Planta*, 233(3), pp.611–619.

Li, R. et al., 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), pp.713–714.

Li, R. et al., 2010. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279), pp.311–317.

De Lucia, F. et al., 2008. A PHD-Polycomb Repressive Complex 2 triggers the epigenetic silencing of FLC during vernalization. *Proceedings of the National Academy of Sciences*, 105(44), pp.16831–16836.

Lulin, H. et al., 2012. The First Illumina-Based De Novo Transcriptome Sequencing and Analysis of Safflower Flowers. *PLoS ONE*, 7(6), p.e38653.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.

Macindoe, S.L. & Brown, C., 1968. *Wheat breeding and varieties in Australia* C. Walken, ed., Sydney.

Michaels, S.D., Bezerra, I.C. & Amasino, R.M., 2004. FRIGIDA-related genes are required for the winter-annual habit in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), pp.3281–3285.

Naim, F. et al., 2012. Advanced Engineering of Lipid Metabolism in *Nicotiana benthamiana* Using a Draft Genome and the V2 Viral Silencing-Suppressor Protein. *PLoS ONE*, 7(12), p.e52717.

Nakamura, Y. et al., 2014. Arabidopsis florigen FT binds to diurnally oscillating phospholipids that accelerate flowering. *Nat Commun*, 5.

National Centre for Biotechnology Information, 2014. Safflower Expressed Sequence Tags. Available at: http://www.ncbi.nlm.nih.gov/nucest/?term=safflower [Accessed January 7, 2015].

Oliver, S.N. et al., 2009. Vernalization-induced flowering in cereals is associated with changes in histone methylation at the VERNALIZATION1 gene. *Proceedings of the National Academy of Sciences*, 106(20), pp.8386–8391.

Oliver, S.N., Dennis, E.S. & Dolferus, R., 2007. ABA Regulates Apoplastic Sugar Transport and is a Potential Signal for Cold-Induced Pollen Sterility in Rice. *Plant and Cell Physiology*, 48(9), pp.1319–1330.

Owen, F. V., Carsner, E. & Stout, M., 1940. Photothermal induction of flowering in suagr beets. *Journal of Agricultural Research*, 61, pp.101–124.

Page, D.R. & Grossniklaus, U., 2002. The art and design of genetic screens: *Arabidopsis thaliana*. *Nat Rev Genet*, 3(2), pp.124–136.

Parra, G., Bradnam, K. & Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9), pp.1061–1067.

Paszkiewicz, K. & Studholme, D.J., 2010. De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11(5), pp.457–472.

Peltola, H., Söderlund, H. & Ukkonen, E., 1984. SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, 12(1 Pt 1), pp.307–321.

Périlleux, C. et al., 2013. A root chicory MADS box sequence and the *Arabidopsis* flowering repressor FLC share common features that suggest conserved function in vernalization and de-vernalization responses. *The Plant Journal*, 75(3), pp.390–402.

Pevzner, P.A., Tang, H. & Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp.9748–9753.

Pin, P.A. et al., 2010. An Antagonistic Pair of FT Homologs Mediates the Control of Flowering Time in Sugar Beet. *Science*, 330(6009), pp.1397–1400.

Pin, P.A. et al., 2012. The Role of a Pseudo-Response Regulator Gene in Life Cycle Adaptation and Domestication of Beet. *Current Biology*, 22(12), pp.1095–1101.

Purushothaman, S., Toumazou, C. & Ou, C.-P., 2006. Protons and single nucleotide polymorphism detection: A simple use for the Ion Sensitive Field Effect Transistor. *Sensors and Actuators B: Chemical*, 114(2), pp.964–968.

Quail, M. et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1), p.341.

Rappapport, L., Wittwer, S.H. & Tukey, H.B., 1956. Seed Vernalization and Flowering in Lettuce (Lactuca sativa). *Nature*, 178(4523), p.51.

Ream, T.S. et al., 2014. Interaction of Photoperiod and Vernalization Determines Flowering Time of *Brachypodium distachyon*. *Plant Physiology*, 164(2), pp.694–709.

Reeves, P.A. et al., 2007. Evolutionary Conservation of the FLOWERING LOCUS C-Mediated Vernalization Response: Evidence From the Sugar Beet (*Beta vulgaris*). *Genetics*, 176(1), pp.295–307.

Rezvani, M. et al., 2000. Apoptosis-related genes expressed in cardiovascular development and disease: an EST approach. *Cardiovascular Research*, 45(3), pp.621–629.

Salisbury, F.B. & Ross, C.W., 1992. *Plant Physiology*, Wadsworth Publishing Company.

Salzberg, S.L. & Yorke, J.A., 2005. Beware of mis-assembled genomes. *Bioinformatics*, 21(24), pp.4320–4321.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), pp.441–448.

Sheldon, C.C. et al., 2000. The molecular basis of vernalization: The central role of FLOWERING LOCUS C (FLC). *Proceedings of the National Academy of Sciences*, 97(7), pp.3753–3758.

Sikora, P. et al., 2011. Mutagenesis as a tool in Plant Genetics, Functional Genomics and Breeding. *International Journal of Plant Genomics*, p.13.

Simpson, G.G., 2004. The autonomous pathway: epigenetic and post-transcriptional gene regulation in the control of Arabidopsis flowering time. *Current Opinion in Plant Biology*, 7(5), pp.570–574.

Simpson, J.T. et al., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), pp.1117–1123.

Sung, S. & Amasino, R.M., 2004. Vernalization and epigenetics: how plants remember winter. *Current Opinion in Plant Biology*, 7(1), pp.4–10.

The *C. Elegans* Sequencing Consortium, 1998. Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science*, 282(5396), pp.2012–2018.

Trevaskis, B. et al., 2006. HvVRN2 Responds to Daylength, whereas HvVRN1 Is Regulated by Vernalization and Developmental Status. *Plant Physiology*, 140(4), pp.1397–1405.

Trevaskis, B. et al., 2007. The molecular basis of vernalization-induced flowering in cereals. *Trends in Plant Science*, 12(8), pp.352–357.

United Nations, 2014. FAOSTAT. Available at: http://data.fao.org/dataset-data-filter?entryId=29920434-c74e-4ea2-beed-01b832e60609&tab=data.

Utturkar, S.M. et al., 2014. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30(19), pp.2709–2716.

Warne, L.G.G., 1947. Vernalization of Lettuce. *Nature*, 159(4027), pp.31–32.

Warren, W.C. et al., 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192), pp.175–183.

Waycott, W., 1995. Photoperiodic Response of Genetically Diverse Lettuce Accessions. *Journal of the American Society for Horticultural Science*, 120(3), pp.460–467.

Wetterstrand, K., 2014. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: http://www.genome.gov/sequencingcosts/ [Accessed December 27, 2014].

Yan, L. et al., 2004. The Wheat VRN2 Gene Is a Flowering Repressor Down-Regulated by Vernalization. *Science*, 303(5664), pp.1640–1644.

Yazdi-Samadi, B. & Zali, A.A., 1979. Comparison of Winter- and Spring-Type Safflower1. , pp.783–785.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zimmerman, L.H., 1973. Effect of Photoperiod and Temperature On Rosette Habit in Safflower. , pp.80–81.

Zohary, D. & Hopf, M., 1993. *Domestication of Plants in the Old World* 2nd Ed., Oxford Scientific Publications.