The genomic and evolutionary analysis of floral heteromorphy in *Primula*

Jonathan Matthew Cocker

Thesis submitted for the degree of Doctor of Philosophy

University of East Anglia

John Innes Centre

September 2016

Abstract

The genetic basis and evolutionary significance of floral heteromorphy in *Primula* has been debated for over 150 years. Charles Darwin was the first to explain the importance of the two heterostylous floral morphs, pin and thrum, suggesting that their reciprocal anther and stigma heights facilitate cross-pollination, and showing that only between morph crosses are fully compatible. This key innovation is an archetypal example of convergent evolution that serves to physically promote insect-mediated outcrossing, having evolved in over 28 angiosperm families.

Darwin's findings laid the foundation for an extensive number of studies into heterostyly that contributed to the establishment of modern genetic theory. The widely accepted genetic model portrays the *Primula S* locus, which controls heterostyly and self-incompatibility, as a coadapted group of tightly-linked genes, or supergene. It is predicted that self-fertile homostyle flowers, with anthers and stigma at the same height, arise via rare recombination events between dominant and recessive alleles in heterozygous thrums. These observations have underpinned over 60 years of research into the genetics and evolution of heterostyly.

The *Primula vulgaris* genome assembly and associated transcriptomic and comparative sequence analyses have facilitated the assembly and characterisation of the complete *S* locus in this species. Here it is revealed that thrums are hemizygous not heterozygous: the *S* locus contains five thrum-specific genes which are completely absent in pins, which means recombination cannot be the cause of homostyles as previously believed. The studies also reveal candidate genes in *Primula veris* and other species, and have facilitated an estimation for the assembly of the *S* locus supergene at 51.7 MYA. These findings challenge established theory, and reveal novel insight into the structure and origin of the *Primula S* locus, providing the foundation for understanding the evolution and breakdown of insect-mediated outcrossing in *Primula* and other heterostylous species.

Contents

Acknowledgements

I would like to thank my supervisor Phil Gilmartin for facilitating my development as a scientist, providing support and encouragement throughout this doctoral project, and giving me the freedom to develop and discuss ideas that have led to the results presented herein. In particular, I thank Jinhong Li for countless discussions, support and the generation of numerous resources, without which much this work would not have been possible. I also thank fellow final year PhD student Olivia Kent for phylogenetic resources and support throughout, including numerous discussions. I thank all members of the Gilmartin lab and others at JIC who have supported this work. The collaborative nature of some elements to this project mean that I am indebted to a number of collaborators, including Jon Wright who helped me to develop as a bioinformatician and carried out many genomic-based analyses that support this work, Mark McMullan for support in phylogenetic analyses, and my third supervisor David Swarbreck and other colleagues at TGAC for input into genomic analyses. I also acknowledge the input of Cock van Oosterhout, whose supervision, enthusiasm and encouragement was invaluable throughout manuscript submission and many of the analyses conducted. I thank my mentors David Westhead and Peter Meyer at the University of Leeds, and others who led me in this direction.

The friends and colleagues I have connected with over the last four years have provided an inspiring and friendly environment in which to work, and in which to unwind. I thank all fellow PhD students, especially friends in my year, and those I have met outside of JIC. I thank my family for encouraging me to follow anything I set my mind to, and for providing a platform from which I was able to take those steps. I cannot thank them enough for the support they have provided over the years. Finally, I thank Emily for immeasurable support and reassurance, for walks and stargazing, and always being there to listen.

Abbreviations

A	*Androecium S* locus allele
AHRD	Automated Assignment of Human Readable Descriptions
ARF	Auxin response factor
as1	*asymmetric leaves 1*
ASGR	Apomixis-specific genomic region
BACs	Bacterial artificial chromosomes
CEGs	Core eukaryotic genes
CFB	*Cyclin-like F box*
cM	Centimorgans
CRT	Cyclic reversible terminator
DEF	*DEFICIENS*
DSB	Double-stranded break
ESS	Effective Sample Size
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
G	*Gynocieum S* locus allele
GLO	*GLOBOSA*
GO	Gene Ontology
GSI	Gametophytic self-incompatibility
HGP	Human Genome Project
HR	Homologous recombination
HSPs	High-scoring segment pairs
ID	Identity
JL	Jinhong Li
JMC	Jonathan Matthew Cocker
JW	Jon Wright
KAT	K-mer Analysis Toolkit
k-mer	sequence of *k* length
KNOX	*KNOTTED-like homeobox*
LH	Long homostyle
LMP	Long mate-pair
LTR	Long terminal repeat
MCMC	Markov Chain Monte Carlo
miRNAs	microRNAs
MYA	Million years ago
NGS	Next generation sequencing
NHEJ	Non-homologous end-joining
ONT	Oxford Nanopore Technologies
OVCK	Olivia Victoria Constance Kent
P	*Pollen S* locus allele
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PMG	Philip Mark Gilmartin
PP	Pin parent
PvKNL	*Primula vulgaris KNOTTED*-like
PvSTL	*Primula vulgaris SHOOT MERISTEMLESS*-like

SFG^{L}	*S* locus *Flanking Gene Left*
SFG^{R}	*S* locus *Flanking Gene Right*
SH	Short homostyle
SI	Self-incompatibility
SNP	Single nucleotide polymorphism
SRY	Sex-determining region Y
SSI	Sporophytic self-incompatibility
TDF	Testis-determining factor
TEs	Transposable elements
TF	Transcription factor
TGAC	The Genome Analysis Centre
TP	Thrum parent
UTR	Untranslated region
WGS	Whole genome sequencing

List of publications

Li, J.*, **Cocker, J.M.*,** Wright, J., Webster, M.A., McMullan, M., Ayling, S., Swarbreck, D., Caccamo, M., Oosterhout, Cv., Gilmartin, P.M. (2016) Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nature plants*, 2: 16188.

Cocker, J.M.*, Webster, M.A.*, Li, J., Wright, J., Kaithakottil, G., Swarbreck, D., Gilmartin, P.M. (2015) *Oakleaf*: an *S* locus-linked mutation of *Primula vulgaris* that affects leaf and flower development. *New Phytologist,* 208: 149–161.

Li, J., Webster, M.A., Wright, J., **Cocker, J.M.,** Smith, M.C., Badakshi, F., Heslop-Harrison, P., Gilmartin, P.M. (2015) Integration of genetic and physical maps of the *Primula vulgaris S* locus and localization by chromosome *in situ* hybridization. *New Phytologist*, 208: 137–148.

Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Dyer, S., Caccamo, M., Gilmartin, P.M. (2017) The *Primula vulgaris* genome (in preparation).

* These authors contributed equally

1

Introduction

1.1 Heterostyly in *Primula*

The floral architecture and breeding habits of *Primula* have been studied for over 150 years (Darwin, 1862, Darwin, 1877). Charles Darwin explained the importance of precise floral organ positioning as a physical method of promoting insect-mediated outcrossing (Darwin, 1862, Darwin, 1876, Darwin, 1877, Gilmartin, 2015). Recent advancements in genomics technology coupled with classical genetics approaches are allowing many of the ideas initially put forward by Darwin to be investigated in a holistic manner (Cohen, 2010). Thus, a bioinformatics project focussing on heterostyly in *Primula* is distinct in that it combines the historical with the cutting-edge: the application of next-generation sequencing (NGS) innovations has revolutionised biological research (van Dijk et al., 2014).

In the phenomenon known as floral heteromorphy, or heterostyly, each primrose plant has one of two forms of flower, the "pin" (long-styled) or the "thrum" (short-styled) (Darwin, 1877, Richards and Barrett, 1992). Darwin showed that intra-morph "illegitimate" fertilizations result in a much-reduced seed set, producing around 20-50% of the seed that results from so-called "legitimate" pin-thrum or thrum-pin cross-fertilizations (Darwin, 1877). He concluded that "seedlings raised from such [illegitimate] unions" were "in some degree sterile, dwarfed, and feeble" and was aware of such effects in other species, going so far as to dedicate an entire volume to the subject, *The Effects of Cross and Self-fertilisation in the Vegetable Kingdom* (Darwin, 1876, Darwin, 1877). Floral innovations and associated interactions with pollinators have since captured the imagination of evolutionary biologists, with the genetic effect of reproductive systems and the high evolvability that results thought to be a driving force

behind the remarkable species-diversity of the angiosperms, outstripping that of their sister group more than 250-fold (de Vos et al., 2014, Puttick et al., 2015).

The development of the two heterostylous floral forms (distyly) in *Primula* is controlled by a coadapted linkage group, or "supergene", known as the *S* locus (Gregory et al., 1923, Gilmartin, 2015), which through the work of Ernst (1955) and others, was predicted to comprise at least three tightly-linked genes (Lewis, 1954, Dowrick, 1956, Lewis and Jones, 1992): the dominant *G* (*Gynocieum*) allele represses style cell elongation, the *P* (*Pollen*) allele acts to increase pollen size, and the *A* (*Androecium*) allele promotes cell division in the anthers (Ernst, 1936c, Lewis and Jones, 1992, Webster and Gilmartin, 2006). Together, the dominant alleles of the three genes (*GPA*) result in flowers with a short style, large pollen, and anthers situated at the mouth of the corolla tube; this morphology is known as the thrum, so-called due to the anthers at the mouth of the floral tube resembling "the ends of weavers' threads" (Darwin, 1877). The thrum is thought to be heterozygous at the *S* locus (*S*/*s*) with the three dominant genes linked in coupling (*GPA*/*gpa*). The pin form, with the protruding stigma resembling the head of a pin, is recognised as homozygous recessive (*gpa*/*gpa*) at the *S* locus (*s*/*s*) and has the opposite morphology: a long style, small pollen, and anthers situated midway-down the corolla tube.

In this way, the stigma and anthers of two floral morphs are in complementary positions (Figure 1.1). The reciprocal architectures suggest that pollen is transferred to the proboscis or body of an insect from the anthers of one form of flower in such a way that it is transferred to the stigma of the other (Darwin, 1877).

Figure 1.1 The reciprocal floral architectures of *Primula vulgaris* pin (*s*/*s*) and thrum (*S*/*s*) heterostylous flowers; stigma and anthers are indicated, as well as the genotype of each morph for the three tightly-linked genes predicted to lie at the *S* locus. Figure reproduced from Cocker et al. (2017).

In addition, an associated biochemical self-incompatibility (SI) system acts as a safeguard against self-fertilization. It is predicted that rare recombination events between the genes at the *S* locus results in flowers with the anthers and stigma situated at the same height (Charlesworth and Charlesworth, 1979a, Li et al., 2015a); these flowers are termed homostyles and are associated with a breakdown in self-incompatibility (Figure 1.2). The above predictions form the backdrop to the last 60 years of research into heterostyly.

Figure 1.2 Disruption of the *S* locus "supergene" results in homostyle flowers with anthers and stigma at the same height; the short homostyle in this photo is in a *Hose in Hose* background, a *GLOBOSA* overexpression mutant that results in conversion of sepals to petals (see Chapter 3). Figure reproduced from Li et al. (2016).

1.2 Heteromorphic self-incompatibility

The majority of flowering plants (~90%) are hermaphrodite, with flowers that have both male and female reproductive structures; as such, they frequently employ diverse mechanisms to prevent self-fertilization through the presence of tightly-linked genes that control the pollen and pistil identity of specific mating types (Allen and Hiscock, 2008, McCubbin, 2008, Barrett and Hough, 2012). In *Primula*, a di-allelic biochemical SI system reinforces the morphological constraints of the heterostylous floral forms by reducing self and intra-morph fertilization through the action of pin- and thrum-specific molecules that interact on or in the stigma (McCubbin, 2008); genetic evidence shows that the pollen and pistil specificities are controlled by separate genes (Dowrick, 1956, Kurian and Richards, 1997). The *GPA* locus could be extended to include two additional genes for pollen and style incompatibility, but it remains that these genes could be regulated by the *S* locus genes rather than linked to them (Barrett, 2008).

The pollen of each *Primula* morph can only fertilise the opposite form effectively (Bateson and Gregory, 1905). Nonetheless, the mechanism has been described as "leaky", with self-incompatibility being stronger in thrum-thrum crosses than in pin, as

observed through comparison of the relative size of resulting seed sets (Darwin, 1877, Ganders, 1979, McCubbin, 2008). Fully uninhibited pollen tube growth occurs only in inter-morph crosses (McCubbin, 2008).

In contrast to heteromorphic SI, the more frequently occurring gametophytic self-incompatibility (GSI) and sporophytic self-incompatibility (SSI) mechanisms are so-called homomorphic SI systems where the flowers of the different mating types are morphologically indistinguishable (Charlesworth, 2010). These homomorphic SI systems are regulated by one or more multi-allelic SI loci (McCubbin, 2008). In contrast, distylous species are characterised as having only two incompatibility types; they are di-allelic (Charlesworth, 2010). There are also heterostylous species that are tri-morphic such as *Lythrum*, where there are three mating types associated with mid-, long- and short-styled morphs (tristyly), with the morphological features potentially under the control of two supergene loci, *S* and *M* (Richards and Barrett, 1992, Charlesworth, 2010).

It is possible to gain some insight into the nature of a species' SI system based on the site of the incompatibility reaction. SSI systems generally show pollen inhibition on the stigma surface, whereas inhibition of incompatible pollen tubes typically occurs within the style in GSI (Allen and Hiscock, 2008); an exception is pollen rejection triggered by a stigma-specific glycoprotein in the GSI system of poppy (Hiscock, 2002). The site of the self-incompatibility reaction appears to differ between heterostylous species in different families but also between species of *Primula*, reportedly occurring on the stigmatic surface, within the stigma, style, or ovary, and even conflicting locations between distinct floral morphs of the same species (Dulberger, 1964, Shivanna et al., 1981, Wedderburn and Richards, 1990, Lewis and Jones, 1992, McCubbin, 2008, Cohen, 2010).

In thrums, the majority of intra-morph pollen fails to germinate on or in the stigma, consistent with SSI, whilst within-pin pollinations can infrequently result in self pollen tubes that penetrate the stigma and reach as far as the stylar transmitting tract (Darwin, 1877, Baker, 1966, Shivanna et al., 1981). This is not necessarily the result of multiple sites of inhibition. Shivanna et al. (1981) showed that there are multiple sites at which growth of the pollen tube ceases, but it could be that the incompatibility reaction does indeed occur on or in the stigma, but varies in effectiveness such that pollen tube

growth stops at different points, with the SI reaction occurring at different speeds (McCubbin, 2008).

Based on such studies, a frequent assertion is that heteromorphic SI might function in a sporophytic manner, with inhibition on the stigma surface (Baker, 1966, Gibbs, 1986, Cohen, 2010). However, another conclusion is that SSI might function in the short-styled thrum morph, and either sporophytic or gametophytic SI in the pin morph. It has even been suggested that multiple gene products could be involved in several different SI mechanisms that act in tandem (Golynskaya et al., 1976, Wedderburn and Richards, 1990, McCubbin, 2008).

Further studies in the heterostylous *Luculia gratissima* (Rubiaceae) suggest inhibition occurs at the stigma. Stevens and Murray (1982) showed that, after removal of the stigma, in pin styles 33% vs. 21% of pollen tubes from legitimate vs. illegitimate pollinations grew from the top of the style to the bottom, and in thrum 66% vs. 72%. This suggests that the female SI determinant is primarily located in the stigma of both morphs, since its removal appears to reduce inhibition of illegitimate pollen tubes, as seen in the comparable percentages above. This effect appears to be clearer in the thrum morph, however, and the possibility of a second site of inhibition remains in pin, where there is a larger difference in the percentages of pollen tubes growing; this could potentially result from expansion of the spatial and temporal expression of the *s* gene products (McCubbin, 2008). It remains that more research needs to be to done reveal the SI components themselves: despite concerted efforts to identify morph-specific genes and associated SI determinants, there is currently no molecular evidence to support the above proposals (McCubbin et al., 2006, Li et al., 2015a).

1.3 The role of heterostyly

The major drawback of a di-allelic self-incompatibility (SI) system, as recognised by Charles Darwin, is that it renders a plant incapable of fertilizing “half its brethren” (Darwin, 1877). In addition to intra-flower pollen deposition, the absence of cross-pollination means that 50% of the incoming pollen will be incompatible based on a typical 1:1 ratio of pin to thrum plants in the wild (Ornduff, 1979, Barrett and Shore, 2008, McCubbin, 2008). Lloyd and Webb (1992b) supported this point of view through

a quantitative model that suggests the problem is great enough such that di-allelic self-incompatibility could not have been selected for prior to the onset of heterostyly. In contrast, homomorphic SI systems comprise multiple mating-types controlled by multi-allelic loci, which implies that much of the pollen that is brought to a stigma will be compatible, therefore resulting in adequate compatible pollen for a good seed set (McCubbin, 2008).

Darwin suggested that heterostyly preceded self-incompatibility; he was forced to regard the evolution of self-incompatibility as an "incidental and purposeless" by-product of the constraints already imposed by heterostyly (Darwin, 1877). Indeed, there are numerous species in primarily heterostylous groups without a self-incompatibility system, but not vice versa (Lloyd and Webb, 1992a), suggesting heterostyly is not a redundant mechanism despite the proposition that a self-incompatibility system in heterostylous species could seemingly act to guarantee outcrossing (Barrett and Shore, 2008). Furthermore, heterostyly is present in at least 28 families and is thought to have evolved independently on over 23 occasions (Barrett, 1992, Lloyd and Webb, 1992a, Naiki, 2012), which implies that it must confer some sort of selective advantage.

The purpose of heterostyly then, appears to be in the promotion of cross-pollination, such that pollen (the male gametophyte) is not wasted through deposition on incompatible stigmas, thereby promoting the male component of fitness through the reduction of "pollen wastage" (Lloyd and Webb, 1992b, Barrett and Shore, 2008). This suggests heterostyly results in a greater proportion of the pollen deposited on a stigma being compatible (Lloyd and Webb, 1992b), therefore mitigating against the detrimental effect that a di-allelic SI system would have on paternal fitness in preventing fertilization with half the population.

Yeo (1975) also proposed that heterostyly could prevent "stigma clogging" effects whereby the access of legitimate pollen might be hindered by illegitimate pollen deposited on stigmas; this hypothesis receives some support in evidence of reduced seed set in some species through application of illegitimate and legitimate pollen on the same stigmas, but other analyses suggest that it has no obvious role in the evolution of distyly (Piper and Charlesworth, 1986, Lloyd and Webb, 1992b).

If heterostyly is in place, and functioning to reduce "pollen wastage" and any other potentially disadvantageous effects of illegitimate pollination, the female component of fitness could then be increased by safeguarding against self-fertilization and reducing inbreeding depression through the action of heteromorphic self-incompatibility. In this way, neither mechanism is redundant. Darwin's early adaptive explanation for the function of heterostyly in promoting insect-mediated cross-pollination between floral morphs is supported, not in spite of, but as a complementary mechanism that seemingly makes the unprecedented evolution of di-allelic SI possible (Darwin, 1877, Lloyd and Webb, 1992b, Barrett and Shore, 2008).

1.4 Morphology of the heterostylous floral morphs

In addition to dimorphic stigma and anther positioning, a range of ancillary features occurs in different *Primula* species. This includes differences in style cross-sectional area, stigma size and shape, pollen size and amount, and corolla mouth size (Dowrick, 1956, Webster and Gilmartin, 2006, McCubbin, 2008). In *Primula* species the anther filaments are fused to the corolla tube (McCubbin, 2008). Differential anther positioning is brought about by increased cell division in the lower corolla tube of thrum flowers. This is accompanied by wider cells in the upper corolla that presumably accommodate the above change by producing a larger corolla mouth (Webster and Gilmartin, 2006).

It has been suggested that developmentally interconnected traits, such as the distinct morphological features pertaining to the style and stigma, or those associated with the pollen or anthers, are probably controlled by single genetic factors (Richards, 1997, Webster and Gilmartin, 2006, McCubbin, 2008). Correspondingly, the observation of recombinants that show disruption of the linkage between specific features and their associated floral morph led to a model that describes three genes at the *S* locus in the order *GPA* (Lewis and Jones, 1992). In some cases recombinants with altered ancillary features would be difficult to observe, but there is no such issue in relation to pollen grain size: the thrum pollen grains are just over two-fold larger in volume than those in pin flowers, whilst the production of pollen in pin is slightly more than two-fold greater than in thrum (Darwin, 1877, Dowrick, 1956, Ornduff, 1979, Piper and Charlesworth, 1986, Barrett and Shore, 2008).

The importance of the above features in the operation of heteromorphic SI is debated, with some suggestions that pollen size and stigmatic papillae may be important (Dulberger, 1975, Darwin, 1877, McCubbin, 2008). Darwin (1877) proposed that the larger pollen of the thrum might provide greater energy reserves suitable for pollen tube growth to the base of the elongated pin style. However, this is refuted by Kurian and Richards (1997) who describe a mutant with a mix of small and large sized pollen that retains thrum pollen specificity in the pollination of pin stigmas. In some species, pollen only adheres to the stigma in inter-morph pollinations, resulting in suggestions that stigmatic papillae length might play a role in the SI reaction, but it seems this is more likely to be the result of biochemical rather than mechanical interaction (Richards, 1997, McCubbin, 2008).

Despite much debate on their potential roles, the remarkably consistent appearance and convergent evolution of the various morph-specific traits across diverse angiosperm species suggests that many of these features may be important in the function of heterostyly (McCubbin, 2008). It seems likely that the morph-associated morphological characters do not directly participate in the active self-recognition and rejection reaction between the pollen and pistil of the reciprocal morphs; perhaps these ancillary features are the result of coadaptation between the interacting surfaces, for the purpose of maximising legitimate and minimising illegitimate transfer of pollen in subtle ways, through the pleiotropic effect of genes at the *S* locus. It is intuitive to suggest that reciprocal anther and stigma positioning might bring about cross-pollination, the role of the ancillary features might be to further enhance this in support of precise pollinator interactions (Darwin, 1877, Lloyd and Webb, 1992a, McCubbin, 2008).

1.5 Inter-morph transfer of pollen

The distinct difference in pollen size between the two heterostylous morphs provides a direct means of assessing pollen loads. The vast majority of heterostylous species undergo biotic (insect) pollination (Lloyd and Webb, 1992). Darwin physically inserted the proboscises of dead bees, as well as bristles and needles, into the corolla tubes of *Primula* flowers in an attempt to ascertain whether heterostyly might result in morph-specific positioning of pollen on an insect body during nectar feeding (Darwin, 1877). In agreement with the positions of the anthers in the two floral morphs, he observed that

the large thrum pollen was largely positioned near the base, and the small pin pollen near the extremity of the proboscis, with some degree of intermingling of pollen types in between.

In support of Darwin's hypothesis of insect-mediated cross-pollination (Darwin, 1877), others have subsequently shown differential deposition of the two pollen types onto different parts of insect bodies through examination of insects captured whilst foraging on heterostylous species, including those in the genus *Fagopyrum*, *Pulmonaria* and *Cratoxylum* (Rosov and Screbtsova, 1958, Olesen, 1979, Lewis, 1982, Lloyd and Webb, 1992b) as well as the tristylous species *Pontederia cordata* (Barrett and Wolfe, 1986, Lloyd and Webb, 1992b).

It has been suggested that the analysis of stigmatic pollen loads only moderately supports Darwin's theory, with excess illegitimate pollen observed on stigmatic surfaces when examining intact flowers in some species (Ganders, 1979, Piper and Charlesworth, 1986, Lloyd and Webb, 1992b). However, it is pointed out that intra-flower pollination would inherently cause more pollen of the same morph to be deposited on a stigma than expected by random pollination (Ganders, 1979, Piper and Charlesworth, 1986, Lloyd and Webb, 1992b). This suggests that analysis of intact flowers says little about pollen flow between plants: intact flowers alone are an unsatisfactory resource for statistically ascertaining the level of disassortative pollination; that is, pollination between flower types (Piper and Charlesworth, 1986, Lloyd and Webb, 1992b). To this end, a number of studies including both intact and emasculated flowers of various species (e.g. *Primula*) revealed a reduction in the proportion of own-morph pollen deposited on the stigmas of emasculated flowers, suggesting that intra-flower pollination does indeed contribute significantly to the observed pollen load on stigmas (Ganders, 1974, Ganders, 1976, Piper and Charlesworth, 1986, Lloyd and Webb, 1992b).

Lloyd and Webb (1992) reanalysed data on pollen loads for stigmas of both intact and emasculated flowers for both floral morphs of the distylous species *Jepsonia heterandra* (Ganders, 1974), accounting for the observation that pin stigmas receive more pollen than thrum stigmas in total, perhaps due to the elevated stigma being more accessible in the precise entry and exit paths of pollinators during nectar feeding (Ganders, 1979, Lloyd and Webb, 1992b, Stone, 1995, Barrett and Shore, 2008). In doing so, the

probability of a pollen grain of each type being deposited on a stigma of each type was analysed, removing the effect of unequal total flower and pollen numbers, and considering the pin and thrum as either maternal or paternal parent in separate crosses. If emasculated flowers are considered and the distorting effects of self-pollination therefore removed, there is an overall twofold greater probability of legitimate transfer in *Jepsonia heterandra*, thus supporting Darwin's cross-pollination hypothesis (Lloyd and Webb, 1992b). In the reanalysis of pollen load data for the tristylous *Pontederia cordata* (Barrett and Glover, 1985) it is shown that all three morphs are more likely to receive legitimate pollen grains, and all three pollen types are more likely to be deposited on legitimate stigmas (Lloyd and Webb, 1992b) thus confirming the above findings.

Piper and Charlesworth (1986) used a natural population of *Primula vulgaris* to analyse factors that could promote the evolution of distyly which were considered in the theoretical framework set out by Charlesworth and Charlesworth (1979). These factors include selection for disassortative pollination, and selection for a reduction in self-pollination, stigma clogging, or pollen wastage. The analysis revealed significantly more legitimate than illegitimate pollen on emasculated flowers, which supports disassortative pollination; they also showed that pin and thrums receive a fifth less self-pollen than long homostyles, suggesting heterostyly also reduces self-pollination to a degree (Lloyd and Webb, 1992b). The small pollen saving effect found was proposed to be biologically insignifciant, and stigma clogging is thought to be unimportant (Piper and Charlesworth, 1986). In summary, the results of various studies taken together with similar considerations to the above confirm that heterostyly promotes cross-pollination across diverse angiosperm families; there are some suggestions for the precise mechanism of transfer, how insect-mediated outcrossing might be physically achieved, and the consequences of pollinator foraging behaviours (Piper and Charlesworth, 1986, Lloyd and Webb, 1992b, Stone, 1995, Harder and Wilson, 1998, Richards et al., 2009, Keller et al., 2014, Zhou et al., 2015).

1.6 Models for the evolution of heterostyly

In defining a species as heterostylous, the presence of two or more floral forms with anthers and stigma at complementary heights (reciprocal herkogamy) is considered

essential (Ganders, 1979, Lloyd and Webb, 1992a). Lloyd and Webb (1992a) suggested that the more varied ancillary features (section 1.4) probably evolved after reciprocal stigma and anther positioning. For this reason, less attention is paid to these traits, and the main focus of theoretical models for the evolution of heterostyly is in establishing the order of the reciprocal architectures and SI, as well as determining the intermediate states in between. It is notable, however, that di-allelic SI is absent in species that have reciprocal herkogamy much more often than the reverse scenario, which means SI need not be present for a species to qualify as heterostylous. This perhaps emphasises the importance of reciprocal herkogamy as a means for promoting cross-pollination (section 1.5) (Lloyd and Webb, 1992a).

In line with the considerable depth of historical research conducted into heterostyly over the course of the last 150 years (section 1.9) there has been a broad range of predictions for the origins of this breeding system (Ernst, 1936a, Mather and De Winton, 1941, Baker, 1966, Charlesworth and Charlesworth, 1979b, Lloyd and Webb, 1992a). Darwin (1877) suggested that reciprocal herkogamy evolved first due to selection for increased precision in reciprocal pollen transfer, thus bringing the stigma and anthers into complementary positions in the two morphs. He postulated that self-incompatibility and other ancillary traits followed reciprocal herkogamy as an incidental by-product of coadaptation between compatible pollen and style. Mather and De Winton (1941) suggested heterostyly and SI arose together in a single step, whilst Ernst (1936a) proposed a series of mutations starting from a homostyle ancestor, but no reasons were given for how such mutations would spread in the population due to a selective advantage (Charlesworth and Charlesworth, 1979b). The majority of other postulates support the evolution of SI prior to reciprocal herkogamy (Charlesworth and Charlesworth, 1979b).

There are two major competing quantitative models considering the evolution of reciprocal herkogamy and di-allelic self-incompatibility; that of Charlesworth and Charlesworth (1979b) and Lloyd and Webb (1992a). Lloyd and Webb (1992a) suggest that heterostyly evolved prior to SI (see below). In contrast, Charlesworth and Charlesworth (1979b) proposed a model that supports two ancestral steps in the evolution of self-incompatibly from a long homostyle-like progenitor that has a long style and elevated anthers. Firstly, a transition to an effectively male-sterile plant based

on mutation to a new pollen type; this pollen mutant is unable to fertilise progenitor-type individuals or itself, and would spread as a result of increased female fitness due to the elimination of self-fertilization. Secondly, a mutation to a new stigma type in the progenitor that is compatible with the new pollen type, thus establishing two mating-types that are reciprocally compatible. This model requires that the second gene is linked to the first, and thus encompasses linkage constraints. For this reason the model is seen to support the supergene hypothesis, such that a "translocation bringing the genes together would be selected for" (see section 1.9) (Charlesworth and Charlesworth, 1979b). It was proposed that a stigma-height dimorphism would then emerge, perhaps due to further reduction in self-fertilization or reduction of pollen wastage. This was followed by the subsequent establishment of heterostyly via a change in anther position in response to selection for disassortative pollination.

The homostyle progenitor is proposed based on a number of monomorphic species within the predominantly heterostylous taxa Plumbaginaceae (Baker, 1966). Baker (1966) revealed species representing possible stages in the evolution of distyly based on a taxonomic analysis of this group, including some homostylous species that retain self-incompatibility (Charlesworth and Charlesworth, 1979b, Ganders, 1979). From this, Baker (1966) concluded that the dimorphisms were imposed sequentially from a homostylous ancestor, with di-allelic self-incompatibility being the first step; a view that is quantitatively supported (Charlesworth and Charlesworth, 1979b). In the Charlesworth and Charlesworth (1979b) model the long homostyle is favoured over the short homostyle as a progenitor based on reciprocal crossing behaviour with pins and thrums (see section 1.9) (Charlesworth and Charlesworth, 1979b). This occurs as most homostyle species retain self-incompatibility reactions; homostyles with no self-incompatibility reaction have been interpreted as being the primitive state (Baker, 1966).

In contrast to Darwin's postulate, proposals such as the above are often based on the view that the evidence for promotion of cross-pollination by heterostyly is insubstantial, and that it confers a secondary selective force as compared to the strong effect of SI in preventing selfing (Lloyd and Webb, 1992a). This view was challenged by Lloyd and Webb (1992b) (section 1.5) and receives support from a reduction of self-pollen seen on the stigmas of homostyles (Piper and Charlesworth, 1986). It has often been assumed

that the ancestral morph was homostylous based on the occurrence of homostyle species in predominantly heterostylous clades, as indicated above (Baker, 1966, Charlesworth and Charlesworth, 1979b, Lloyd and Webb, 1992a). However, it is reasonable to suggest that these self-pollinating species may be so-called "secondary homostyles" that have derived from founding heterostylous species on multiple occasions, perhaps due to selection for reproductive assurance based on unreliable pollinator service (Mast et al., 2006, Barrett and Shore, 2008). Furthermore, where SI reactions are absent they may have arisen due to relaxation of the selective pressure to maintain them in homostyle species that are predominantly self-fertilizing, rather than this being the primitive state (Ganders, 1979).

Lloyd and Webb (1992a) proposed a contrasting model based on an ancestor with a pin-like approach herkogamous morphology that comprises stigma(s) situated above the anthers. This ancestral morph was suggested based on a comprehensive sampling of character states across heterostylous families, and occurs in addition to other characteristic features that were recognised, such as having a so-called depth-probed flower with a floral tube, as described below (section 1.7). The opposite arrangement with anthers above the stigma (reverse herkogamy) is much less common, whilst approach herkogamy occurs relatively frequently in the angiosperms as compared to heterostyly (Webb and Lloyd, 1986). The model supports the evolution of reciprocal herkogamy from this approach herkogamous ancestor, which in turn preceded the evolution of self-incompatibility. The first step is postulated to be a stigma-height dimorphism as in the Charlesworth and Charlesworth (1979b) model, which produces two herkogamous forms in one step from a herkogamous ancestor, as opposed to two steps from a homostylous ancestor. This may be one reason why SI was proposed as the first step in the Charlesworths' model that assumes a homostylous ancestor.

In the Lloyd and Webb (1992b) proposal the initial morphologies of the two herkogamous morphs comprise stigma and anthers at different heights, but they are not in exactly complementary positions. The quantitative model suggests that the stigma-height dimorphism could evolve due to the selective advantage of legitimate- over illegitimate-pollinations, but not from a reduction in self-fertilization, suggesting Darwin's cross-pollination hypothesis provides a powerful selective force. The foundation of a stigma-height dimorphism was proposed to allow further physiological

and morphological characteristics to evolve more readily, SI may subsequently evolve depending on the importance of selecting against the genetic costs of self-fertilization (Lloyd and Webb, 1992b). The importance of SI in heterostylous groups appears to vary considerably, with some species being self-compatible, others showing differing strengths in the incompatibility reaction between morphs, and rigid SI in all morphs (Barrett and Cruzan, 1994, Barrett and Shore, 2008). Instead of a homostylous precursor, the presence of a reciprocal herkogamy arrangement that promotes cross-pollination is shown to make the selection for self-incompatibility much more likely because it reduces the inherent problem of a di-allelic system in that each plant is unable to fertilize half the population (Lloyd and Webb, 1992b). In this model, additional dimorphic features such as an anther-height polymorphism or ancillary traits can be selected for if they decrease self-fertilization or increase legitimate pollination (Lloyd and Webb, 1992b).

The Lloyd and Webb model supports Darwin's proposal that reciprocal herkogamy preceded SI, but it is clear that SI is not an incidental by-product; the two systems are complementary (section 1.3). Lloyd and Webb (1992a) extended the taxonomic assessment of heterostylous groups to show that the situation in Plumbaginaceae is atypical; this family represents one instance of di-allelic SI in the absence of heterostyly, as opposed to numerous heterostylous species that are self-compatible in an array of other angiosperm families (Ganders, 1979, Lloyd and Webb, 1992a). Furthermore, stigma-height polymorphisms are frequent in non-heterostylous groups, but there are no such occurrences for di-allelic SI, suggesting this was not the ancestral state prior to heterostyly and further acknowledging the supporting role of heterostyly in the evolution of di-allelic SI (Ganders, 1979, Lloyd and Webb, 1992a).

Despite the mathematical plausibility of the selective events above, theoretical models rely heavily on their assumptions (Ganders, 1979); changing the assumptions could dramatically affect the outcomes. For example, Lloyd and Webb (1992b) point out that selection against self-fertilization is likely to have played a more important role than assigned to it if the ancestors were not herkogamous. In order to distinguish between the theoretical evolutionary models, refinement of the character states associated with heterostylous families and an in-depth analysis of their sister clades is required (Lloyd and Webb, 1992b). This, alongside data-driven analysis of evolutionary events at the *S*

locus, will determine the order of the ancestral morphological and physiological states in the evolution of heterostyly and homostyly: only once the complete *S* locus sequence and heterostyly-determining genes are uncovered will this be possible.

1.7 Phylogenetic context

Floral heteromorphy is present in over 28 angiosperm families, with the most well studied distylous genuses including *Fagopyrum*, *Primula*, *Turnera* and *Linum* (Figure 1.3) (Cohen, 2010). The characteristic features of heterostylous families reveal that reciprocal herkogamy is more likely to evolve in species with actinomorphic flowers and a floral tube, typically with pollen presented peripherally and stigmas centrally, such that pollinators contact the reproductive organs in succession when feeding on nectar at the base of the flower (depth-probed). In contrast, species with numerous free carpels or stamens, exposed nectar, irregular corollas, or open dish-shaped flowers are in most cases thought to lack the precision required for the precise pollinator contacts (Darwin, 1877, Ganders, 1979, Lloyd and Webb, 1992a, Barrett and Shore, 2008). The taxonomic distribution of heterostyly is therefore strongly non-random (Lloyd and Webb, 1992a).

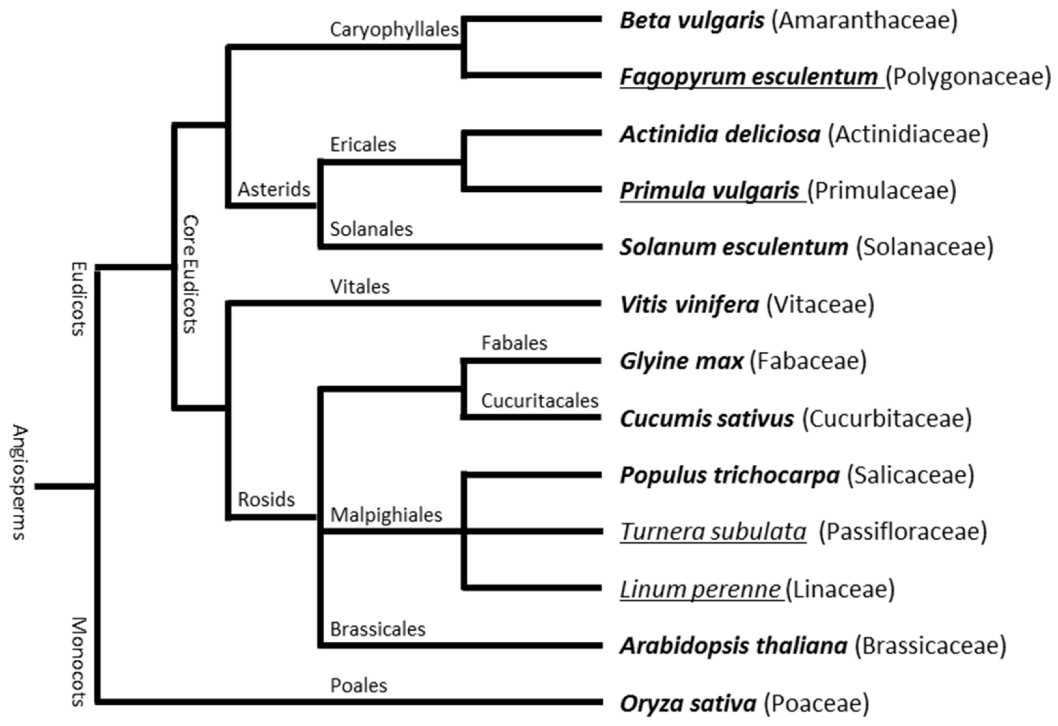


Figure 1.3 Phylogeny of *Primula vulgaris* and other angiosperm species; species with sequenced and assembled genomes are shown in bold, heterostylous species are underlined. Figure reproduced from Cocker et al. (2017).

Primula has probably attracted the most attention as a model of heteromorphic flower development. However, perhaps due to the agronomic importance of *Fagopyrum esculentum* (buckwheat), much research has also gone into the identification of the genes controlling the development of morph-specific features in this so-called pseudocereal (Yasui et al., 2012). Due to the meticulous work carried out in *Primula* towards determining the number of *S* locus genes and their order, other heterostylous taxa are also discussed in terms of control by a supergene (Barrett and Shore, 2008, Cohen, 2010). The evidence for supergene control of heterostyly in families other than the Primulaceae is limited due to a lack of unambiguous recombinants. In agronomically important *Fagopyrum* species this is perhaps due to the difficulties in carrying out association studies in crop plants due to their complex breeding histories (Flint-Garcia et al., 2003, Pasam et al., 2012, Yasui et al., 2012). *Fagopyrum* also represent an exception to the character states identified above, with species having open dish- or bowl-shaped flowers (Björkman et al., 1995).

In around ten heterostylous species the inheritance pattern of distyly is the same as in *Primula*; in some cases the dominance relationship is reversed such that the pin allele of the *S* locus is dominant to thrum (Lewis and Jones, 1992, Ornduff, 1992, Barrett and Shore, 2008). However, this does not imply that there is a supergene comprising multiple tightly-linked genes, and it remains a possibility that the "supergene" in these species comprises a single gene as in the control of butterfly mimicry for example (Barrett and Shore, 2008, Kunte et al., 2014) (see Chapter 4 discussion). In crosses between distylous and long homostyle *Fagopyrum* and *Turnera* species, thrum dominance is maintained, thus revealing allelic relationships equivalent to those at the *Primula S* locus (Barrett and Shore, 1987, Lewis and Jones, 1992, Fesenko et al., 2006, Yasui et al., 2012). The observed long homostyle species could therefore be the result of recombination at the *S* locus in thrum (*GPA*/*gpa*) to give a *gPA* haplotype as in *Primula*, but this interpretation only supports a supergene comprising two genes due to a lack of recombinants with dissociation between pollen size and anther height (Barrett and Shore, 2008).

1.8 Potential applications for the study of heterostyly

Primula are important horticultural crops in Europe, the United States and Japan, with a 100 million euro market in Germany alone (Hayta et al., 2016). These plants are produced mainly for autumn and early spring bedding and pot plant markets; as a cool-season crop they offer good return for low inputs and high plant densities, in growing temperatures that are unsuitable for most ornamental crops (Karlsson, 2001, Hayta et al., 2016). Breeding programs focus on flower size and colour, improved germination rate, and flowering time (Karlsson, 2001). From a genomics perspective, the sequencing and assembly of the *Primula vulgaris* genome could ultimately facilitate genome-assisted breeding (Varshney et al., 2005, Varshney et al., 2013, Kole et al., 2015), whilst an understanding of the genes controlling heterostyly could allow manipulation of floral architectures to further aid breeding programs.

The identification of the complete *S* locus will reveal the genes controlling both style and anther height. Perhaps these genes could allow precise floral engineering in a number of agriculturally important crop plants: self-pollinators could be engineered into out-breeders through the separation of anthers and stigma to aid hybrid seed set; or

anthers and stigma could be brought closer together in order to facilitate self-pollination. In addition, the breakdown to homostyly in heterostylous clades offers the opportunity to discover how and why self-fertilization evolves; an important consideration in light of declining populations of insect pollinators (Klein et al., 2007, Potts et al., 2010). It is thought that the transition of heterostylous species from an outcrossing system to inbreeding is a result of absent or unreliable pollen vectors such that self-fertilization is required to ensure seed set (Baker, 1966, Ganders, 1979). This decline is of particular concern for crops that rely heavily on insect pollinators to set seed or maximise fruit set (Richards, 2001). It is estimated that 20-50% of extant plant species are autogamous, including many crop plants. However, the loss of self-incompatibility is not sufficient to guarantee self-fertilization as changes in floral morphology are often required to facilitate self-pollination (Chen et al., 2007). In many crops, spatial separation of the anthers and stigma can mean that autogamy and fruit set is uncertain in the absence of animal visitation (Richards, 2001). The quality of fruit can also depend on the size or number of seeds; seed set is often reduced without insect-mediated cross-pollination due to inbreeding depression (Charlesworth and Charlesworth, 1987, Richards, 2001). The group of plants for which this is a concern includes species of agricultural importance such as oilseed rape (*Brassica napus*), flax and linseed (*Linum usitatissimum)*, sunflowers (*Helianthus annuus*), cotton (*Gossypium* spp.), soya (*Glycine max*), strawberry (*Fragaria x ananassa*), aubergine (*Solanum melanocarpum*), pepper (*Capsicum annuum*), tomato (*Lycopersicum esculentum*), olives (*Olea europea*) and grapes (*Vitis vinifera*) (Richards, 2001). In the case of poor pollination by insect pollinators, costly techniques to enforce fruit set may have to be applied such as the introduction of bees, and in extreme cases plants may require hand pollination, which is particularly expensive (Westerkamp and Gottsberger, 2000, Richards, 2001). The manipulation of floral architectures could bring anthers and stigma closer together to aid pollination.

Linum grandifolium presents flowers with a stigma-height dimorphism but no difference in anther positioning (Darwin, 1863, Barrett, 2010, Ushijima et al., 2012). The closely related *Linum usitatissimum* is important for flaxseed oil production (Wang et al., 2012). This species is predominantly self-pollinating and has hypogynous rather than heterostylous flowers (Jhala et al., 2011), but perhaps its cultivation might benefit from a greater understanding of the breeding systems in this genus. Interestingly, the

outbreeding wild relatives of the autogamous cultivated tomato *Solanum lycopersicum* are approach herkogamous, whilst the cultivated tomato itself has recessed stigmas; flowers in which the stigma is recessed relative to the anthers are more likely to self-pollinate (Chen et al., 2007).

Finally, the identification of the *S* locus and its constituent genes in *Primula* would find a more direct application in an agronomic setting by providing a platform for directed studies into isolation of the heterostyly-determining genes in *Fagopyrum esculentum*. Buckwheat species have a short vegetation period (~60 days) and are important as low input crops due to higher yield returns in sandy, stony, and nitrogen poor soils, as well as being naturally gluten-free and having an excellent nutrient profile (Kreft and Luthar, 1990, Alvarez-Jubete et al., 2010). It has been estimated that heterostyly evolved independently on 23 occasions across diverse angiosperm families (Lloyd and Webb, 1992a). Despite the potential for disparity in the underlying genes, the isolation of the *S* locus in any species would represent a significant leap forward.

1.9 The history of heterostyly

The primrose has fascinated botanists for over four centuries. Perhaps the first attempt to describe the two forms of flower was documented in the late 16th Century by Clusius (Clusius, 1583). Clusius thought the distinct floral morphs were separate varieties (or species) and is known as a botanist who paid a great deal of attention to the intricate details of floral morphologies (van Dijk, 1943, Gilmartin, 2015). It is perhaps surprising then, that his illustrations of the primrose do not show a dissected thrum flower, suggesting he was seemingly unaware of the anthers hidden in the depths of the flower (van Dijk, 1943, Gilmartin, 2015). Two centuries later, the primrose featured centrally in studies of species definition spearheaded by Carl Linnaeus, who cited the plant as an example of where flower enthusiasts concentrate on small details that "no sane botanist" would consider important, such as those between the two forms of heterostylous flower (Linnaeus, 1792, Gilmartin, 2015). The result was that the floral morphs were thereafter recognised as being of the same species; a train of thought which carved the path for Darwin, who sought to answer the inevitable question of why those differences existed at all.

Darwin noted the "balancement of long and short pistils and stamens" in 1860 and wrote to his close-friend and confidant Hooker to explain his observations (Darwin, 1860, Gilmartin, 2015). He subsequently read a paper before the Linnaean Society on the dimorphic condition and remarkable sexual relations of *Primula* species, in which he contested the general consensus amongst botanists that heterostyly represented "mere variability" (Darwin, 1862, Gilmartin, 2015). In Darwin's landmark book, *The Different Forms of Flowers on Plants of the Same Species* (Darwin, 1877), the true significance of the reproductive organs' reciprocal positions was recognised. Darwin reported that within-morph (or "illegitimate") crosses yielded a much-reduced seed set in comparison to "legitimate" crosses between morphs, and likened heterostyly to the male and female sexes in humans (Darwin, 1877).

Darwin's first idea was that *Primula* were tending towards a dioecious condition, the long styled plants with longer pistils, rougher stigmas and smaller pollen were more feminine and thus produced more seed, and the short styled plants with shorter pistils, longer stamens and larger pollen grains were more masculine (Darwin, 1877). On rare occasions the transition from heterostyly to dioecy is thought to have occurred (Ganders, 1979). Darwin concluded, however, that the morphological differences between pin and thrum were in place as part of a system to promote insect-mediated outcrossing: "the benefit which heterostyled dimorphic plants derive from the existence of the two forms is sufficiently obvious, namely, the intercrossing of distinct plants being thus ensured…"

Darwin somewhat remarkably foresaw that the benefit of such a system was in reducing the propensity for inbreeding. This, despite the fact that he presumably knew nothing of Mendel's studies at the time, and thus had no insight into the genetic basis of inbreeding depression that results from recessive deleterious traits being exposed (Charlesworth and Charlesworth, 1987). In fact, he demonstrated the negative effects of self-fertilization through his writings in an entire volume on the subject, as highlighted at the start of this introduction (Darwin, 1876).

It may be true that Darwin was aware of Henslow's earlier drawings of the two forms of primrose flower in 1826 (Kohn et al., 2005), and still before that Curtis depicted whole and dissected forms of the two floral morphs in *Flora Londinensis*, using the descriptive terms pin-eyed and thrum-eyed 100 years before Darwin (Curtis, 1777-1798); but it was

Darwin who first sought to explain the true significance of the two morphs in promoting cross-pollination where others had previously only observe (Gilmartin, 2015).

The rediscovery of Gregor Mendel's work in the early 20th Century sparked a chain of genetic studies into the genetic basis of heterostyly (Bateson, 1902, Bateson and Gregory, 1905). Bateson and Gregory (1905) were the first to define floral heteromorphy in *Primula* as a classic example of Mendelian inheritance, defining the dominance relationship of thrum and pin at the *S* locus. This was just months before Bateson coined the term "genetics" (Ornduff, 1992).

Despite little attention being paid to the ancillary features of heterostyly in determining its evolutionary origins (section 1.6), perhaps the main fascination of this system is the Mendelian segregation of a whole suite of phenotypic characters akin to the separation of the sexes (Ornduff, 1992). Pellow (1928), in an annual meeting of the John Innes Horticultural Institute, proposed that the morphological characters of the two heteromorphs might be controlled by more than one tightly-linked gene. Haldane (1933) expanded this to a series of closely linked genes. Darlington and Mather (1949) subsequently coined the term supergene to describe such a region, as well as suggesting that the *S* locus could lie on the largest pericentric chromosome based on an excessive number of *S*-linked phenotypic traits (Darlington, 1931).

The extensive inheritance studies of Alfred Ernst ultimately led to a proposal of three genes at the *S* locus, thus establishing the genetic architecture of distyly in *Primula* (Ernst, 1928a, Ernst, 1955). Ernst (1928a) described anomalous combinations of phenotypic characters and deduced that these abnormal plants were the result of mutation. It has been concluded, however, that the rate of appearance for these novel floral phenotypes is in keeping with recombination between the three genes thought to lie at the *S* locus rather than mutation, hence the prediction that homostyles are the result of recombination as described at the start of this introduction (Lewis and Jones, 1992, Barrett and Shore, 2008). Darwin described homostyles as early as 1877 (Darwin, 1877). There are natural populations in the UK near Sparkford, Somerset, and in the Chilterns, that captured the attention of population geneticists (Crosby, 1940, Fisher, 1949, Dowrick, 1956, Bodmer, 1960, Piper et al., 1984). In one case, Crosby (1949) counted 15,555 plants from heterostyled populations of *Primula vulgaris*, all of which were pin or thrum, suggesting the occurrence of homostyles is very rare, perhaps due to

recombination suppression at the *S* locus (Mather and De Winton, 1941, Barrett and Shore, 2008).

It is interesting that throughout Ernst's studies of *Primula* species (Ernst, 1955, Ernst, 1957), the vast majority of "secondary homostyle" species are long homostyles; although some short homostyle species are known, long homostyles have been reported much more frequently (Charlesworth and Charlesworth, 1979a, Ganders, 1979). This is very surprising, as crossing-over should produce long and short homostyles in equal frequency (Ganders, 1979). However, if homostyles are almost completely autogamous, then they can act as male parents (pollen donors) but are unlikely to be female parents (Charlesworth and Charlesworth, 1979a, Ganders, 1979). In this case, it has been suggested that the establishment of an inbred homozygous population of long homostyles would be favoured over short homostyles as in outcrossed matings the short homostyle functions as recessive and the long homostyle functions as dominant. This is based on the following: pollen from the long homostyle (*gPA*/*gPA*) is compatible with the pin (*gpa*/*gpa*) stigma exclusively, with crosses resulting in a *gPA*/*gpa* (long homostyle) genotype. In contrast, short homostyle pollen (*GPa*/*GPa*) is only compatible with the thrum (*GPA*/*gpa*), suggesting they are less likely to spread and become established in a heterostylous population (Dowrick, 1956, Charlesworth and Charlesworth, 1979a, Ganders, 1979). The dominance relationships at the *S* locus determine the relative frequency of short and long homostyles. This conclusion was supported by the quantitative model of Charlesworth and Charlesworth (1979a) which predicts that a self-fertile homostyle with the dominant allele for pollen type could spread into and eventually replace a heterostylous population. This view is apparently supported due to long homostyles being more frequent in the *Primulaceae*, and short homostyles more frequent in heterostylous families where the dominance relationship of pin and thrum is reversed. Indeed, the short homostyle phenotype is noted in the literatue as having been fixed only once in *Primula* (*P. septemloba* var. *minor*) (Richards, 2014).

This broad and fundamental body of work highlights the primrose as an early model for genetics; a cornerstone in the establishment of the Darwinian evolutionary synthesis, featuring prominently in the development of Mendelian and population genetics, including research focusing on the generation of linkage maps, patterns of inheritance,

epistasis, supergenes and polymorphic equilibria by the likes of W. Bateson, R.A. Fisher, J.B.S. Haldane, C.D. Darlington, C.B. Bridges, A. Ernst, K. Mather, and D. Lewis (Darlington and Mather, 1949, Barrett and Shore, 2008). The breadth and detail of the above work, including meticulous inheritance studies in *Primula* over the course of 30 years by A. Ernst (Ernst, 1928a, Ernst, 1957, Lewis and Jones, 1992), suggests that heterostyly is one of the most well studied plant sexual polymorphisms (Barrett and Shore, 2008, McCubbin, 2008). Despite this, 150 years since Darwin first explained the importance of heterostyly in the promotion of cross-pollination, the molecular basis of the system is still to be uncovered (Darwin, 1862).

1.10 Molecular studies of floral heteromorphy

Efforts towards isolating the genes at the *S* locus have focused on examining differential expression between morphs, and the generation of linkage maps and associated BAC assemblies. The most well studied species in this regard are *Primula vulgaris* (Primulaceae), *Turnera subulata* (Turneraceae), *Fagopyrum esculentum* (Polygonaceae) and *Linum grandiflorum* (Linaceae) (Cohen, 2010).

In *Primula* recent investigations include the analysis of differentially expressed floral genes using subtractive hybridisation, and the characterisation of *S* locus-linked sequences and mutant phenotypes, with a view towards generating genetic and physical maps spanning the *S* locus (Manfield et al., 2005, McCubbin et al., 2006, Li et al., 2007, Li et al., 2008, Li et al., 2010, Li et al., 2011b, Yoshida et al., 2011). In *Linum grandiflorum* (heterostylous flax) analysis of morph-specific cDNA fragments and proteins revealed morph-related genes including a potential downstream target of the *S* locus that reduces style and anther height when overexpressed in *Arabidopsis thaliana* (Ushijima et al., 2012).

In *Turnera subulata* significant progress towards isolating the key genes was made using a high-resolution genetic map (Labonne et al., 2008). This facilitated the assembly of three BAC contigs that enabled the positional cloning of the recessive *s* haplotype through the use of X-ray deletion mutants that delimit the *S* locus region (Labonne et al., 2009, Labonne et al., 2010, Labonne and Shore, 2011). The application of similar approaches in *Fagopyrum esculentum* revealed *S*-linked markers and enabled

construction of a genetic map and associated BAC assembly (Yasui et al., 2004, Konishi et al., 2006, Ota et al., 2006, Yasui et al., 2008). In addition, several proteins specifically expressed in long or short styles were identified (Miljus-Dukic et al., 2004). The analysis of a short-style chromosome deletion mutant that produces long-styled flowers identified a candidate gene named *S-ELF3* that was lost in the deletion (Yasui et al., 2012).

The classical genetic and molecular-based studies above provide a significant resource going forward that will facilitate the identification and characterisation of the heterostyly-determining genes, whilst the identification of differentially expressed genes has potentially revealed downstream targets of the *S* locus.

1.11 Next-generation sequencing in the study of heterostyly

It is perhaps surprising given the impressive number of focused molecular studies described above, that prior to the analyses described in this thesis, the genes responsible for the development of heterostyly had not been uncovered in any species; such efforts suggest that the identification of the so-called "supergene" represents a significant challenge. The application of next-generation sequencing (NGS) and associated technologies has revolutionised biological research and enabled complex tasks such as the assembly of the ~17 Gb hexaploid bread wheat (*Triticum aestivum*) genome and population level human genome sequencing through the 1000 Genomes Project (van Dijk et al., 2014, Goodwin et al., 2016). The technologies used in such studies represent a relatively untapped potential in the identification of the heterostyly determining genes.

Following the completion of the Human Genome Project (HGP) in 2003 (Collins et al., 2003), high-throughput sequencing has led to a 50,000-fold reduction in the cost of human genome sequencing, with arguably the greatest issues going forward leaning towards ethical and data storage considerations rather than sequencing costs (Goodwin et al., 2016). This makes next-generation sequencing a viable option for fundamental studies in non-model plants (Unamba et al., 2015). The HGP utilised Sanger sequencing; although this approach is costly and low-throughput it produces relatively long sequencing reads (up to 1000 bp) that allow genuine overlaps to be distinguished from repetitive sequences present in multiple copies throughout the genome. NGS

approaches on the other hand produce characteristically short reads; even accurately sequenced repetitive reads will be present in multiple copies throughout the genome (Goodwin et al., 2016).

Despite the above drawback, NGS platforms developed by Illumina currently dominate the sequencing industry, accounting for the largest market share and boasting a comparatively high-throughput and low per-base cost in comparison to alternative systems (van Dijk et al., 2014, Goodwin et al., 2016). In the cyclic reversible terminator (CRT) technology used in such platforms both ends of random DNA fragments are ligated with adapter oligonucleotides that bind to an optically transparent chip or "flow cell". The high density of primers on the flow cell results in a massively parallelised approach, where each fragment forms a bridge with an adjacent primer to allow PCR amplification, and the build-up of dense clusters of identical fragments that permit sequencing. CRT is a so-called "sequencing by synthesis" approach where four fluorescently labelled deoxynucleoside triphosphates (dNTPs) are washed over the flow cell with DNA polymerase and incorporated into a growing chain; after laser excitation, the emitted fluorescence is imaged and the fluorophore that doubles as a reversible terminator is cleaved to allow the sequencing cycle to continue with incorporation and identification of the next base (Schatz et al., 2010, Koren et al., 2012). The maximum paired-end read length of Illumina HiSeq 2500 is 250 bp with the lower throughput Rapid Run Mode, whilst the high-throughput mode only produces reads up to 150 bp (Rhoads and Au, 2015).

In contrast to Illumina CRT, some third-generation sequencing technologies do not require cyclical hybridisation and wash steps as they aim to directly detect the composition of single-stranded DNA molecules without the need for costly sequencing reagents or base-incorporation steps (Goodwin et al., 2016). These platforms are based on the transit of the DNA molecule through a pore that allows sequencing based on an electric current or optical signal that is characteristic of a particular DNA sequence (Clarke et al., 2009). The single flowcell MinION from Oxford Nanopore Technologies (ONT) is the first commercially available platform for such a technology and produces considerably longer reads than Illumina CRT (up to 200 kb), but so far it is comparatively expensive and has much lower throughput. In lieu of the above, long mate-pair (LMP) libraries with large insert sizes can be used to bridge gaps in an

assembly by traversing repetitive regions. The sequencing of LMP libraries is carried out in the same way as for paired-end reads with short insert sizes, but preparation of a library comprising long fragments of DNA can prove problematic (Koren et al., 2012). This therefore suggests that the production of long reads by the so-called third-generation sequencing methods is a technology worth pursuing.

Future proposals include the massively parallelised ONT PromethION platform that is reported to comprise 48 improved flow cells that could produce ~2-4 Tb of sequencing data in a two day run. These figures would put this system in potential competition with the ultra high-throughput Illumina HiSeq X, currently the highest throughput platform available (Schadt et al., 2010, Goodwin et al., 2016). In response, Illumina's synthetic long-read sequencing platform combines existing sequencing technologies with barcoded wells that contain long fragments (~8-10 kb) that are amplified and sheared to ~350 bp for pooled short read sequencing that retains a record of reads associated with each long fragment (Goodwin et al., 2016). Pacific Biosciences (PacBio) currently offer the most widely used (third-generation) platform for the sequencing of single molecules in real time, based on fluorescently labelled probes; unlike the ONT system this approach relies on sequencing reagents and optical equipment for base detection, but it does not require a pause between each base-read as it leverages the split-second delay during base-incorporation (Schadt et al., 2010, Rhoads and Au, 2015, Goodwin et al., 2016). This platform therefore allows faster runs than CRT approaches, and it is continually improving in terms of throughput and read lengths, but the per-base-pair costs and high single-pass error rates (11-15%) means data is often augmented with short reads from Illumina CRT that have an overall accuracy rate of > 99.5% (Nagarajan and Pop, 2013, Goodwin et al., 2016, Mostovoy et al., 2016). The PacBio RS II system boasts average read lengths of over 10 kb in single-pass runs, but multiple (15) passes of the circular DNA molecule are required to improve accuracy to > 99%, therefore prompting a trade-off between read length and accuracy based on limits stipulated by the lifetime of the DNA polymerase; longer sequences have lower accuracy (Rhoads and Au, 2015). NGS technologies are increasingly becoming a routine part of biological research, but the developing technologies described above are yet to achieve the necessary throughput or accuracy to compete with the high-throughput platforms from Illumina that are currently used in most genomics studies (Nagarajan and Pop, 2013, Goodwin et al., 2016, Jackman et al., 2016).

For the assembly of short sequencing reads associated with cost effective next-generation sequencing platforms, a number of application-specific algorithms were developed. These approaches are based on the de Bruijn graph data structure, first used in the EULER assembler (Schatz et al., 2010). Some of the most popular tools are Velvet, SOAPdenovo, ABySS and ALLPATHS (Butler et al., 2008, Zerbino and Birney, 2008, Simpson et al., 2009, Luo et al., 2012, Nagarajan and Pop, 2013). The first assemblies of long Sanger sequencing reads were accomplished using algorithms that find overlaps between reads allowing for sequence differences (1-10%), then merging sequences with the longest overlap to form contiguous sequences or "contigs" (Schatz et al., 2010) These simple "greedy" algorithms where all reads are compared with each other are not tractable for the millions of short reads generated for the assembly of large genomes (Schatz et al., 2010).

In de Bruijn graph based tools, reads are divided into sequences of *k* length (*k*-mers) that form the nodes of the graph, with *k*-mers that overlap by *k*-1 bases connected by edges such that a path through the graph can be traversed to form contigs that end on the boundaries of repeats (Schatz et al., 2010). The advantage of this approach is that *k*-mer overlaps representing perfect complementarity between reads are implicitly captured by the graph, rather than explicitly computed, saving a substantial amount of computing time (Simpson et al., 2009, Jackman et al., 2016). The method reveals sequencing errors as "tips" (dead-ends) or "bubbles" in the graph, with "forks" further highlighting repetitive or duplicated regions that may be simplified into one sequence or removed. In dividing reads into a set of *k*-mers pre-processing and error correction is possible, which results in an easier assembly process; a contig with a large number of associated reads can also be flagged as repetitive based on an erroneously high coverage (Chaisson and Pevzner, 2008, Zerbino and Birney, 2008, Schatz et al., 2010, Nagarajan and Pop, 2013).

Following contig assembly, information gleaned from paired-end reads is incorporated; these reads are generated from sequencing the DNA fragments from both ends, with reads separated by the average fragment (insert) size of the library. If a unique path in the assembly graph is found to connect reads at the end of a read-pair with length equal to the fragment size, then the path is thought to be correct; heuristic measures help to facilitate the intensive task of finding paths of pre-defined length (Jackman et al., 2016).

This approach enables contigs to be linked together across repetitive regions less than the insert size to generate "scaffolds" through the merging of linked contigs (Schatz et al., 2010, Jackman et al., 2016).

Recent advancements in genome assembly have focused on improving memory efficiency through new approaches to construct and process assembly graphs, thus facilitating the analysis of larger datasets without the need for high-performance computational infrastructure (Simpson et al., 2009, Nagarajan and Pop, 2013, Goodwin et al., 2016, Jackman et al., 2016). In addition, the incorporation of complementary information derived from multiple sequencing technologies and/or mate-pair data is increasingly being leveraged (Nagarajan and Pop, 2013, Jackman et al., 2016). The central challenge in genome assembly is resolving repetitive regions (Schatz et al., 2010). The use of LMP (paired-end read) libraries generated from libraries with larger fragment sizes enables larger gaps in short paired-end read assemblies to be bridged with "N"s representing the insert size between reads, but as a result of the difficulties in getting DNA to circularize efficiently as fragments get longer, these LMP libraries often comprise too little DNA to cover the genome at a depth comparable to that of paired-end read libraries generated from short fragments, and there is also a higher associated variance in their length distribution (Collins and Weissman, 1984, Schatz et al., 2010, Jackman et al., 2016).

If further linkage information in the form of a physical map is available, this also permits ordering and orientation of contigs; one widely applied approach to this end is optical mapping, which involves stretching linear DNA fragments across a glass surface or in a "nanochannel" array such that locations of restriction digest sites can be visualised and fragment lengths estimated with the aid of dye or fluorescent labelling (Schwartz et al., 1993, Dong et al., 2013, Tang et al., 2015, Jackman et al., 2016). The cost of generating a physical map is often prohibitive, but use of NGS generated paired-end reads in combination with LMP reads is now attainable for a non-model species such as *Primula vulgaris*, using libraries with a mixture of insert sizes can be very effective (Schatz et al., 2010, Jackman et al., 2016).

The most direct application of the above technologies in this project is for the whole genome sequencing (WGS) and assembly of *Primula vulgaris*. WGS offers the most comprehensive view of genomic information (Goodwin et al., 2016), whilst

transcriptomics applications through RNA sequencing (RNA-Seq), and sequence capture approaches, provide an alternative that requires much less throughput per sample for applications where it is neither practical nor necessary to sequence an entire genome (ten Bosch and Grody, 2008, Goodwin et al., 2016). For example, whole-exome sequencing is designed to selectively enrich for coding regions that make up only 2% of the human genome, but contain 85% of known disease-causing variants (Goodwin et al., 2016). The capture of a specific set of sequences is often achieved with in-solution capture methods (e.g. MYbait oligo-capture), where a biotinylated probe is designed for hybridization to an a priori genomic target, followed by fragment retrieval with streptavidin-coated magnetic beads, and a wash step to discard fragments that were not captured. PCR-amplification and sequencing of the captured fragments can then be carried out (Ávila-Arcos et al., 2015, Warr et al., 2015).

RNA-Seq reads are derived from the high-throughput sequencing of fragments prepared from total mRNA content in cells, comprising a mixture of different spliceoforms and transcripts at various levels of abundance that represent the genes that are transcribed at the time of sampling. The resulting reads can be used for the prediction of gene structures, identification of splice variants, and due to the correlation between the number of RNA-Seq reads and the abundance of the transcript that the reads were derived from, whole-genome expression profiling (Trapnell et al., 2009, Trapnell et al., 2012). RNA-Seq assembly approaches can be *de novo* (no reference assembly required) as in Trinity, trans-ABySS, Oases, and SOAPdenovo-Trans (Robertson et al., 2010, Grabherr et al., 2011, Schulz et al., 2012, Xie et al., 2014), using a similar *k*- mer based approach to the graph-based WGS assemblers above; or referenced based, in which case a high-quality reference genome assembly is desirable (Grabherr et al., 2011, Chang et al., 2015).

NGS tools produce hundreds of millions of short-sequencing reads; these reads may contain sequencing errors, or genuine mismatches, insertions and deletions as a result of genomic variation (Trapnell et al., 2012, Dobin et al., 2013). Phred quality scores provide a measure of read quality in the FASTQ sequence files produced using NGS technologies, and are calculated using several parameters related to peak shape and peak resolution of DNA sequencing traces at each base following NGS. These parameters are used to detect corresponding quality scores in a large lookup table. The lookup table

contains information on how often the correct base was called for a particular set of parameter values where the correct base was known. This was carried out for a test set comprising tens of thousands of known sequences (Ewing et al., 1998). The probability value obtained from the table is converted into a Phred score, with a higher score indicating a smaller probability of error (Cock et al., 2010).

In DNA resequencing efforts, the aim is to align NGS reads to a previously assembled reference genome, often in order to capture information on Single Nucleotide Polymorphisms (SNPs); the mapped reads are assigned Phred-scaled mapping quality values. The most popular alignment tools for this purpose are Bowtie, BWA and SOAP (Li et al., 2009b, Langmead and Salzberg, 2012, Li, 2013). These tools use the Burrows Wheeler Transform (BWT) which allows sequences to be more easily compressed such that repeats in an indexed genome can be collapsed (Li and Durbin, 2009). This means reads can be aligned against each copy of a repeat simultaneously instead of against each one individually, which makes these algorithms more efficient; sequences can be retrieved as the BWT works in such a way that the compression can be reversed (Nong and Zhang, 2007, Ruffalo et al., 2011). There are continual advancements to improve the efficiency of such tools (Liu et al., 2012, Xie et al., 2014). There is often a compromise between mapping accuracy and computational resource requirements, with the speed of read alignment increasingly representing a serious bottleneck in throughput (Dobin et al., 2013). Following short-read alignment, SNPs can be called using tools such as SAMtools and GATK if there is a sufficient number of high-quality reads covering the SNP position (Li et al., 2009a, DePristo et al., 2011).

The above tools are not sufficient for the mapping of RNA-Seq reads as transcriptome information is reorganised into non-contiguous exons that must be spliced together to produce mature transcripts (Dobin et al., 2013). For this reason “splice-aware” alignment methods have been developed to discover splice sites, including Tophat, GSNAP, STAR and HISAT (Wu and Nacu, 2010, Trapnell et al., 2012, Dobin et al., 2013, Kim et al., 2015). The problem of aligning sequences with mismatches to the genome is shared with DNA resequencing, but this must be extended to consider splicing in RNA-Seq alignment tools in order to fully reconstruct intron and exon structures based on spliced RNA molecules. This is a particular problem where a read maps to a splice site such that a short (< 10 bp) overhang is left. The above issues are

further confounded due to families of related genes that share similar sequences, thus resulting in multiple mapping locations that are seemingly correct (Dobin et al., 2013). Following alignment of RNA-Seq reads, tools such as Cufflinks or Scripture can be used to assemble the reads into transcripts by merging sequences with overlapping alignments and reconstructing splice variants (Guttman et al., 2010, Trapnell et al., 2012). The alignment of sequencing reads is another scenario where the long sequencing reads offered by third-generation sequencing platforms would provide great advancement in capabilities; particularly in the reconstruction of haplotype information and RNA connectivity.

Transcriptome assembly approaches are often the platform for detection of differential expression between genes using tools such as CuffDiff, DESeq and edgeR (Anders and Huber, 2010, Robinson et al., 2010, Trapnell et al., 2013). This task, and indeed RNA-Seq assembly itself, is complicated by non-uniform coverage between and within transcripts, and across samples with different library sizes. For this reason normalization schemes are applied, for example transformed quantities such as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) are used to normalize the counts for differing library sizes and transcript sizes; long transcripts will have more associated reads than a short transcript with equal expression (Soneson and Delorenzi, 2013). RNA-Seq based transcriptome assemblies can be also used as evidence for intron positions and exon-noncoding region boundaries in genome annotation software such as AUGUSTUS and MAKER2, alongside other sources of support such as proteins from related-organisms and repeat annotations (Stanke and Morgenstern, 2005, Holt and Yandell, 2011, Hoff et al., 2016). RNA-Seq coverage alone is not sufficient to predict protein coding regions; algorithms incorporating statistical models for gene prediction usually require a training step where a set of high quality example genes defines species specific parameters (Hoff et al., 2016).

Following assembly and the generation of gene annotations and associated RNA-Seq datasets, functional annotation and cross-species comparative analyses are often used to facilitate evolutionary analyses and assign potential functions to genes; such studies can be based on local alignments of protein and nucleotide sequences using tools such as Exonerate, BLAT or BLAST+ (Kent, 2002, Slater and Birney, 2005, Camacho et al., 2009). These programs first scan for short matches (words), and then extend them into

larger alignments termed high-scoring pairs (HSPs). In contrast, multiple sequence alignment (MSA) tools such as MUSCLE or Clustal seek to maximize the sum of sequence similarities across the entire length of a set of sequences, so-called global alignment, with penalties for gaps (Edgar, 2004, Edgar and Batzoglou, 2006, Larkin et al., 2007). MSA is the starting point for the inference of phylogenetic distances and protein structure prediction (Edgar and Batzoglou, 2006).

For whole genome assembly, annotation, differential expression analysis, and associated comparative analyses in *Primula vulgaris*, a combination of the methods described in this section will form the first focus. In the study of heterostyly, the use of these approaches could help to bridge gaps in BAC assemblies, whilst RNA-Seq approaches will provide the precision to define differentially expressed sequences between pin and thrum, including those at the *S* locus.

1.12 Summary and research aims

Research into primroses and heterostyly has straddled major advances in science over the last century and a half. First, is an era characterised as descriptive, based on increasingly meticulous observations of floral morphologies, and the fine balance between this and Linnaean thinking (Clusius, 1583, Curtis, 1777-1798, Linnaeus, 1792, Gilmartin, 2015); second is a focus on Darwin's explanation of the detrimental effects of self-fertilization, and the adaptive value of heterostyly in promoting outcrossing and generating novel variation as a substrate for natural selection (Darwin, 1862, Darwin, 1877, Li et al., 2016). Following this was the rediscovery of Mendel's work, with a central role for the primrose in the subsequent advancement and synthesis of subjects related to genetics and the principles of Darwinian evolution (Bateson and Gregory, 1905, Bridges, 1914). Finally, the development of molecular-based technologies has accelerated progress towards isolation of the *S* locus genes (Labonne and Shore, 2011, Yasui et al., 2012, Li et al., 2015a) (section 1.10). This thesis is situated at the intersection between the molecular-era and a new phase of genomics-based research. The aim is to present an array of assemblies and associated genomic resources that have contributed towards the isolation and characterisation of the elusive *S* locus supergene, as well as an understanding of its evolutionary origin and maintenance.

2

Assembly and genomic analysis of *Primula* genomes

2.1 Relevant publications

Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Dyer, S., Caccamo, M., Gilmartin, P.M. (2017) The *Primula vulgaris* genome (in preparation).

2.2 Introduction

The study of floral heteromorphy in *Primula* dates back over 150 years. It was Charles Darwin, in his book *The Different Forms of Flowers in Plants of the Same Species*, who first explained the importance of the system in promoting outcrossing (Darwin, 1877). Darwin described how each primrose plant had one of two forms of flower; the pin or the thrum, with the anthers and stigma situated in reciprocal positions, such that insect-mediated cross-pollination between the two floral morphs is physically promoted. In the majority of Primulaceae species this functions alongside a di-allelic self-incompatibility (SI) system termed heteromorphic SI, which acts to reduce self-fertilization (Lewis, 1954, Dowrick, 1956, Lewis and Jones, 1992).

Darwin was aware of the "inferiority" of seedlings resulting from self-fertilization in terms of their reduced height and vigour (Darwin, 1876), and proposed that the primary purpose of heterostyly was in preventing these detrimental effects (Darwin, 1877). This particular adaptation of floral reproductive structures is a remarkable example of the evolutionary innovations that angiosperm species have undergone in order to promote outcrossing, often involving coevolution with insect pollinators; the resulting genetic diversity may facilitate adaptation to changing environments due to favourable

combinations of alleles, thus mitigating extinction and accelerating species diversification (Dodd et al., 1999, de Vos et al., 2014). Floral innovations are thought to have enabled this group of plants to become the most diverse on earth, with over 350,000 living species classified into 416 families (Dodd et al., 1999, Paton et al., 2008, de Vos et al., 2014, The Angiosperm Phylogeny, 2016).

The apparent adaptive value of heterostyly in the promotion of outcrossing marks it as one of the most fascinating examples of convergent evolution, having evolved independently on at least 23 occasions, with representative species in 28 families of flowering plants (Lloyd and Webb, 1992a). The *Primula* genus sits within the Primulaceae family of the order Ericales; a highly evolved sub-clade of the Asterids lineage, within which the majority (90%) of species are heterostylous (Richards and Edwards, 2003). Like the closely related heterostylous species *Primula veris* (Nowak et al., 2015), the *Primula vulgaris* genome has 11 chromosome pairs (2n =2x = 22), with estimates by flow-cytometry placing the genome size at either 459 Mb (Temsch et al., 2010) or 489 Mb (Siljak-Yakovlev et al., 2010), which is comparable to the 479.22 Mb predicted for *Primula veris* (Siljak-Yakovlev et al., 2010).

The draft genome of *P. veris* was recently published (Nowak et al., 2015), with suggestions for the potential basis of heterostyly based on *S* locus-linked sequences previously identified by Li et al. (2011b). However, it did not facilitate the assembly of the *S* locus, nor was complete linkage to the heterostylous floral morphs shown for any of the assembled sequences. The assembly covers only 301.8 Mb (63%) (Nowak et al., 2015) of the estimated 479.22 Mb genome. In addition, the gene annotation and differential expression studies are based on RNA-Seq datasets without biological replicates: samples are only from leaf and flower tissue, as opposed flowers, leaves, roots, fresh seed, and seedlings in our analysis of *P. vulgaris*. To proceed with analyses of heterostyly in *Primula*, an improved reference assembly for the Primulaceae would be a valuable resource.

The comparison of assembly size with the predicted size of the genome based on flow-cytometry approaches is a useful gauge of genome-completeness, and is one reason why the NG50 measure of genome assembly quality is preferred over the N50 (Earl et al., 2011, Bradnam et al., 2013). The N50 is an often used metric that describes the contig length above which 50% of the total genome assembly content is contained, whereas the

NG50 is the contig length above which 50% of the total predicted size of the genome is contained; hence the latter takes into account the actual size of the genome being sequenced as opposed to the assembly size, which may fall somewhat short of this (Earl et al., 2011). These measures are an improvement over comparisons using the average contig size due to the high frequency of short contigs in many genome assemblies (Parra et al., 2009). The assembly of short sequencing reads from non-haploid organisms that are highly heterozygous due to widespread polymorphism often leads to a highly fragmented assembly with a total size larger than estimated (Small et al., 2007, Pryszcz et al., 2014, Pryszcz and Gabaldón, 2016). This suggests that the *P. veris* assembly could feasibly represent less than 63% of the genome content, with much of the polymorphic content from a heterozygous read library likely to be represented by alternate contigs that artificially inflate the size of the assembly. The use of *k*-mer frequency abundance graphs can provide a visual depiction of the heterozygosity of a genome by fragmenting genomic reads into *k* sized sequences (*k*-mers) and determining the number of distinct *k*-mers appearing in the genome at each frequency. Since a heterozygous genome will result in polymorphic regions being represented by two sets of alternate sequences, *k*-mers unique to these regions will have a frequency that is roughly half that in homozygous regions. This results in a distribution of *k*-mer frequency abundances containing two peaks, one with double the average *k*-mer frequency of the other (Liu et al., 2013). Alongside the N50 and NG50 metrics, this is another method that is useful for the validation of whole-genome assemblies (Mapleson et al., 2016). Further to this, genome assembly-completeness can be assessed by mapping genes known to be present in the majority of eukaryotic species. For this purpose, sets of core eukaryotic genes (CEGs) and associated methods for ascertaining their presence in a genome assembly have been developed: highly paralogous genes are removed from the CEG dataset to reduce false positives when trying to accurately identify orthologues (Parra et al., 2007, Simão et al., 2015). This chapter will employ the above methods for validation of the *Primula* genome assemblies described herein.

In the assembly of the potato (*Solanum tuberosum*) genome, for example, considerable effort was taken to overcome the key issue of heterozygosity by using a doubled-monoploid form of potato that can be derived through tissue culture approaches (Paz and Veilleux, 1999, Xu et al., 2011). Here, our chosen reference assembly for *P. vulgaris* is generated from a long homostyle *P. vulgaris* plant derived from an in-bred

population that originates from the Wyke champflower population in Somerset, UK (Crosby, 1949). This plant was selfed for five generations in the hands of the PMG lab (Jinhong Li, personal communication), but is likely to have undergone further selfing prior to collection. In contrast to pin or thrum plants, homostyle flowers are characterised by the display of anthers and stigma at the same height (Crosby, 1940); the long homostyle presents both of these organs at the mouth of the flower. In such an individual, there is a breakdown in the reciprocal positions of the reproductive structures and associated heteromorphic SI system, such that the long homostyle plant is self-fertile. This allowed an inbred and presumably highly-homozygous line to be generated for use in this study: the homozygosity of this individual will be explored using *k*-mer frequency abundance plots (Mapleson et al., 2016). Self-pollination of homostyle flowers will produce offspring with greater allelic homozygosity than obligate out-crossing pin or thrum plants. For genome assembly, greater homozygosity reduces fragmented contig assemblies and gene models, as well as duplicate redundant contigs and gene paralogues that could result from highly polymorphic regions of the genome (Pryszcz & Gabaldón, 2016). The use of this individual will therefore avoid problems of heterozygosity in the genome: alongside comprehensive *P. vulgaris* RNA-Seq datasets from multiple tissues, this will facilitate the generation of a high-quality genome assembly, which should enable the annotation of a more complete geneset than that of the draft *P. veris* genome assembly (Nowak et al., 2015), as well as contributing to the advancement of comparative analyses within the Primulaceae.

Efforts since Darwin's work (Darwin, 1862, Darwin, 1877) have centred on defining the Mendelian inheritance of the genes responsible for heterostyly (Bateson and Gregory, 1905, Gregory et al., 1923, Dowrick, 1956), generating linkage maps around the controlling locus based on classical genetics approaches (Bridges, 1914, Altenburg, 1916, De Winton and Haldane, 1931, Ernst, 1936c), and the assembly and annotation of a BAC assembly initiated by genetic markers that co-segregate with the pin and thrum phenotypes (Li et al., 2011b, Li et al., 2015a); as described in Chapter 3 of this thesis. The combination of genetics and genomics promises to unravel the molecular basis of heterostyly in *Primula* by building on the above studies; this chapter presents the foundation of this work through multiple *Primula* genome assemblies and associated genomic analyses.

2.3 Methods

2.3.1 Genomic DNA and RNA-Seq paired-end read libraries

Plant material is from the population of *Primula* plants maintained by the Philip Mark Gilmartin (PMG) lab; for the inbred self-fertile *P. vulgaris* long homostyle, DNA was originally isolated from the wild population of long homostyles at Wyke Champflower, Somerset, UK (Crosby, 1940). DNA and RNA preparation was carried out by Jinhong Li (JL) prior to genomic DNA and RNA-Seq library preparation at The Genome Analysis Centre (TGAC) (now Earlham Institute) using standard Illumina protocols. Paired-end sequencing of read libraries (Table 2.1) was carried out with either Illumina HiSeq2000 or HiSeq2500 at TGAC, as described in Li et al. (2016). Genomic DNA was isolated from leaf tissue, and RNA from various tissues as listed in Table 2.1. The mean insert-sizes of the read libraries were determined (by TGAC) using an Agilent 2100 Bioanalyzer, which analyses wells on a chip based on the principles of gel-electrophoresis (http://www.genomics.agilent.com/).

Library	Material	Type	Mean insert size (bp)	Read count
LIB2558	Long homostyle	Genomic	522	140114901
LIB5215	Long homostyle	Genomic	4131	40450009
LIB5216	Long homostyle	Genomic	6675	54234033
LIB5217	Long homostyle	Genomic	8818	53977718
LIB3565	Short homostyle	Genomic	464	168866674
LIB1732	Pin parent	Genomic	423	205079089
LIB1167	Thrum parent	Genomic	368	147913358
LIB1474	Thrum parent	Genomic	9780	225676870
LIB1730	Pin pool	Genomic	349	197538574
LIB1731	Thrum pool	Genomic	180	204029422
LIB3564	*P. veris* thrum	Genomic	556	179871647
LIB4735	Root (pin & thrum)	RNA	199	77754060
LIB4736	Fresh seed (pin & thrum)	RNA	205	63843125
LIB4737	Seedlings (pin & thrum)	RNA	235	80545495
LIB8234	Pin flower (1)	RNA	296	39027602
LIB8235	Pin flower (2)	RNA	278	26497129
LIB8236	Pin flower (3)	RNA	318	31792278
LIB8237	Pin flower (4)	RNA	284	34838987
LIB8238	Thrum flower (1)	RNA	296	30318958
LIB8239	Thrum flower (2)	RNA	295	32575413
LIB8240	Thrum flower (3)	RNA	269	23971121
LIB8241	Thrum flower (4)	RNA	286	38550763

Table 2.1 DNA and RNA-Seq paired-end read libraries used for genome assembly and transcriptomic analyses. Numbers in brackets indicate replicates for RNA-Seq libraries; flower replicate libraries were generated from 15-20 mm *P. vulgaris* buds. Read count = number of paired-end reads in the library.

2.3.2 Genomic DNA paired-end read assemblies

Genomic paired-end read assemblies listed in Table 2.2 were carried out by Jonathan Matthew Cocker (JMC) or Jon Wright (JW) as listed. To generate the LH_v2 assembly, the long homostyle paired-end reads (LIB2558) were assembled with SOAPdenovo (v2.04) by JW (*k*-mer length (-K) 81), followed by scaffolding (*k*-mer length (-k) 41) with long mate-pair (LMP) libraries (LIB5215, LIB5216, LIB5217) and removal of contaminated contigs. All other paired-end read assemblies (Table 2.2) were generated by either JMC: SH (short homostyle), PP (pin parent), and VT (*P. veris* thrum), or JW: TP (thrum parent) and LH_v1 (long homostyle), using ABySS (v1.3.4) (Simpson et al., 2009), and scaffolded (where applicable) with SOAPdenovo (v2.04) (Luo et al., 2012), as described in Li et al. (2016). For the short homostyle (SH_v2) and pin parent (PP_v2) assemblies, *k*-mer lengths of 85 (-*k* 85) and 71 (-*k* 71) respectively, were specified for contig assembly, followed by scaffolding with the 9 kb thrum parent LMP reads (LIB5217) (prepare command options: -K 85 (SH), -K 71 (PP), map command options: -k 63 (SH), -k 71 (PP)). *P. veris* thrum (VT_v1) was assembled with ABySS (v1.3.4) (Simpson et al., 2009) using a *k*-mer length of 81 (by JMC). For all assemblies, only sequences ≥ 200 bp were retained for further analysis.

Assembly name	**Assembled by**	**Flower form**	**Libraries used for assembly**
LH_v2	JW (TGAC)	Long homostyle	LIB2558 (scaffolded with LIB5215, LIB5216, LIB5217)
LH_v1	JW (TGAC)	Long homostyle	LIB2558 (scaffolded with LIB1474)
SH_v2	JMC	Short homostyle	LIB3565 (scaffolded with LIB1474)
TP_v2	JW (TGAC)	Thrum	LIB1167 (scaffolded with LIB1474)
TP_v1	JW (TGAC)	Thrum	LIB1167 (not scaffolded)
PP_v1	JMC	Pin	LIB1732
PP_v2	JMC	Pin	LIB1732 (scaffolded with LIB1474)
VT_v1	JMC	*P. veris* Thrum	LIB3564 (not scaffolded)

Table 2.2 *Primula vulgaris* (and *P. veris*, where listed) paired-end read assemblies carried out by JMC or JW (TGAC).

2.3.3 Assembly validation

To evaluate the LH_v2 assembly for duplicated content, and to compare the content of the unprocessed paired-end reads against the final assembly, the long homostyle assembly (LH_v2) and read library (LIB2558) were used to generate *k*-mer hashes with the Jellyfish (v2.2.0) "count" function: *k*-mer length (-m) = 31 (Marçais and Kingsford, 2011). The K-mer Analysis Toolkit (KAT) (v2.1.0) (Mapleson et al., 2016) was used to produce a matrix of *k*-mers shared between the FASTA file of LH_v2 and the long homostyle read library (LIB2558) (with the "comp" function), and to produce a *k*-mer spectra plot using the "plot" function. Jellyfish (v2.2.0) was also used to produce *k*-mer hashes for the pin (LIB1732), short homostyle (LIB4568), thrum (LIB1167), and published *P. veris* (Nowak et al., 2015) read libraries; the "histo" function was then used to generate a frequency-abundance distribution of *k*-mer occurrences, which was plotted in R (v3.2.0) with x- (frequency) and y-axis (*k*-mer count) limits of 100 and $2x10^7$ respectively. Separate plots were produced in the same way for the *P. veris* reads, one on the same scale as the above for comparison, and the other with altered x- and y-axis limits of 400 and $5x10^6$, respectively, to improve visualisation of the distinct peaks in the distribution. KAT was also used to produce a plot of *k*-mers shared between the *P. veris* reads and assembly (x- and y-axis limits of 400 and $5x10^6$).

CEGMA (v2.5) (Parra et al., 2007) was used to evaluate the presence of a set of 248 core eukaryotic genes in the *Primula vulgaris* LH_v2 genome assembly, as a proxy for completeness of the genome; the genome assembly FASTA file was used as input (intron_size=50000).

2.3.4 Repeat library construction and repeat masking

Only the long homostyle (LH_v2) genome was annotated. RepeatModeler (open v1.0.7) (http://www.repeatmasker.org/RepeatModeler.html) was used to identify *de novo* repetitive sequences in the LH_v2 scaffolds.

To reduce the chance of protein coding genes being annotated as repeats, sequences from the *de novo* repeat library were aligned to Pfam-A (using curated thresholds) and Pfam-B (e-value $1x10^{-4}$) with HMMer hmmscan (v. 3.1b1) (Eddy, 2009, Finn et al., 2014). The sequences with at least one alignment to a transposition-associated domain,

or with no alignments to any Pfam domains (transposition-associated or otherwise), were retained. Pfam domains were considered transposition-associated based on alignment to a database of transposable elements included in the RepeatRunner package using HMMer hmmscan (v3.1b1) (e-value $1x10^{-4}$) (http://www.yandell-lab.org/software/repeatrunner.html). BlastX alignments to the NCBI "nr" database were carried out for sequences in the repeat library identifying a non-transposition-associated domain in addition to a transposition-associated domain; those with no BlastX (v2.2.28+) (Camacho et al., 2009) hits or that appeared to be transposon-associated based on manual review were retained, the remainder were removed from the repeat library.

Short-interspersed and simple repeats were identified in the genome assembly using the curated *de novo* repeat library with a local installation of RepeatMasker based on the RMBlast algorithm (open v4.0.1) (http://www.repeatmasker.org/). Additional classification of repeat elements was performed with TEclass (Abrusan et al., 2009).

2.3.5 Alignment of proteins from related species

FASTA files of protein sequences were obtained for the species listed in Table 2.3. Additional protein sequences from species within the Asterids were obtained from the NCBI protein database (http://www.ncbi.nlm.nih.gov/) using "asterids[orgn]" as the search term, with the aforementioned species excluded from the results. The sequences containing Selenocysteine ("U") or Pyrrolysine ("O"), or ambiguous amino acids, were removed from the dataset; low-complexity regions were masked in the remaining sequences using the "segmasker" tool (available as part of Blast 2.2.28+) (Camacho et al., 2009).

The above proteins were aligned to the LH_v2 assembly with Exonerate 2.2.0 (https://www.ebi.ac.uk/~guy/exonerate/) using the "protein2genome" model with soft-masked query and target (minintron=20, maxintron=50000) to produce a GFF format file that was converted to an AUGUSTUS hints file using the exonerate2hints Perl script (available at http://augustus.gobics.de/binaries/scripts/exonerate2hints.pl); this script was modified to exclude alignments below 90% identity and 70% coverage.

Species	Access link	Version
Solanum lycopersicum	http://solgenomics.net/organism/Solanum_lycopersicum/genome	2.4
Solanum tuberosum	http://solgenomics.net/organism/Solanum_tuberosum/genome	3.4
Mimulus guttatus	http://phytozome.jgi.doe.gov/pz/#!info?alias=Org_Mguttatus	2.0
Capsicum annuum	http://peppersequence.genomics.cn/page/species/index.jsp	2.0
Actinidia chinensis	http://bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi	NA
Nicotiana benthamiana	http://solgenomics.net/organism/Nicotiana_benthamiana/genome	0.4.4

Table 2.3 Protein databases of species closely related to *Primula vulgaris* used for alignment to the LH_v2 genome assembly.

2.3.6 Annotation of genes

The *ab initio* annotation software AUGUSTUS (Stanke and Morgenstern, 2005) was trained (by JW) with a set of full-length transcripts assembled using TopHat (Trapnell et al., 2012) (v2.0.11) and Cufflinks (Trapnell et al., 2012) (v2.1.1). These transcripts are based on alignment of RNA-Seq datasets derived from leaves, flowers, roots, fresh seed, and seedlings from both pin and thrum *P. vulgaris* plants (Table 2.1, with leaf/flower libraries from Table 3.1). This facilitated prediction (by JW) of protein-coding genes, with the repeat annotations, protein alignments from related species (JC; above), and RNA-Seq transcript models (assembled by JW) as evidence.

2.3.7 Functional annotation of predicted genes

Functional annotation of predicted genes in LH_v2 was carried out using AHRD (Automated assignment of Human Readable Descriptions) (https://github.com/groupschoof/AHRD) (Hallab, 2015). Alignments were performed using BLASTP (v2.2.26) (e-value 1x10-4) with the following target databases: Uniprot/trEMBL, Uniprot/Swissprot (http://www.uniprot.org/) and TAIR10 (https://www.arabidopsis.org/). AHRD aims to algorithmically select the best scoring descriptions from Blast alignments, with greater weight being given to descriptions from more trusted sources (weighting applied for this analysis; TAIR = 50, trEMBL =

10, Swissprot = 100); descriptions were truncated to a maximum of 100 characters long. GO (gene ontology) terms were assigned using Blast2GO (Conesa et al., 2005) with BlastX searches against the NCBI "nr" database as input (e-value 1x10-4); domains were annotated with InterProScan5 (Jones et al., 2014).

TransposonPSI (http://transposonpsi.sourceforge.net/) was used with the *Primula vulgaris* and *Primula veris* (Nowak et al., 2015) coding sequences to detect degenerate repetitive elements based on protein homology, as well as repetitive sequences which otherwise escaped detection with RepeatMasker.

2.3.8 Comparative analysis of *P. vulgaris* and *P. veris* genes

Primula veris coding sequences predicted in Nowak et al. (2015) were aligned to *P. vulgaris* LH_v2 coding sequences and genomic contigs in two separate alignments using TBLASTX (v2.2.31+) (Camacho et al., 2009). In corresponding TBLASTX alignments LH_v2 coding sequences were mapped against the published *P. veris* genomic contigs and coding sequences. In cases where there were multiple isoforms available for each gene, the isoform with the suffix ".1" was regarded as the principal isoform and was used in this analysis. The resulting output files from the alignments were parsed to extract any HSPs with over 95% sequence identity; the total percentage coverage across each coding sequence region for these alignments was recorded, and the cumulative number of coding sequences at each coverage plotted for each of the four alignments using R (v3.2.0) (https://www.r-project.org/).

2.3.9 Analysis of orthologous gene groups

Orthologous and paralogous groups were determined using OrthoMCL (v2.0.9) with the method described on the OrthoMCL website (inflation factor=1.5) (http://orthomcl.org/common/downloads/software/v2.0/UserGuide.txt). Two comparable analyses were performed, using proteins from (i) *Primula vulgaris* LH_v2, and (ii) *Primula veris* (Nowak et al., 2015). The all-vs-all alignments required for analysis with OrthoMCL was carried out with BLASTP (v2.2.28+) with the "seg" and "soft masking" options applied (e-value=$1x10^{-5}$); protein sequences used in the alignments were from *Actinidia chinensis* (http://bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi), *Orzya sativa* (version 7,

http://rice.plantbiology.msu.edu/downloads_gad.shtml), *Arabidopsis thaliana* (TAIR10, https://www.arabidopsis.org/) and *Solanum lycopersicum* (version 2.4, http://solgenomics.net/organism/Solanum_lycopersicum/genome), as well as either *Primula vulgaris* (LH_v2, current study) (i) or *Primula veris* (Nowak et al. 2015) (ii). In cases where there were multiple isoforms available for each gene, the isoform first listed (with the suffix ".1") was regarded as the principal isoform and was used in this analysis. From the OrthoMCL output, Venn diagrams of orthologous genes were drawn using the tool available at http://bioinformatics.psb.ugent.be/webtools/Venn/

2.3.10 RNA-Seq differential expression analysis

RNA was isolated in biological replicates from 15-20 mm buds of four wild-type pin plants and four wild-type thrum plants for sequencing with Illumina HiSeq2000 (Table 2.1), as described in Li et al. (2016). The resulting reads were screened for rRNA removal using SortMeRNA and quality-trimmed with trim galore (Q20) (Kopylova et al., 2012) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore), before alignment to the long-homostyle genome assembly with TopHat (v2.0.13) and assembly with Cufflinks (v2.2.1) (Trapnell et al., 2012) using LH_v2 gene models as a guide after manual curation of all *S* locus genes (see Chapter 4). Differential expression was carried out using the Cufflinks (v2.2.1) tool "Cuffdiff"; genes differentially expressed between pin and thrum flowers, as well as genes specifically expressed in only one morph, were extracted from the "gene_exp.diff" file. R (v3.2.0) (https://www.r-project.org/) was used to plot the $\log_2$ fold change in FPKM for each differentially expressed gene, and the $\log_{10}$ difference in FPKM + 1 for genes with morph-specific expression.

GO term enrichment analysis based on Fisher's exact test was performed using the goatools package for Python (https://github.com/tanghaibao/goatools). GO terms corresponding to the subset of genes identified as differentially expressed, or morph-specific in expression, were compared with GO terms associated with genes across the whole genome, based on the functional annotation with Blast2GO (Conesa et al., 2005).

2.4 Results

2.4.1 Analysis of paired-end reads

Primula vulgaris plants are normally outbreeding (Li et al., 2011b). However, in rare cases homostyle plants with the anthers and stigma at the same height are produced; these plants are associated with a breakdown in self-incompatibility (SI) (Charlesworth and Charlesworth, 1979a). The long homostyle individual used for the LH_v2 assembly is from an inbred line; this plant was highlighted as a possible source for generating the paired-end read library required for a high-quality *Primula vulgaris* genome assembly, based on potentially reduced polymorphic content as compared to an outbreeding *P. vulgaris* plant (Pryszcz and Gabaldón, 2016).

The *k*-mer frequency distribution (spectra) of the LIB2558 library generated from the long homostyle individual reveals that the genome of this plant is highly homozygous, with a unimodal distribution beyond the first local minima (Figure 2.1). The distribution for LIB2558 is free from any noteworthy secondary peaks, indicating minimal heterozygosity within the sequencing reads. The distributions of the pin and thrum read libraries indicate heterozygosity within their genomes (Figure 2.1); the pin distribution suggests it might be difficult to resolve between true genomic content and erroneous sequences for this read library, as there is some overlap between the exponential phase of the distribution and the presumed true genomic content which lies beyond the first local minima (Mapleson et al., 2016).

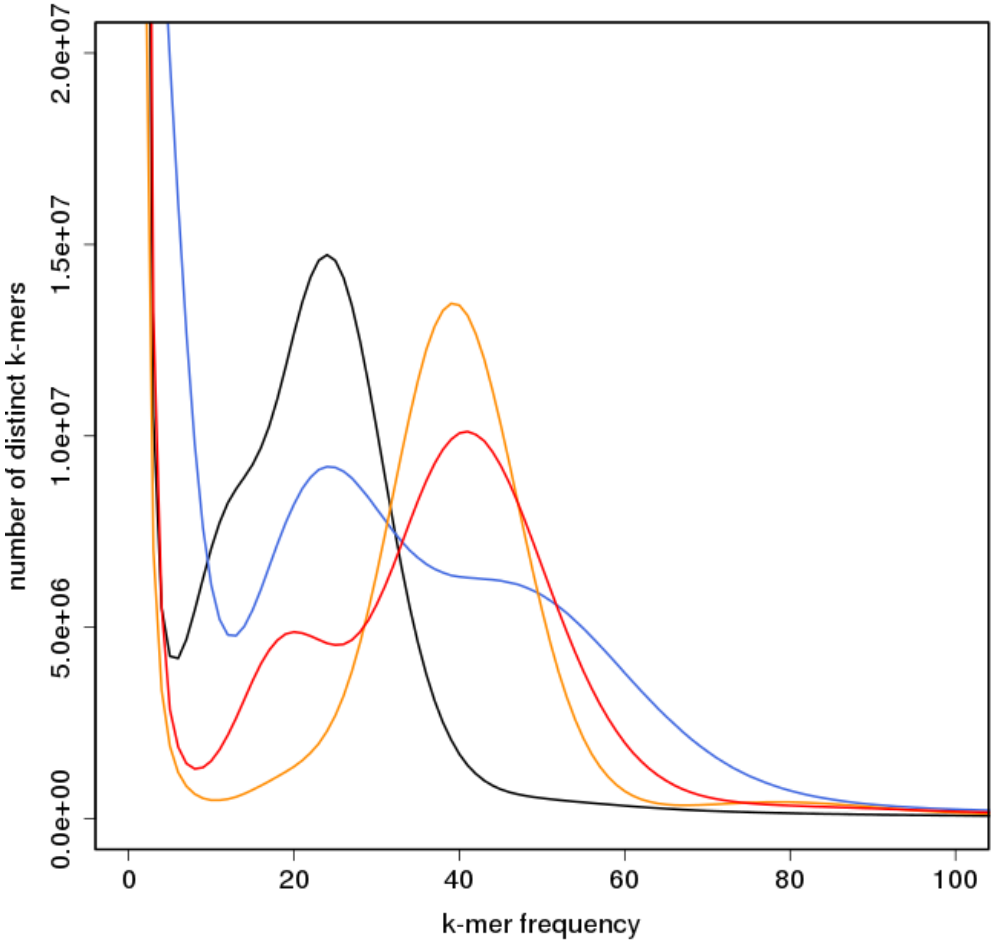


Figure 2.1 Frequency distribution (spectra) of *k*-mers (k=31) showing the number of distinct *k*-mers occurring with each frequency in paired-end read libraries for pin (black), long homostyle (orange), short homostyle (blue) and thrum (red) individuals (LIB1732, LIB2558, LIB4568, LIB1167; Table 2.1).

These results provide a good indication that a long homostyle assembly generated from the LIB2558 library might comprise longer contigs, and an assembly size that more accurately reflects genome size predictions, based on reduced polymorphic content and an improved capacity to recognise sequencing errors in comparison to the other sequenced genomic read libraries (Table 2.1). In a more heterozygous genome with increased polymorphism, such as the thrum sequenced here (LIB1167) (Table 2.1), the resulting assembly would typically be more fragmented, comprising redundant contigs that cannot be resolved by a single path due to the small size of the short sequencing reads. This would potentially result in fragmented gene models and a greater number of

apparent paralogs and duplicated genomic regions than are actually present in the genome (Pryszcz and Gabaldón, 2016).

The *k*-mer spectra for the short sequencing read library used to generate the published *P. veris* genome assembly suggests there is a high level of heterozygosity in this library (Figure 2.2). This is probably due to the use of a read library comprising paired-end reads derived from genomic DNA isolated from both pin and thrum individuals (Nowak et al., 2015). Since wildtype *Primula* plants are usually outbreeding (Li et al., 2011b), both the pin and the thrum genome would be heterozygous with numerous polymorphic sites between the distinct *P. veris* individuals, thus resulting in what appears to be four peaks in the distribution. This would result in numerous groups of alternate contigs that comprise the same polymorphic genomic regions (Pryszcz and Gabaldón, 2016).

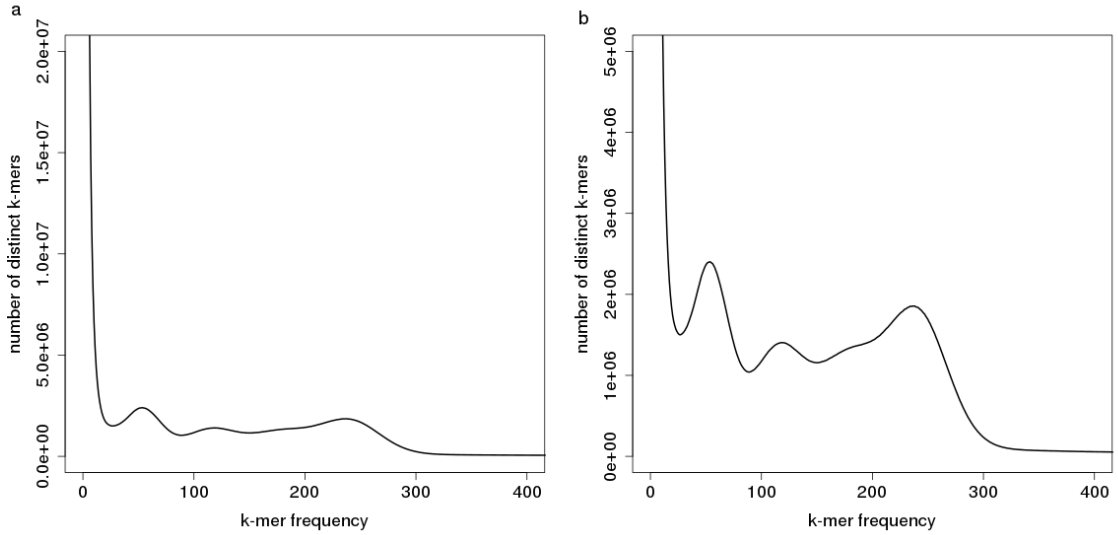


Figure 2.2 Frequency distribution (spectra) of *k*-mers (k=31) showing the number of distinct *k*-mers occurring with each frequency in the paired-end read library used in the published *P. veris* assembly (Nowak et al., 2015); a = on the same scale as Figure 2.1 for comparison, b = reduced y-axis maximum to facilitate visualisation of the distinct peaks in the distribution.

2.4.2 Assembly validation

Genomic paired-end reads (Table 2.1) were used to generate the assemblies listed in Table 2.2. The long homostyle LIB2558 paired-end read library, as well as the long

homostyle long mate-pair (LMP) reads (LIB5215, LIB5216, LIB5217) (Table 2.1) were used in the long homostyle (LH_v2) assembly which, with an N50 of 294.8 kb and NG50 of 229.8 kb, was chosen for subsequent annotation (Table 2.4). The LH_v2 NG50 value was calculated using the higher of the two flow-cytometry estimates for the *P. vulgaris* genome size (Siljak-Yakovlev et al., 2010, Temsch et al., 2010), that is 489 Mb (Siljak-Yakovlev et al., 2010).

The published *Primula veris* genome assembly is 309.7 Mb and has a reported N50 of 164 kb; it covers 65% of the estimated genome size (479.22 Mb) (Siljak-Yakovlev et al., 2010). However, the NG50 calculated with GenomeTools (v1.5.8) (Gordon, 2013) using the estimated genome size of 479.22 Mb, is 73.3 Kb; this considerable difference in N50 and NG50 values (55% decrease) is most likely due to the apparent removal of contigs less than 888 bp, which would result in the removal of a significant portion of the true genomic content from the *P. veris* assembly (Figure 2.5), thereby decreasing the assembly size and increasing the N50 statistic; this is an example of the NG50 metric offering a more appropriate measure of genome assembly completeness, as it is more robust to the removal of true genomic content (Earl et al., 2011, Bradnam et al., 2013). The *P. vulgaris* assembly covers more of the genome and has a higher NG50.

Assembly name	Assembled by	Flower form	Libraries used for assembly	Contig count	Total (Mb)	N50 (kb)	NG50 (kb)
LH_v2	JW (TGAC)	Long homostyle	LIB2558 (scaffolded with LIB5215, LIB5216, LIB5217)	67491	411.1	294.8	229.8
LH_v1	JW (TGAC)	Long homostyle	LIB2558 (scaffolded with LIB1474)	30792	401.0	57.6	37.8
SH_v2	JMC	Short homostyle	LIB3565 (scaffolded with LIB1474)	125497	552.8	12.0	14.2
TP_v2	JW (TGAC)	Thrum	LIB1167 (scaffolded with LIB1474)	124558	625.5	13.7	19.8
TP_v1	JW (TGAC)	Thrum	LIB1167 (not scaffolded)	156809	482.2	8.3	8.1
TP_v1.1	JW (TGAC)	Thrum	LIB1167 (scaffolded with LIB1474)	159254	614.9	12.8	19.3
PP_v1	JMC	Pin	LIB1732	136961	499.6	9.1	9.4
PP_v2	JMC	Pin	LIB1732 (scaffolded with LIB1474)	105238	581.6	14.9	20.0
VT_v1	JMC	*P. veris* Thrum	LIB3564	145617	441. 5	10.8	9.5

Table 2.4 Genome assembly statistics for *Primula vulgaris* (and *P. veris*, where listed) paired-end read assemblies carried out by JMC or JW (TGAC). NG50 is calculated using a genome size of 489 Mb (*Primula vulgaris*) or 479.22 (*Primula veris*) (Siljak-Yakovlev et al., 2010). In all cases contigs < 200 bp were removed prior to these calculations.

In summary, the LH_v2 genome assembly of 411 Mb covers between 84 and 90% of the 459-489 Mb genome (Siljak-Yakovlev et al., 2010, Temsch et al., 2010) and offers a considerable improvement over the *P. veris* assembly in terms of NG50 value, suggesting most of the *Primula* genome content is in contigs of comparatively larger size in the LH_v2 assembly. Furthermore, the 309.7 Mb *P. veris* assembly contains 40.7 Mb (13.14%) "N"s (ambiguous bases) as opposed to 29.9 Mb (7.26%) in the 411 Mb *P. vulgaris* assembly, suggesting LH_v2 comprises a greater proportion of resolved base calls. This is perhaps due to the homozygosity of the long-homostyle allowing highly polymorphic sites to be assembled into contigs (Pryszcz and Gabaldón, 2016); this would result in fewer contigs having to be scaffolded together with LMP libraries, which would result in intervening "N"s between joined sequences.

The final LH_v2 assembly discussed above was filtered to remove contigs less than 200 bp. The comparison of *k*-mers present in the LH_v2 genome assembly (without 200 bp contigs removed) and the paired-end reads used to generate it suggests that this precursor of the LH_v2 assembly incorporates the majority of true genomic content that lies beyond the first local minima (Figure 2.3); the *k*-mers in the exponential phase of the plot, which represent unique sequencing errors present in low frequencies (Sato et al., 2012), are for the most part not included in the genome assembly (0x coverage in black), with the homozygous content present once in the assembly (red), suggesting the assembly is highly "collapsed", with few alternate contigs that would otherwise result from unresolvable polymorphic regions of the genome (Mapleson et al., 2016, Pryszcz and Gabaldón, 2016).

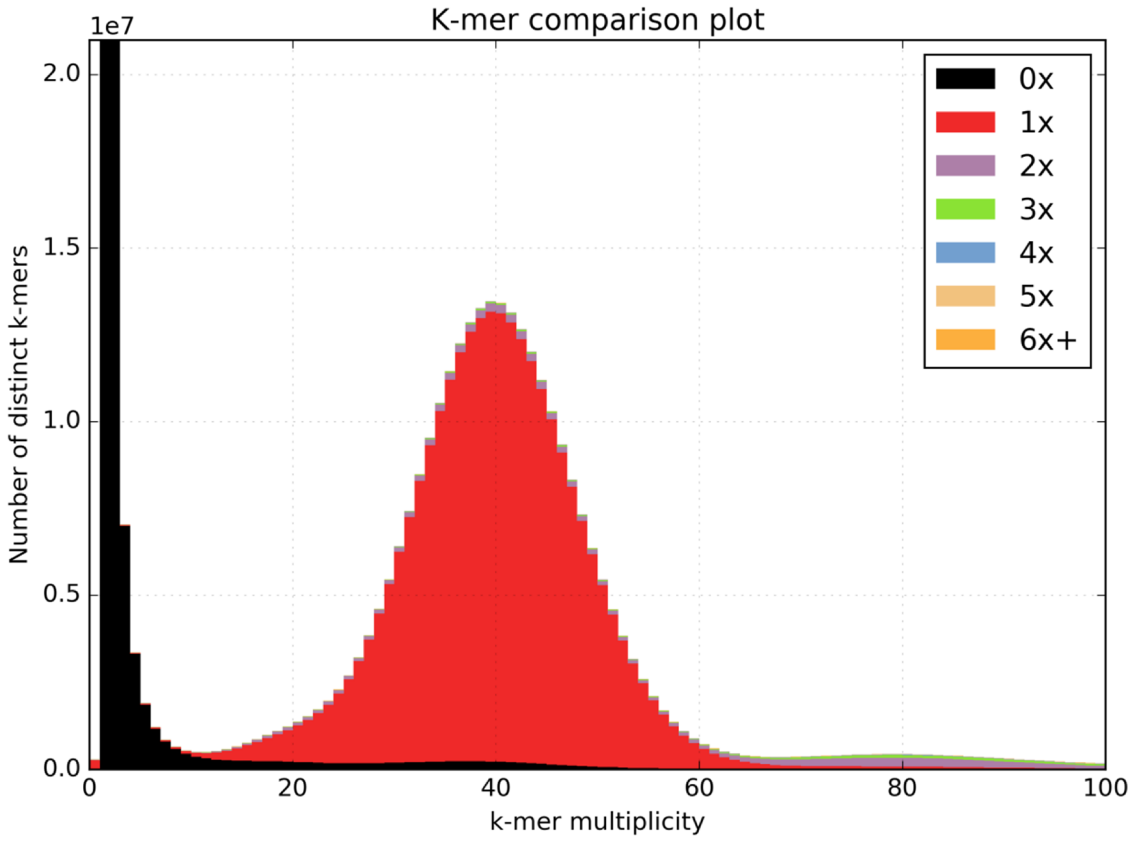


Figure 2.3 Frequency distribution (spectra) of *k*-mers in the LIB2558 read library, and the copy number of these *k*-mers in the LH_v2 assembly (prior to removal of contigs < 200 bp); black = content absent from the assembly, red = content present once in the assembly, purple = twice, etc.

Based on the validation steps described above, the LH_v2 genome with contigs < 200 bp removed was chosen for annotation and analysis of genes and repetitive sequences; the removal of "ultra-small" contigs < 200 bp is a common step (e.g. Mayer et al., 2014) as small contigs might comprise highly repetitive sequences, or sequencing errors, and are not incorporated into larger contigs, or would otherwise slow down downstream analyses. In this instance the removal of contigs < 200 bp does not remove a significant portion of genomic content in the LH_v2 assembly as there is little change in the *k*-mer spectra of the assembly with these contigs removed (Figure 2.4) in comparison to Figure 2.3.

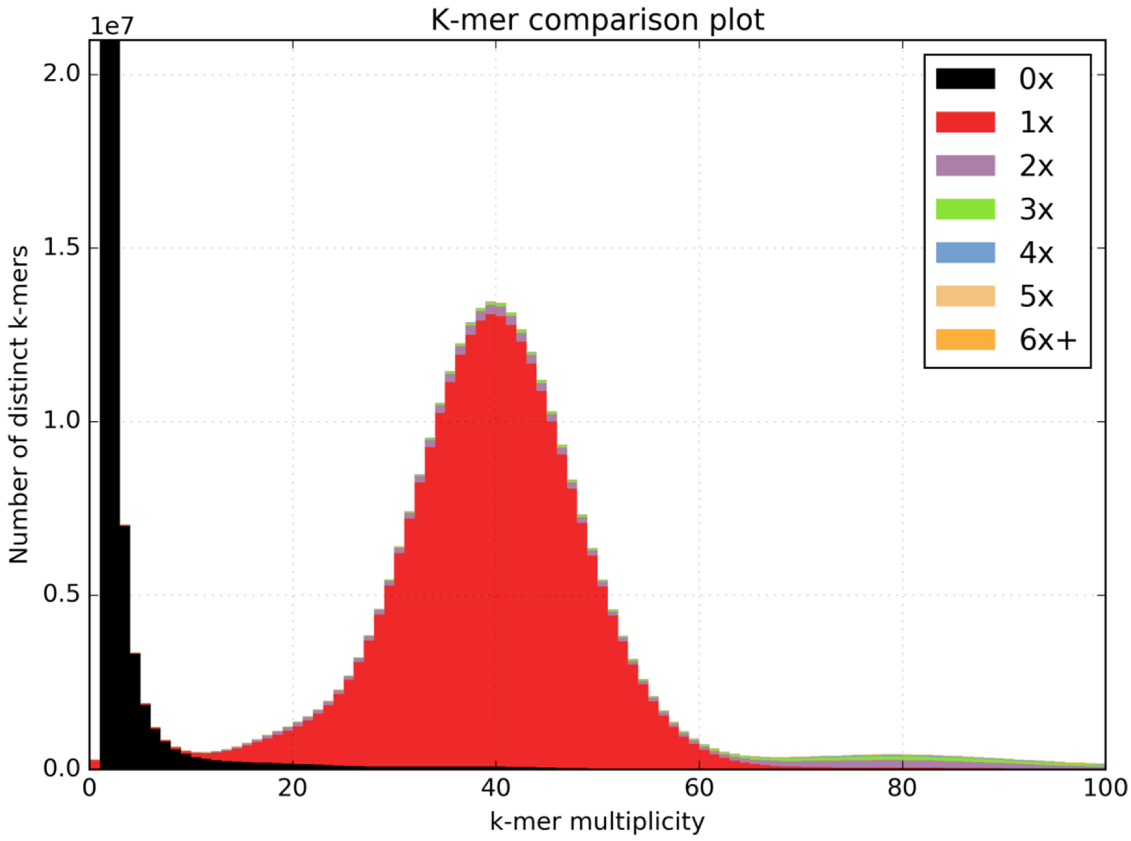


Figure 2.4 Frequency distribution (spectra) of *k*-mers in the LIB2558 read library, and the copy number of these *k*-mers in the LH_v2 assembly (with removal of contigs < 200 bp); black = content absent from the assembly, red = content present once in the assembly, purple = twice, etc.

For completeness, the equivalent analysis for *P. veris* paired-end reads against the published genome assembly (Figure 2.5) was performed. This indicates that a significant proportion of the true genomic content in the read libraries is missing from the assembly. This is perhaps a result of the difficulty in distinguishing between erroneous *k*-mers and true genomic content in the highly-polymorphic read library, as well as the apparent use of a ≥ 888 bp cut-off for contig size. The *k*-mers that are present only once in the read library most likely result from reads containing sequencing errors; if the true genomic content is also represented by *k*-mers with a low copy number as compared to the main distribution, then a partial assembly will result.

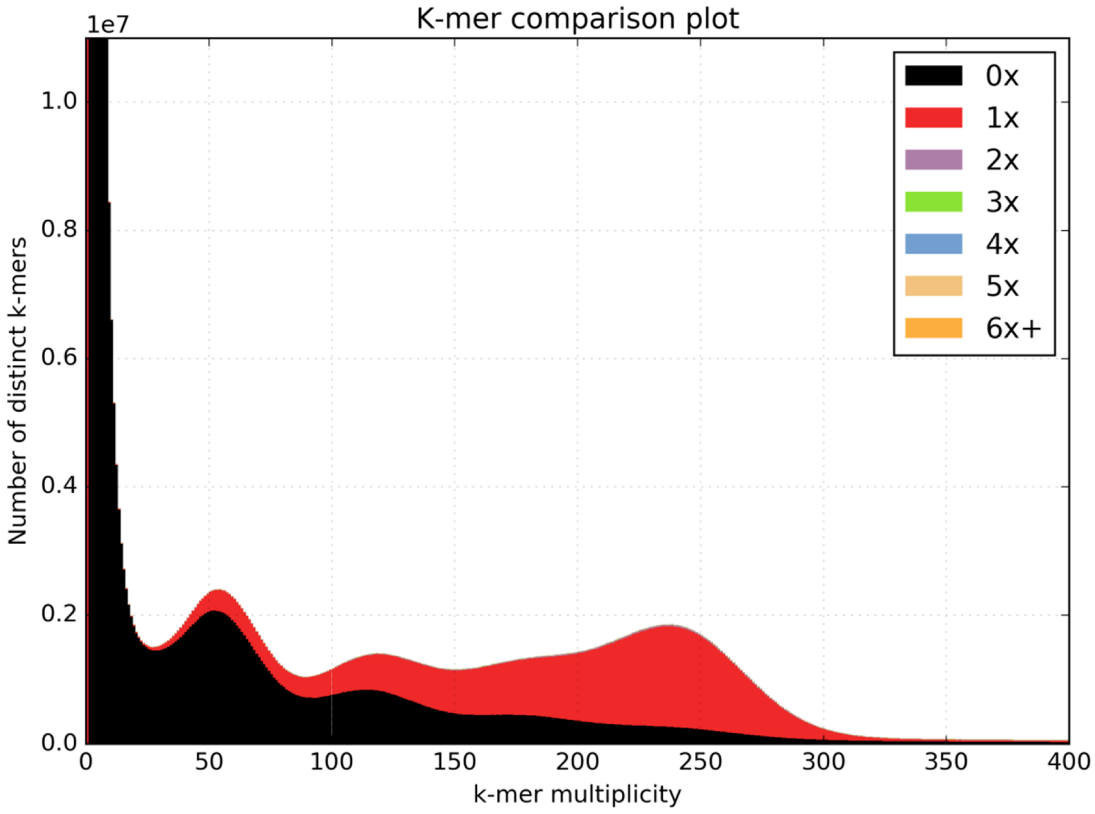


Figure 2.5 Frequency distribution (spectra) of *k*-mers in the *P. veris* SRR1658103 paired-end read library, and the copy number of these *k*-mers in the published *P. veris* assembly (Nowak et al., 2015); black = content absent from the assembly, red = content present once in the assembly, purple = twice, etc.

2.4.3 Evaluation of the genespace captured in the assembly

The LH_v2 assembly was inspected for the inclusion of 248 CEGs; these genes are expected to be present in the majority of eukaryotic genomes (Parra et al., 2007). This is a useful measure of whether the genespace of an organism has been captured by the assembly (Simão et al., 2015). The vast majority (97.18%) of CEGs are at least partially present in the LH_v2 assembly, which suggests a high level of completeness for a non-model species. The published *Primula veris* genome (Nowak *et al.*, 2015) partially covers 94.18% of CEGs, which by proxy suggests a greater number of genes may be absent from this assembly. The use of partial matches is to avoid biases in relation to method used for alignment (Parra et al., 2007, Parra et al., 2009).

	No. proteins	Completeness (%)
Complete	223	89.92
Group 1	61	92.42
Group 2	46	82.14
Group 3	53	86.89
Group 4	63	96.92
Partial	241	97.18
Group 1	63	95.45
Group 2	53	94.64
Group 3	61	100.00
Group 4	64	98.46

Table 2.5 The number and percentage of 248 ultra-conserved CEGs present (either complete or partial) in the *Primula vulgaris* LH_v2 genome, as determined by CEGMA (v2.5) (Parra et al., 2007).

RNA-Seq datasets derived from leaves, flowers, roots, fresh seed, and seedlings from both pin and thrum *P. vulgaris* plants (Table 2.1 and Table 3.1) (see section 2.3.6) were mapped to the LH_v2 genome (by JW) using the DRAGEN co-processor (http://www.edicogenome.com/dragen_bioit_platform/). The mean and mean concordant (paired) mapping rate was 98.5% and 88.1% respectively. The TopHat alignment of RNA-Seq reads (used in our expression analysis; see section 5.3.3) to the published *P. veris* assembly gives a mean overall mapping rate of 82.5%, and a concordant pair alignment rate of 75.7%. In both cases RNA-Seq reads were filtered for ribosomal RNA with sortmeRNA v1.9 (Kopylova, E. et al., 2012) and quality-trimmed (Q20) using trim galore v0.3.3 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), prior to short-read alignment. These RNA-Seq reads were used in gene prediction for the respective genomes. In conclusion, despite RNA-Seq reads from a much wider range of tissues and

a greater number of libraries being mapped to LH_v2, the mapping rates suggest a more complete genespace for this assembly.

It has been suggested that a good gauge of the "usefulness" of an assembly is to determine the number of scaffolds with a length greater than that of an average gene (Halder et al., 2010). Based on assembled and annotated vertebrate genomes, this was calculated as 25,000 bp (Halder et al., 2010). The average gene length in rice is 3,223 bp (RGAP 7; http://rice.plantbiology.msu.edu/), and in *Arabidopsis* 2,300 bp (TAIR10; https://www.arabidopsis.org/). Here, the average of these two values (2,761.5 bp) was used as the approximate length of an average angiosperm gene. To this end, ~303 Mb of sequence in the *P. veris* genome assembly comprises contigs > 2761.5 bp, versus ~388 Mb for *P. vulgaris* LH_v2, suggesting both genomes are adequate for gene prediction, but that the *P. vulgaris* assembly contains more total sequence of sufficient length for annotation (Figure 2.6). However, the approximate average gene length of 2,761.5 bp for angiosperms is much lower than that calculated for vertebrates (25,000 bp), suggesting the measure of relative "usefulness" used by Bradnam et al. (2013) is not as applicable for plants due to expansion-limited intron sizes in plants as compared to vertebrates (Wu et al., 2013) and the apparent ease of assembling contigs (and thus genes) less than 5 kb in length (Figure 2.6).

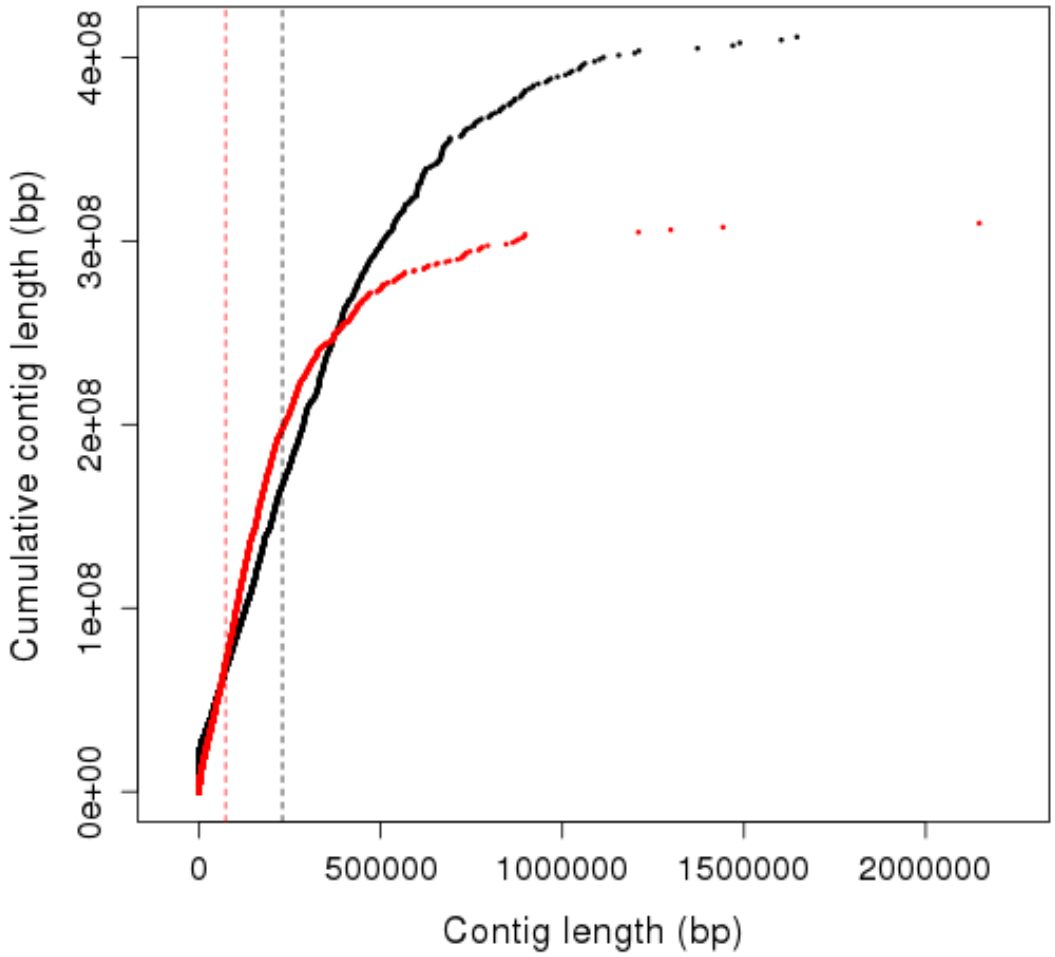


Figure 2.6 Contribution of genomic contigs of different lengths (x-axis) to total cumulative size of genome assembly (y-axis), for *P. veris* (red) (Nowak et al., 2015) and *P. vulgaris* LH_v2 (black). The dashed lines indicate the respective NG50 values for the assemblies (red = veris, black = vulgaris).

The validation methods above suggest that comprehensive annotation and analysis of the assembled *P. vulgaris* LH_v2 genome is worth pursuing based on a high-quality and relatively complete genome assembly.

2.4.4 Repeats in the *Primula vulgaris* genome

The species-specific *de novo* repeat library generated for *P. vulgaris* through curation of the RepeatModeler output allowed 37.03% of the 411 Mb *P. vulgaris* LH_v2 assembly to be annotated as repetitive (Table 2.6), with up to 36% comprising transposable elements (TEs); this is comparable with the predicted TE content in the similar sized (370 Mb) rice assembly (> 35%), as well as the closely-related and more recently annotated 616.1 Mb *Actinidia chinensis* (kiwifruit) genome (36%). In *P. vulgaris* up to 66.14% of annotated TEs were categorized as long terminal repeat (LTR) retrotransposons, in line with previous observations that LTRs are the most widespread TEs in plants (Kubat et al., 2014).

Repeat type	Length (bp)	% in genome	% in repeat
Interspersed repeats	153945885 (153945885)	37.44 (37.44)	96.15 (96.15)
Class I: Retroelement	62615656 (95912158)	15.23 (23.33)	39.11 (63.68)
LTR Retrotransposon	54323780 (66959990)	13.21 (16.28)	33.93 (41.82)
Copia	39875915 (39875915)	9.70 (9.70)	24.91 (24.91)
Gypsy	12544567 (12544567)	3.05 (3.05)	7.83 (7.83)
Other LTR	1903298 (14539508)	0.46 (3.54)	1.19 (9.08)
non-LTR Retrotransposon	8291876 (28952168)	2.02 (7.04)	5.18 (18.08)
SINE	1085346 (1363804)	0.26 (0.33)	0.68 (0.85)
LINE	7206311 (20717729)	1.75 (5.04)	4.50 (12.94)
other non-LTR	219 (6870635)	< 0.01 (1.67)	< 0.01 (4.29)
Other Class I	219 (6045423)	< 0.01 (1.47)	< 0.01 (3.78)
Class II: DNA transposon	19522480 (52043722)	4.75 (12.66)	12.19 (32.50)
Unclassified interspersed	71807749 (5990005)	17.46 (1.46)	44.85 (3.74)
Simple repeats	6164398 (6164398)	1.50 (1.50)	3.85 (3.85)
Total (uncorrected)	160110283 (160110283)	38.94	100.00 (100.00)
Total (corrected for overlaps)	152243370	37.03	

Table 2.6 Repetitive sequences annotated in the *Primula vulgaris* genome using RepeatMasker with a *de novo* repeat library for *P. vulgaris*. Length (base-pairs), and % in genome and repetitive portion of the genome is shown for each repeat class; additional classification with TEclass (Abrusan et al., 2009) is shown in brackets. Total length of all repeat types (bp) is shown either uncorrected or corrected for overlapping annotations based on the RepeatMasker output. The sequence percent for each individual repeat type is not corrected for overlaps.

The *de novo* repeat library generated for *P. vulgaris* was used to annotate 34.62% of the draft VT_v1 *P. veris* thrum genome (Table 2.2) as repetitive, which is much greater than the reported 7.7% for the published *P. veris* genome, suggesting the true proportion of repeats might be higher than 7.7% if a species-specific repeat library is generated for *P. veris* (Nowak et al., 2015). Indeed, using the *P. vulgaris* repeat library, ~26% of the published *P. veris* genome was annotated as repetitive. This perhaps highlights that a significant portion of the repeat content has been removed, or was not assembled due to comparatively high polymorphism between the multiple individuals used in sample preparation, as opposed to the preliminary VT_v1 assembly that adopts a more conservative ≥ 200 bp cut-off and was generated from a read library of one sequenced individual; this is consistent with findings in Figure 2.5 that show a large portion of the read library unrepresented in the published *P. veris* assembly.

The comprehensive repeat library generated for *Primula vulgaris* will facilitate comparisons between the *S* locus and surrounding genomic regions once the genes at this locus have been identified (see Chapter 4).

2.4.5 Gene annotations in the *Primula vulgaris* genome

In total 24,600 genes (principal isoforms) were predicted in the *P. vulgaris* LH_v2 genome assembly with an mean coding sequence length of 1,466 bp. This is a similar number of predicted genes to the 27,655 predicted in *A. thaliana* (27,411; TAIR10, https://www.arabidopsis.org/), but less than that in the *Actinidia chinensis* (kiwifruit) genome (39,040) (Huang et al., 2013). Kiwifruit is the most closely related sequenced plant species other than *P. veris*, which has a reported 19,507 (18,301 published) predicted genes. The comparatively large number of genes in the *Actinidia chinensis* genome is most likely due to recent whole genome duplication events in this species (Huang et al., 2013), as also noted by Nowak et al. (2015).

Of the 24,600 predicted *P. vulgaris* protein sequences, 84% were functionally annotated with AHRD based on homology to SwissProt, TrEMBL (http://www.uniprot.org/) and TAIR10 (https://www.arabidopsis.org/) protein databases using BLASTP (Camacho et al., 2009), and additional searches with Interproscan and BLAST2GO (Conesa et al., 2005, Jones et al., 2014). Of these, 90% contain at least one domain, and 58% have GO terms attached; 8% of the 24,600 genes have no descriptions based on homology to

proteins via the BLASTP alignments, but are nonetheless annotated with GO terms or domains.

Based on the results of additional homology-based searches for potentially degenerate repeat elements using TransposonPSI (http://transposonpsi.sourceforge.net/), as well as an accompanying lack of non TE-related descriptions in the functional annotations, up to 762 of the 24,600 predicted genes in LH_v2 could contain TEs. If this is the case, then the actual number of genes in the *P. vulgaris* genome is perhaps closer to 23,838. Unless, of course, some of these genes are in fact protein coding, but contain domains that share similarity to TE-related domains, as is the case with the AP2 binding domain that is present in both plant developmental transcription factors (TFs) and integrases such as tn916 (Balaji et al., 2005). There is a suggestion that recruitment of DNA binding domains in TFs from transposases or integrases could be a recurrent theme in evolution (Balaji et al., 2005), with evolutionary mobile protein domains seen in different sequence contexts (Triant and Pearson, 2015).

2.4.6 Comparison of genes in *P. vulgaris* and *P. veris* genomes

The above *k*-mer analyses show that the *P. vulgaris* LH_v2 genome assembly incorporates most of the genomic content present in the reads (Fig. 2.4). CEGMA (Table 2.5) and RNA-Seq alignment analyses suggest that the genespace has been successfully captured. The *P. vulgaris* assembly is therefore expected to include most of the genes in the genome. The *P. veris* assembly has a slightly lower percentage of CEGs partially mapped to it (94.18% vs. 97.18%), and a much lower percentage of CEGs with complete matches (79.84% vs. 89.92%) (Nowak et al., 2015). The RNA-Seq read mapping rate is also lower than that for *P. vulgaris*, and *k*-mer analyses show that some of the genomic content may have been lost due to the polymorphic read library (Fig. 2.5).

To evaluate the different number of genes annotated in the *P. vulgaris* LH_v2 and published *P. veris* genome assemblies (24,600 vs. 18,301) the predicted coding sequences were compared against each other and the respective genome assemblies, as shown in Figure 2.7. The results show that 6,502 coding sequences out of 24,600 genes predicted in the LH_v2 assembly (26.43%) have no coverage (> 95% identity) in the *P. veris* coding sequences (Figure 2.7). Furthermore, the alignment of *P. veris* coding

sequences to *P. vulgaris* coding sequences (red line) and *P. vulgaris* contigs (blue) results in a shallower distribution compared to reverse alignments of *P. vulgaris* coding sequences to *P. veris* (black/orange). Therefore, this analysis also suggests that *P. vulgaris* coding sequences mapped to the *P. veris* assembly do so with lower coverage than *P. veris* coding sequences mapped to *P. vulgaris*, as supported by a reduced number of CEGs with complete coverage in the *P. veris* genome (79.84% vs. 89.92%). It should be noted that the *P. veris* genome manuscript (Nowak et al., 2015) quotes the number of genes in the assembly as 19,507, whilst final available file of coding sequences contains 18,301 sequences. This is presumably due to the apparent use of a ≥ 888 bp cut-off for contig size in the assembly, which would result in some of the annotated genes being removed from the final assembly (Figure 2.5).

Of these 6,502 coding sequences, 1,166 are completely absent from the *P. veris* genome assembly (blue line). The absence of these genes might be expected based on the failure of the *P. veris* assembly to incorporate all of the genomic content in the reads (Fig. 2.5). This analysis therefore clarifies that the lower number of genes annotated in the *P. veris* assembly is not a result of gene duplication events in *P. vulgaris* or genes that are otherwise species-specific, but that the *P. veris* annotations are not as comprehensive as those in LH_v2, and do not include a sizeable portion of the genes in the genome. This could be due to the use of single library RNA-Seq datasets from a narrow range of tissues (leaves and flowers) compared to the multiple-replicate datasets used for the LH_v2 annotations (flowers, leaves, roots, fresh seed, and seedlings): unreplicated RNA-Seq data may fail to capture biological variance (see section 3.4.3). The broader range of tissues for *P. vulgaris* is likely to offer a more comprehensive sampling of the expressed genes. The use of a highly polymorphic genomic read library containing both pin and thrum reads (Figure 2.2) may also have confounded the *P. veris* assembly itself due the difficulty of assembling heterozygous genomes, resulting in the assembly excluding true genomic content (and thus genes) from the reads (Figure 2.5) (see section 2.4.2).

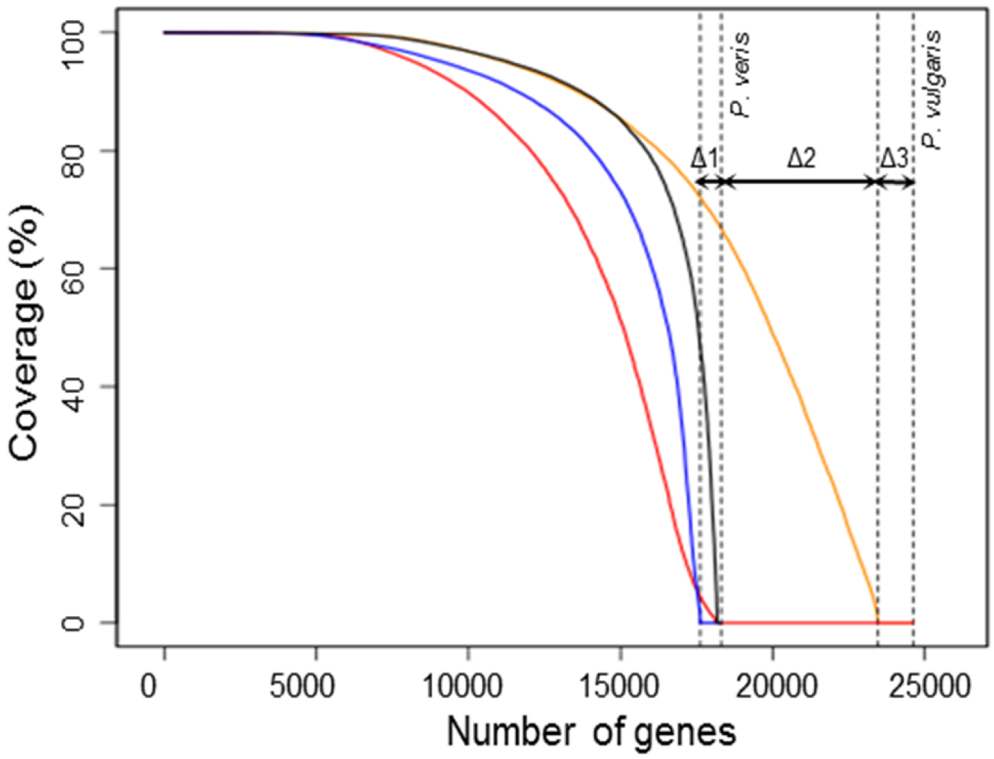


Figure 2.7 Comparison of genes annotated in *P. vulgaris* LH_v2 and published *P. veris* genome assemblies: total coverage of HSPs (High Scoring Pairs) with ≥ 95% sequence identity in TBLASTX alignments. LH_v2 coding sequences aligned to *P. veris* genome assembly (orange), and *P. veris* coding sequences (red). *P. veris* coding sequences mapped to LH_v2 genome assembly (black), and LH_v2 coding sequences (blue). Dotted lines indicate total number of genes annotated in each genome (*P. veris* = 18,301, *P. vulgaris* = 24,600). Δ1 = number of *P. veris* coding sequences with no coverage in LH_v2 coding sequences (685; 17,616 of 18,301 present); Δ2 = number of LH_v2 coding sequences with no coverage in *P. veris* coding sequences (6,502; 18,098 of 24,600 present); Δ3 = number of LH_v2 coding sequences with no coverage in *P. veris* genome assembly (1,166; 23,434 of 24,600 present). The small number of *P. veris* genes with no coverage in the *P. vulgaris* genome (130; 18,171 of 18,301 present) is not indicated. In contrast, there are only 685 (as opposed to 6,502) *P. veris* coding sequences with no coverage (> 95% identity) in the *P. vulgaris* geneset; with a minimal number (130) absent from the genome assembly as a whole. This suggests that the majority of *P. veris* coding sequences are covered by the *P. vulgaris* genome and geneset; conceivably those that are missing could be species-specific.

In addition to the above, TransposonPSI (http://transposonpsi.sourceforge.net/) annotations of the *P. veris* geneset suggest that 226 of the genes could be TE-related; as noted above, up to 762 genes in the *Primula vulgaris* LH_v2 annotations could be TE-related. This could explain at least some of the 6,502 coding sequences that are absent from the *P. veris* annotations.

Nowak et al. (2015) note a discrepancy between the number of assembly-based gene predictions (19,507), and the number of transcripts obtained in the *de novo* transcript assembly for *P. veris* (25,409); this was produced using Trinity (Grabherr et al., 2011). These *de novo* assembled transcripts without use of the assembly as a reference may represent the full complement of genes in the genome without limitation from missing genomic content, but it seems more likely that genes are still missing due to use of fewer tissue samples, and that a good proportion are instead alternative spliceoforms or fragmented transcripts, as is often the case with *de novo* transcriptome assemblers that can overestimate the number of transcripts compared to that expected for a given organism (Zhao et al., 2011, Bankar et al., 2015, Sayadi et al., 2016). Our final annotations comprise 29,088 coding sequences, with 4,488 recognised as alternative spliceoforms.

2.4.7 OrthoMCL analysis of orthologous genes

Further analysis of the full complement of genes in *P. vulgaris* to determine orthologues in well-annotated and closely related angiosperm species will provide a platform for future understanding of the gene families in which the *S* locus genes lie. This was undertaken through alignment of proteins from each of these species and analysis with OrthoMCL to produce clusters of related proteins (Li et al., 2003) (Figure 2.8). To provide a comparison, *P. veris* proteins were also aligned to these species, and vice-versa; due to differences in the number of predicted genes that may affect the clustering stage of the analysis, the two *Primula* species were investigated in two separate sets of alignments.

These analyses resulted in a sizeable number of orthologues being identified in both *Primula* species. In the *P. vulgaris* analysis, a total of 19,861 groups were identified, as compared to 19,448 in *P. veris*. *P. veris* performs relatively well in this analysis, presumably because the genes in question are those that are well-conserved across

angiosperm species; such genes are used as evidence in annotation pipelines (Holt and Yandell, 2011).

There are slight differences in the number of elements within each grouping. It is difficult to speculate on the exact cause of such differences, but the most noticeable change lies in the number of *P. vulgaris*- and *P. veris*-specific groups (853 vs. 224). It appears that the number of *P. vulgaris*-specific gene families (853) is more in line with that seen for the other species, which suggests that the missing genes in *P. veris* may have affected the clustering algorithm in such a way that some missing paralogous genes are no longer identified in this group, leading to their partners being recognised as more closely related to the groups of genes in other species.

This resource will facilitate downstream pairwise comparison of paralogous genes in the *P. vulgaris* genome, as well as phylogenetic analyses of the *S* locus genes once they have been identified.

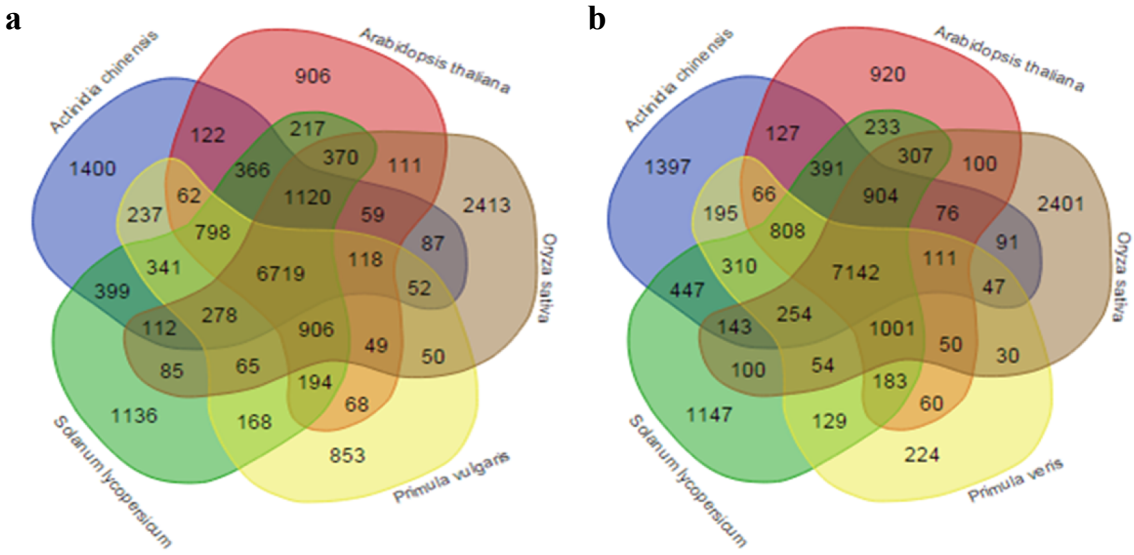


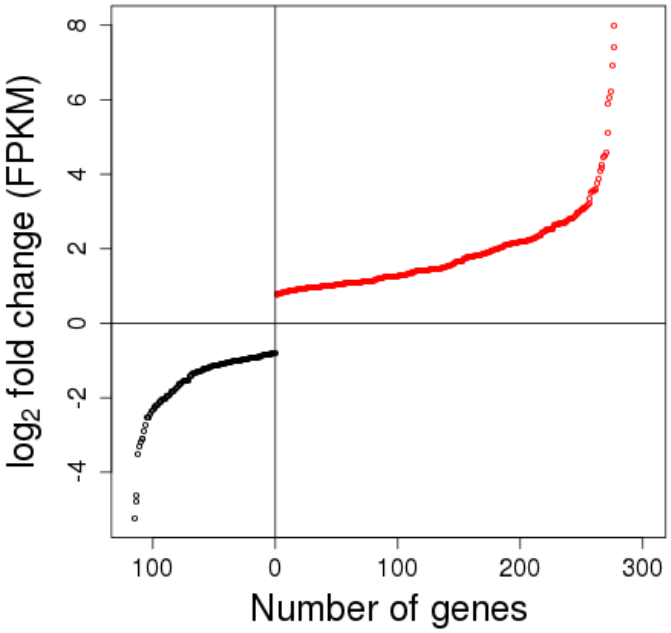
Figure 2.8 OrthoMCL analysis showing orthologous genes between *P. vulgaris* and four other angiosperm species (a), with alignments based on predicted protein sequences in the LH_v2 genome. For comparison, the same analysis is shown (b) using published *P. veris* protein sequences (Nowak et al., 2015).

2.4.8 Differential expression

Differential expression analysis for all genes predicted in the *P. vulgaris* genome was carried out between pin and thrum flowers using RNA-Seq reads in 4x biological replicate. This study identified 401 genes differentially expressed between pin and thrum flowers, and 994 with morph-specific expression (Figure 2.9).

The most distinct feature of this analysis is that there are many more genes upregulated in thrum (383), as compared to the number upregulated in pin (118) (Figure 2.9). This includes eight genes expressed in a morph-specific manner, with no expression in one of the floral forms. There are an approximately equal number of genes expressed only in thrum flowers (525), or only in pin flowers (468).

a



b

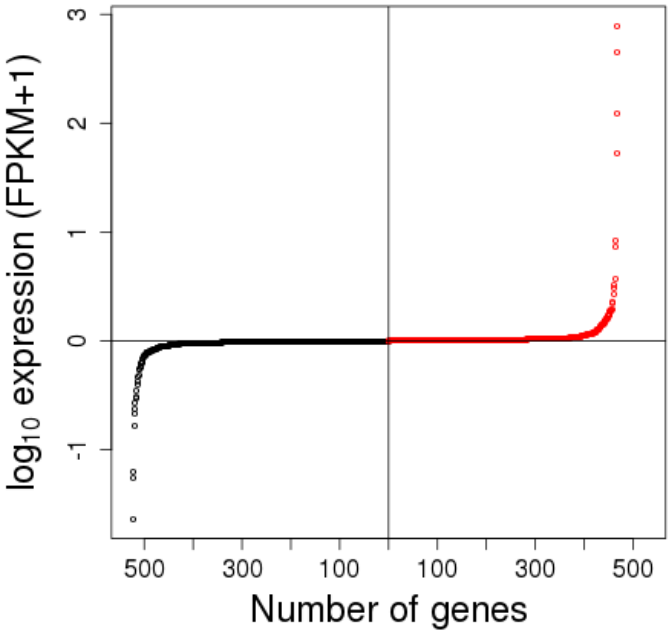


Figure 2.9 (a) genes upregulated in pin (black) and thrum (red) flowers, with $\log_2$ fold-change in expression (FPKM) shown relative to thrum; genes with morph-specific expression are not plotted; (b) genes specifically expressed in either pin (black) or thrum (red) flowers, $\log_{10}$ expression (FPKM+1) relative to thrum.

GO terms associated with genes in the sets of differentially expressed and morph-specific genes (Figure 2.9) were extracted from the LH_v2 functional annotations. GO-term enrichment analysis was carried out to identify over-represented GO terms as compared to the frequency of occurrence for GO terms attached to the full complement of genes in the LH_v2 assembly.

In the functional annotations for the set of 401 differentially expressed genes, GO terms involved in cell wall modification and potentially SI-related pathways are overrepresented (Figure 2.10). This highlights the genes as putative targets in the downstream regulatory pathway of the *S* locus genes that control differential cell division and elongation between the pin and thrum floral forms (Webster and Gilmartin, 2006).

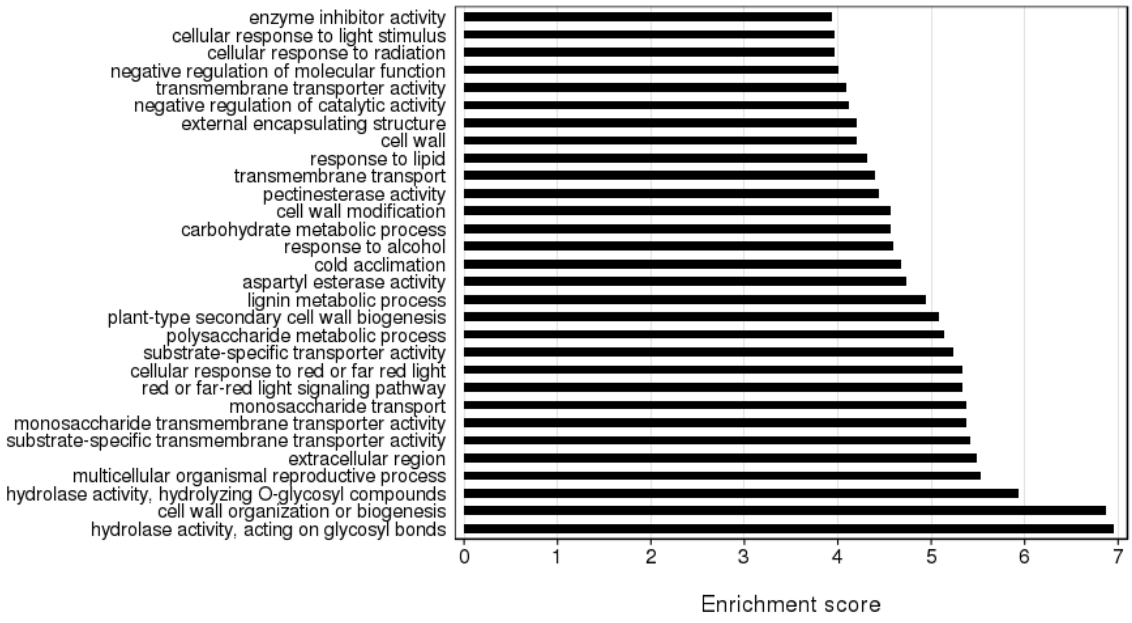


Figure 2.10 Gene Ontology (GO) terms in the set of 401 genes upregulated in thrum flowers (Figure 2.9), and their associated GO-term enrichment scores (top 30 over-represented GO terms (False Discovery Rate (FDR) < 0.1) are shown). Enrichment score = –log10(p-value), p-value as calculated in goatools GO-term enrichment analysis versus GO term occurrence in the *P. vulgaris* LH_v2 gene annotations (https://github.com/tanghaibao/goatools).

In contrast, overrepresentation of such GO terms was not found for the genes with morph-specific expression, casting doubt on the importance of such genes in the downstream regulatory pathway. However, this set of genes includes all genes expressed in only one of the floral forms, including those with extremely low (in many cases < 0.1 FPKM) expression. If subsets of these genes with > 0.1 FPKM or > 1 FPKM expression are taken, then it remains that there is no enrichment of GO terms of apparent interest in relation to heterostyly. In scrutinising the functional annotations directly, it appears that some of the genes identified, for example those with similarity to F-box transcription factors, may be of potential interest with regards to a role in SI; nonetheless, genes expressed in a morph-specific manner by and large seem to be of limited importance in terms of identifying the broad downstream targets of the *S* locus.

2.5 Discussion

This chapter describes a broad range of assemblies and associated annotations that are poised to serve as a robust platform for future genomic analyses within the Primulaceae. *Primula vulgaris* is an outbreeding angiosperm with a putatively sporophytic self-incompatibility (SI) system that acts as a safeguard against self-fertilization (McCubbin, 2008, Li et al., 2011b), as such it might reasonably be expected that a wild-type primrose would have a highly heterozygous genome composition. The level of polymorphism or heterozygosity within a genome further confounds problems associated with the assembly of short sequencing reads, as the assembly algorithm is charged with resolving multiple polymorphic versions of a genome within one sequencing library (Pryszcz and Gabaldón, 2016).

In some cases it is inevitable that a heterozygous genome must be assembled, as with the heterozygous diploid *Trifolium pratense* (red clover) that is difficult to inbreed without severe loss of viability and vigour due to a gametophytic SI system (De Vega et al., 2015). In contrast, homostyle primroses are associated with a breakdown in SI; they are self-fertile, so given a sufficient level of inbreeding, perhaps could provide a highly-homozygous source for genome assembly. In choosing a *P. vulgaris* genome based on an inbred long-homostyle for annotation, the results presented here show that the homozygosity of this individual has been exploited, thus producing a reference genome sequence that is more contiguous, complete and compact than the published *P. veris*

assembly, covering 411 Mb (84-90%) of the estimated 459-489 Mb genome (Siljak-Yakovlev et al., 2010, Temsch et al., 2010) as compared to 65% for *P. veris*. In addition, draft assemblies and short sequencing read libraries for separate pin, thrum and short-homostyle individuals will offer invaluable insight into the specific differences between pin and thrum floral-morphs in trying to formulate a complete picture of the *S* locus and the surrounding regions.

Here, it is shown through an integrated approach combining RNA-Seq, repeat annotations and evidence from closely related angiosperm species, that the diploid *Primula vulgaris* long-homostyle (LH_v2) genome has 24,600 predicted genes. This is perhaps a reasonable estimate for the total number of genes to be expected in the closely related *Primula veris* genome (Nowak et al., 2015), as well as other diploid species in the Primulaceae that have not undergone whole-genome duplications or otherwise been subject to expansion of discrete gene families. The set of comprehensive *P. vulgaris* genes and associated functional annotations will serve as a blueprint for defining the genes in other Primulaceae species, as well as for the rapid functional classification of genes resulting from various expression and exploratory genomic analyses in this family.

The *P. veris* genome assembly has 18,301 annotated genes (Nowak et al., 2015), which contrasts with 24,600 genes in the LH_v2 assembly. From the results presented in this chapter, it appears that the majority of these genes are unannotated, perhaps resulting from the minimal number of tissues represented in the *P. veris* RNA-Seq dataset. Furthermore, a subset of those genes are completely absent from the *P. veris* genome. Due to the close relationship of *P. veris* and *P. vulgaris*, which can interbreed to produce hybrids known as false oxlip (Gurney et al., 2007) one might expect their similarly-sized diploid genomes (Siljak-Yakovlev et al., 2010, Temsch et al., 2010) to contain a similar number of genes. This suggests that some of the genomic DNA content is not incorporated into the *P. veris* assembly, as supported by the apparent absence of many seemingly error-free *k*-mers that are present in the associated genomic read library. The *Primula vulgaris* LH_v2 genome presented here represents a significant leap forward in terms of the accessible functionally-annotated genespace for the Primulaceae. In addition, a more comprehensive repeat analysis for *P. vulgaris* with 37.03% of the genome annotated as repetitive in comparison to 7% in *P. veris*,

highlights the potential usefulness of the repeat library generated in this study for the annotation of repeats in other *Primula* species, as well as for the analysis of the genomic architecture of the *S* locus once it is revealed.

The data presented in this chapter provide the first broad, robust set of differentially expressed genes between pin and thrum flowers for the Primulaceae, using RNA-Seq data based on four biological replicates from *Primula vulgaris* pin and thrum flowers, as well as a comprehensive gene prediction strategy encompassing multiple sources of evidence, including RNA-Seq libraries from a wide range of tissues. The differentially expressed geneset contains a large proportion of genes of potential interest as downstream targets, with a number of overrepresented GO terms that might relate to the cell division and elongation processes affecting the corolla tube and style of *P. vulgaris* plants in order to bring about the respective morphologies of the distinct heterostylous floral forms (Webster and Gilmartin, 2006). In addition, a large array of genes distributed throughout the genome is apparently regulated in response to the *S* locus, which raises the possibility that the SI determining genes might not reside at the *S* locus itself (Barrett, 2008).

Intriguingly, the differential expression analysis reveals more than double the number of upregulated genes in thrum as compared to pin. The thrum is characterised by large pollen, and increased cell division below the point of anther attachment. In addition, cells are widened above the point of anther attachment, which presumably acts to maintain the precision of pollinator interactions by counteracting the increased number of cells below the point of anther attachment; in pin flowers there is no such contrast in cell size (Webster and Gilmartin, 2006). It could be that the specific morphological features of the thrum are more intricate and complex to attain. This may involve the downregulation of some genes, but perhaps high levels of expression for the downstream genes is ultimately required to drive the increased size of the pollen and corolla tube cells, with a greater number of genes being required to manipulate the differential cell sizes above and below the point of anther attachment; additional developmental pathways may be upregulated by the *S* locus to generate the specific morphology of the thrum flowers.

Genes with morph-specific expression seem to be of largely limited importance in the regulatory pathway downstream of the *S* locus; there is no associated GO term

overrepresentation, and the majority of associated functional annotations reveal no immediately apparent role in cell division and elongation processes that act to bring about the distinct heterostylous floral morphologies (Webster and Gilmartin, 2006). Though it may prove that a small proportion of these morph-specific genes are important in this pathway, perhaps it is not surprising that the majority of the determining genes appear to be differentially expressed rather than subject to morph-specific silencing; precise adjustments in expression rather than complete downregulation of genes in the specific pathways leading to the fine-scale tuning of reciprocal organ height in the two forms of *Primula* flower could be seen as in keeping with the intricate developmental patterns that might be associated with such a system, as well as the precise nature of insect-mediated pollen transfer (Cohen, 2010, de Vos et al., 2014).

In summary, the *P. vulgaris* long-homostyle assembly and associated genomic analyses represent a significant improvement over existing resources for the Primulaceae, as well as highlighting this genome as a potential platform for the identification and assembly of the *Primula S* locus. These results exemplify the necessity for good starting material in the assembly, annotation and analysis of a non-model plant species; without such material it would not have been possible to assemble a good quality genome sequence without significant further funding. The current study describes a comprehensive set of genomic data that will facilitate the drive towards identifying the regions underpinning heterostyly in diverse families across the angiosperms.

3

Annotation and characterisation of *S*-linked genes

3.1 Relevant publications

Cocker, J. M.*, Webster, M. A.*, Li, J., Wright, J., Kaithakottil, G., Swarbreck, D., Gilmartin, P. M. (2015) *Oakleaf*: an *S* locus-linked mutation of *Primula vulgaris* that affects leaf and flower development. *New Phytologist,* 208: 149–161.

Li, J., Webster, M. A., Wright, J., **Cocker, J. M.,** Smith, M. C., Badakshi, F., Heslop-Harrison, P., Gilmartin, P. M. (2015) Integration of genetic and physical maps of the *Primula vulgaris S* locus and localization by chromosome *in situ* hybridization. *New Phytologist*, 208: 137–148.

* These authors contributed equally

3.2 Introduction

Primula vulgaris plants have one of two forms of flower, pin or thrum. The *S* locus, which controls the development of these floral morphs, comprises a tightly linked cluster of at least three genes (Lewis and Jones, 1992); the dominant alleles of these genes co-segregate with the thrum phenotype, which presents anthers with large pollen at the mouth of the corolla tube and the stigma midway down. The pin has a reciprocal floral morphology, with the stigma at the mouth of the flower and the anthers halfway down the floral tube. This arrangement of reproductive structures serves to physically promote insect-mediated outcrossing between plants with the reciprocal forms of flower. The architectural constraints imposed by the complementary positioning of the

male and female reproductive structures is reinforced by a putatively sporophytic self-incompatibility (SI) system, for which the determining genes are thought to be linked to or under the control of the *S* locus and its constituent genes (Lewis, 1949, Dowrick, 1956de Nettancourt, 1997, Lewis and Jones, 1992). There is typically a 1:1 ratio of pin to thrum plants in field populations based on equally frequent pin and thrum genotypes (Ornduff, 1979).

It can be easily observed whether a plant has flowers of one form or the other, as such it is straightforward to determine whether a phenotypic character is linked to the *S* locus based on co-segregation of the character with the pin or thrum phenotype. As a result, over the many years of study since Darwin first explained the importance of the system in promoting outcrossing (Darwin, 1877), several genes have been identified as linked to the *S* locus in *Primula sinensis* and *P. vulgaris*, including flower pigment genes (De Winton and Haldane, 1933, Kurian, 1996), *Hose in Hose* (Fig. 3.1) (Ernst, 1942, Webster and Grant, 1990, Webster and Gilmartin, 2003, Webster, 2005, Li et al., 2010) and *sepaloid* (Webster and Gilmartin, 2003, Webster, 2005, Li et al., 2008).

Figure 3.1 *Hose in hose* wood block print (van de Passe, 1614) showing characteristic conversion of sepals to petals; arrow indicates flower with normal sepals that has reverted to wildtype. Figure reproduced from Li et al. (2010).

The above genes, as well as an array of *S* locus-linked genes and markers identified by random amplified polymorphic DNA (RAPD) and fluorescent differential display analyses (Manfield et al., 2005, Li et al., 2008) facilitated the formulation of a genetic map flanking the *Primula vulgaris S* locus (Figure 3.2) (Li et al., 2011, Li et al., 2015). In a drive towards identifying the key genes responsible for the development of the reciprocal floral architectures, the linked genes and markers were used as probes in the identification of founding BACs (Bacterial Artificial Chromosomes) in a BAC-walking strategy that resulted in BACs spanning the *S* locus (Li et al., 2011, Li et al., 2015). In one of the crosses used to generate the genetic map, a self-fertile short homostyle was produced, the sequencing and assembly of which was described in Chapter 2 of this thesis. This was surprising as homostyles are notably rare, with Richards (1997) estimating that they might occur at frequencies of less than 1% in wild *P. vulgaris* populations. Darwin (1877) noted them, as did Bodmer (1960) in Somerset, and Crosby (1940) in the Chilterns, but these are long homostyle populations.

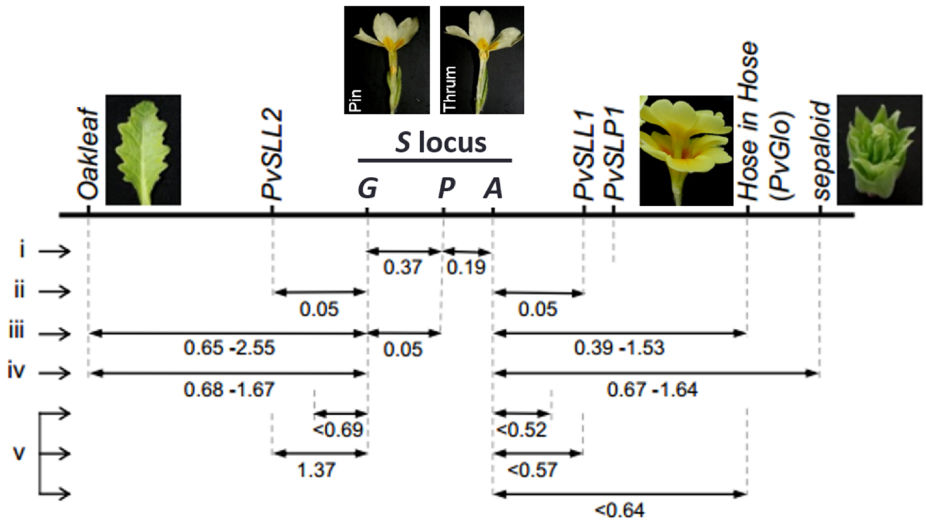


Figure 3.2 Genetic map of the *Primula S* locus; mapping distances (cM) for *S* locus-linked genes and markers are shown using (i) data from Lewis and Jones (1992) for the predicted distance between *S* locus constituent genes (*GPA*), (ii-v) data from analyses in Li et al. (2015a). Figure reproduced from Li et al. (2015a).

Following the rediscovery of Mendel's work on pea plants (e.g. Bateson, 1902), Bateson and Gregory's studies in defining the dominance relationship of pin and thrum

flowers (Bateson and Gregory, 1905) led to a series of studies which mark heterostyly in *Primula* as one of first genetic systems for which genetically-linked phenotypes were identified (Gregory, 1911, Bridges, 1914, Altenburg, 1916); these phenotypes include *Hose in Hose* (Ernst, 1942) and four loci in *P. sinensis*: magenta (*b*), red stigma (*g*), red leaf back (*l*) and double (*x*) (Gregory, 1911, De Winton, 1928); genes responsible for sporophytic SI are also linked (Lewis, 1949, de Nettancourt, 1997, Lewis and Jones, 1992). De Winton and Haldane (1935) constructed the first genetic map for a distylous species. However, the identification of floral variants in *Primula* dates back over 400 years (van de Passe, 1614). This is perhaps owing to the primrose's status as one of the most popular garden plants in the world, known in European gardens since the time of medieval herbalists (Mast et al., 2001, Richards and Edwards, 2003). It seems quite remarkable that *Hose in Hose*, shown illustrated by van de Passe in 1614 (van de Passe, 1614) (Figure 3.1) was subsequently identified as *S* locus-linked (Ernst, 1942, Webster and Grant, 1990), although perhaps not so surprising in the eyes of Darlington (1931), who noted that there were a great number of *S*-linked loci and used this as a basis to suggest that the *S* locus might lie on the largest chromosome, as confirmed in analyses associated with the current study (Li et al., 2015); *in situ* hybridisation analyses reveal that the *P. vulgaris S* locus is located close to the centromere of the largest metacentric chromosome, with *PvGLO* orientated proximal to the centromere. This confirms previous evidence of linkage to the centromere based on "double reduction" in autotetraploid plants, and is in line with suggestions of recombination suppression at the *S* locus to maintain the distinct pin and thrum genotypes (Darlington, 1929, Ornduff, 1992).

In addition to the *Hose in Hose* and *sepaloid* mutants that were established as *S* locus-linked in more contemporary studies (Ernst, 1942, Li et al., 2008), a third *S* locus-linked developmental phenotype exists in *Primula vulgaris* that produces lobed leaves and distinctive attenuated petals with variable shape and character, possibly due to differential expression of the locus in different genetic backgrounds (Figure 3.3). This mutant phenotype was named *Oakleaf* and is revealed as a single dominant locus through the observation of the expected ratio of progeny in crosses between *Oakleaf* and wildtype plants, supported by chi-squared analysis (Cocker et al., 2015). The phenotype is sometimes visible in seedlings as lobed cotyledons, with the first true leaves consistently showing a lobed phenotype that is similar in appearance to that of *Quercus*

species, such as the oak tree. Ectopic meristems, typically vegetative, can emerge from the leaves of *Oakleaf* plants. In addition, an increase in the separation and size of sepals can sometimes be observed, but the sepals are not lobed. The floral phenotype ranges from petals similar to wild type with splits in the corolla, to attenuated and separated petals; the most severe of which presents straight and narrow petals that look like the spokes of a wheel (Cocker et al., 2015). In combination with mutants that result in a disruption of floral-organ identity, the attenuation of petals or lobed appearance of *Oakleaf* leaves are observable in whorls that are converted to petals or leaves (Cocker et al., 2015). For example, in combination with *Hose in Hose*, where the first whorl is converted from sepals to petals, the first whorl petals show attenuation; thus, the effect of *Oakleaf* is organ-specific rather than whorl-specific.

Figure 3.3 *Oakleaf* developmental phenotypes (bars = 1 cm), (a) seedlings with wild-type and *Oakleaf* phenotypes (arrows), (b) leaf from *Oakleaf* plant, (c) *Oakleaf* mutant with distinctive lobed leaves and attenuated petals, (d) extreme attenuated petals phenotype, (e) partially attenuated petals, (f) *Oakleaf* plant with near normal petals, (g) *Oakleaf* leaf with leaves developing from ectopic meristem (arrow), (h) ectopic flower (arrow), (i) ectopic flower in (h) with developing seed capsule following pollination (arrow). Figure reproduced from Cocker et al. (2015).

The above features (lobed leaves with occasional ectopic meristems) are highly similar to the phenotype of *Arabidopsis thaliana* with ectopically expressed Class I *KNOX* genes (Lincoln et al., 1994, Chuck et al., 1996, Hay and Tsiantis, 2010), and other species including tobacco (Hareven et al., 1996), which like *P. vulgaris* resides in the asterids. This suggests *KNOX*-like genes in *P. vulgaris* are a good starting place for identifying the cause of *Oakleaf*.

Gregory (1911) described a *P. sinensis* mutation named "o" that also affects the flowers and causes oak-shaped leaves, but it is distinct from *Oakleaf* in that it was found to be recessive and is not linked to the *S* locus. The *Oakleaf* phenotype was identified in 1999

amongst commercial ornamental *Primula* plants and was developed in polyanthus form as a commercial variety by Margaret Webster, with a division of the original mutant plant being used to establish the *Oakleaf* population used in the current study (Cocker et al., 2015).

This chapter describes transcriptomic and genomic analyses towards the characterization of the new *S* locus-linked mutant named *Oakleaf*, as well as the *ab initio* annotation of genes in the BAC region spanning the *S* locus, in an effort to orientate and validate the region as a prelude to the identification of the genes controlling the development of heterostyly.

3.3 Methods

3.3.1 Plant material

Plants used in this study are wild-type *Primula vulgaris* Huds. and derived commercial cultivars. *Primula vulgaris Oakleaf* plants were originally obtained from Richards Brumpton (Woodborough Nurseries, Nottingham, UK) in 1999 and maintained by Margaret Webster as part of the National Collection of *Primula*, British Floral Variants. Plants were grown as described previously by Margaret Webster and JL (Webster & Gilmartin, 2006).

3.3.2 Differential gene expression between *Oakleaf* and wild-type

RNA was isolated from leaves and open flowers of *Oakleaf* and wild-type pin plants, as well as mixed stage pin and thrum flowers for RNA-Seq using Illumina HiSeq2000 (Table 3.1). RNA-Seq reads were aligned to LH_v1 contigs using TopHat (v2.0.8) (http://ccb.jhu.edu/software/tophat/index.shtml), followed by construction and merging of the transcriptome using Cufflinks (v2.1.1) (Trapnell et al., 2013) (http://cole-trapnell-lab.github.io/cufflinks/). RNA-Seq reads from mixed stage pin and thrum flowers were used for transcriptome assembly but not subsequent expression analysis. HTSeq (Anders et al., 2014) was used to count raw read numbers per gene with RNA-Seq data from *Oakleaf* and wild-type leaf and flower samples. DESeq (v1.16.0) was used to

normalise these read counts by estimating the effective library size (Anders & Huber, 2010) and to carry out differential expression analysis. Genes upregulated by a ×2 log2 fold-change in both *Oakleaf* leaves and *Oakleaf* flowers were characterised by BLASTX analysis (e-value $1 \times 10{-}4$) (Camacho et al., 2009) to identify related sequences in the TAIR10 (https://arabidopsis.org/) and NCBI "nr" protein databases, the latter alignments being used as an input for Blast2GO (Conesa et al., 2005). Sequences were submitted to NCBI under Bioproject number PRJNA260472.

Library	**Description**	**Type**	**Insert size**	**Read count**
LIB668	Pin mixed flower buds	RNA	223	100768798
LIB669	Thrum mixed flower buds	RNA	180	81045610
LIB976	*Oakleaf* flower	RNA	220	24995179
LIB977	Pin mature open flower	RNA	209	33400153
LIB978	*Oakleaf* leaf	RNA	219	14310589
LIB979	Pin leaf	RNA	215	45723021

Table 3.1 RNA-Seq libraries used for transcriptome assembly and differential expression analyses

3.3.3 Gene model predictions for *P. vulgaris KNOX* (*PvKNOX*) genes

Arabidopsis thaliana KNOX proteins, KNAT1, KNAT2 (Lincoln et al., 1994), KNAT3, KNAT4, KNAT5 (Serikawa et al., 1996), KNAT6 (Belles-Boix et al., 2006), KNAT7 (Li et al., 2011), and STM1 (Long et al., 1996) were aligned to the LH_v1 *Primula vulgaris* genome assembly (see Chapter 2 for assembly details) with Exonerate (v2.2.0) (Slater & Birney, 2005). *Primula vulgaris KNOX* loci were identified and the gene models confirmed by transcript evidence from TopHat (v2.0.8) and Cufflinks (v2.1.1) (Trapnell et al., 2013) and by homology of the predicted proteins to KNOX proteins from the TAIR10 protein database (https://www.arabidopsis.org/). Parameters for protein sequence comparisons were ≥ 50% identity with ≥ 30% coverage of the KNOX query sequence. Gene models were curated manually where necessary with

GenomeView to resolve the intron and exons sizes of all eight predicted genes (http://genomeview.org/). *PvKNL1* was initially identified as two sequences on separate genomic contigs, but was resolved as one locus by alignment to a single transcript sequence spanning the boundary between the two sequences; this transcript was present in a *de novo* transcriptome assembly generated with Trinity (Grabherr et al., 2011) by TGAC using RNA-Seq data from *P. vulgaris* pin and thrum mixed stage flower buds, as used in the Cufflinks assembly.

3.3.4 Generation of the PvKNOX phylogenetic tree

Multiple sequence alignment of *Zea mays* KNOTTED1, *A. thaliana* KNOX proteins, and predicted PvKNOX protein sequences was carried out in MEGA6 using MUSCLE (Edgar, 2004, Tamura et al., 2013). To obtain phylogeny support, Bayesian analyses were performed using MrBayes (v3.2.2) (Ronquist et al., 2012) and output files visualised in FigTree (v1.4.0) (http://tree.bio.ed.ac.uk/software/figtree/). The mixed amino acid substitution model was used, and the first 25% of samples discarded as burn-in. The consensus tree was obtained after 1,000,000 generations, with the average standard deviation of split frequencies below 0.01 to ensure convergence.

3.3.5 Identification of variant sites between *Oakleaf* and wild-type

BAM files produced with TopHat (in the above transcriptome assembly) were used as an input for the SAMtools (v0.1.18) (Li et al., 2009a) "mpileup" tool for the identification of variant sites between *Oakleaf* and LH_v1, and the wild-type pin plant and LH_v1, for both leaves and open flowers. The positions of the *KNOX*-loci based on the curated *PvKNOX* gene models predicted above were used to determine the resulting amino acid substitutions that would result from the identified SNPs. The predicted impact on protein function of each amino acid substitution identified above was determined by PMG using the SIFT prediction tool (Ng and Henikoff, 2003b).

3.3.6 Linkage analysis of *PvKNOX* candidate genes

Four libraries of genomic paired-end sequencing reads for individual pin and thrum plants (LIB1732 and LIB1167; Table 2.1), and separate pools of 24 pin and 28 thrum progeny resulting from a cross between the individual pin and thrum plants (LIB1730

and LIB1731; Table 2.1), were mapped to the LH_v2 genome assembly using BWA v0.6.2 (Li and Durbin, 2009). This analysis was carried out at a later stage to the analyses described above, hence the use of the LH_v2 assembly. SAMtools "rmdup" was used to reduce amplification biases (PCR duplicates) for the mapped read libraries (Li and Durbin, 2009).

SAMtools (v0.1.18) (Li et al., 2009a) was used to call SNPs between the four read libraries and the LH_v2 contigs; sites were excluded with mapping quality (MQ) < 20 and depth (DP) < 10 in the Variant Call Format (VCF) files for each of the four read libraries; sites with genotype quality (GQ) < 30 in the thrum or pin parent were also omitted. *PvKNOX* genes were mapped to the *P. vulgaris* long-homostyle LH_v2 genome assembly using Exonerate (v2.2.0) (Slater and Birney, 2005), and SNPs extracted for the contigs to which they aligned. Sites that were heterozygous in the thrum parent and homozygous for an alternate (non-reference) allele in the pin parent were analysed to ascertain the ratio of reference to alternate alleles in the progeny, and thus calculate the mean genetic distance from the *S* locus in centimorgans (cM) for each contig based on the number of recombinants represented by the minor (alternate) allele frequency at each site. The distance used for each site was the lower of the two putative distances calculated for either the pin or thrum progeny, assuming one must be homozygous and one heterozygous if the contig is linked. Sites were only included where the reference allele frequency was lower than the alternate ( ) allele frequency in either pin or thrum progeny, whichever was selected based on a smaller associated genetic distance. Exonerate (v2.2.0) (Slater and Birney, 2005) was used to map *PvKNOX* genes to the LH_v2 contigs, and genetic distances for the associated contigs extracted.

3.3.7 BAC contig assembly

BAC library construction, screening and sequencing was carried out by JL, and is documented in Li et al. (2011 and 2015). The contig assemblies of BAC sequences from the resulting 454 and Illumina reads were generated by JW using gsAssembler (v2.6) and ABySS (v1.3.6) respectively (Simpson, 2009). BAC contig assembly to combine the small NGS-derived contigs above was carried out by JW using minimus2 (Sommer et al., 2007) to merge overlapping BAC sequences and BAC-end sequences (with

sequence identity > 98%); this was facilitated by the inclusion of contigs from the draft thrum genome assembly TP_v2 (see Chapter 2), and further merging of contigs based on regions overlapping > 500 bp using BLAT v3.5 (Kent, 2002).

3.3.8 Repeat masking of the BAC contig assembly

RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) was used to identify *de novo* repetitive sequences in the draft thrum genome assembly TP_v2; the resulting sequences were hard-marked for inclusions from protein coding genes using BLASTX (v2.2.28) (Camacho et al., 2009) alignments (e-value $1x10^{-4}$) to protein databases from *Actinidia chinensis* (http://bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi), *Mimulus guttatus* (v2.0), *Solanum tuberosum* (v3.4) and *Solanum lycopersicum* (v2.4) (http://phytozome.jgi.doe.gov/).

The repeat library of *de novo* repetitive sequences comprised all hard-masked sequences with at least one alignment to a transposition-associated domain from Pfam-A (curated thresholds) or Pfam-B (evalue $1x10^{-4}$); alignments were carried out using HMMer hmmscan (v3.1b1) (http://pfam.sanger.ac.uk/; http://hmmer.janelia.org/search/hmmscan). Pfam domains were considered transposition-associated if they aligned with any of the sequences contained in the database of transposable elements included in the RepeatRunner package (http://www.yandelllab.org/software/repeatrunner.html).

Repeats and low complexity regions were identified in the BAC contig assembly using the repeat library with a local installation of RepeatMasker based on the RMBlast algorithm (version open-4.0.1: http://www.repeatmasker.org/).

3.3.9 Prediction of genes in the BAC contig assembly

The self-training gene annotation program GeneMark-ES (http://opal.biology.gatech.edu/) was used with the draft thrum genome assembly to produce a training file which served as an input for the de novo gene finder GeneMark-E in the annotation of the repeat-masked BAC assembly. The gene models obtained for each contig were scanned with the software package Full-LengtherNEXT (e-value $1x10^{-}$

[4]) (http://www.scbi.uma.es/site/scbi/downloads/313-full-lengthernext) in order to classify them as full- length, 5'-end, 3'-end or internal.

3.3.10 Functional annotation of the BAC contig

BLASTX (e-value $1x10^{-4}$) (Camacho et al., 2009) was used to query the putative genes against the TAIR10 *Arabidopsis thaliana* protein database (http://www.arabidopsis.org/) and the NCBI non-redundant protein database, the result of the latter being an input for further annotation with Blast2GO (https://www.blast2go.com/) (Conesa et al., 2005) (see Table S2 (Li et al., 2015)).

3.4 Results

3.4.1 Prediction of *Primula vulgaris KNOX*-like (*PvKNL*) gene models

Oakleaf plants present lobed leaves and sometimes produce ectopic meristems on the veins of the leaves that can be floral or vegetative. *Oakleaf* is dominant to wildtype. These features have similarities to the role of Class I *KNOX* homeodomain genes in tomato and *Cardamine hirsuta* (Hareven et al., 1996, Bharathan et al., 2002, Hay and Tsiantis, 2006, Shani et al., 2009) and are particularly reminiscent of ectopically expressed Class I *KNOX* genes in *Arabidopsis thaliana* (Lincoln et al., 1994, Chuck et al., 1996, Hay and Tsiantis, 2010). We therefore hypothesized that *Oakleaf* might result from mutation of a *KNOX*-like gene in *P. vulgaris*.

On this basis, the current study proceeded with an exploration of *KNOX*-like genes in the *Primula vulgaris* genome to consider the following possibilities: (i) that the *Oakleaf* phenotype is the result of constitutive overexpression (as in Arabidopsis) of a *KNOX*-like homeodomain gene in mature leaves and flowers, (ii) that expression is unchanged but a mutation results in a dominant gain of function in protein activity, (iii) that the phenotype is the result of upregulation of a gene with no similarity to *A. thaliana KNOX* family genes.

The first reference-based transcriptome for *Primula vulgaris* was assembled prior to LH_v2 gene prediction (Chapter 2). This assembly comprised RNA-Seq datasets from

leaves and flowers of *Primula vulgaris* wild-type and *Oakleaf* plants that were used for subsequent expression analyses, as well as additional RNA-Seq reads from pin and thrum mixed stage flower samples, to maximise coverage of the transcribed regions of the genome (Table 3.1). The draft *Primula vulgaris* LH_v1 genome assembly available at the start of the current study was assembled prior to the final LH_v2 assembly (Chapter 2) and was used as a reference assembly for the above transcriptome (Table 2.2).

The seven *Arabidopsis thaliana KNAT* and *STM* coding sequences were aligned to the *P. vulgaris* genome with Exonerate (v2.2.0) (Slater and Birney, 2005) using evidence from transcripts in the Cufflinks-generated (Trapnell *et al.,* 2013) transcriptome (above) to define the *Primula vulgaris KNOX*-like gene models. This analysis revealed five Class I and three Class II *KNOX* genes in the *Primula vulgaris* (LH_v1) genome (Figure 3.4). Class I *KNOX* family genes have roles in shoot apical meristem identity (Hay and Tsiantis, 2010); this is consistent with the ectopic meristems that are sometimes produced on the veins of *Oakleaf* leaves (Cocker et al., 2015). Class II *KNOX* family genes *KNAT3*, *KNAT4* and *KNAT5* are involved in root development (Truernit et al., 2006) and *KNAT7* in the formation of secondary cell walls (Li et al., 2011a, Li et al., 2012). If these classifications hold true in *P. vulgaris*, then this suggests genes showing homology to Class I *KNOX* genes might be the best candidates for the *Oakleaf* phenotype.

PvKNL1 was originally identified on two separate contigs, supported by two transcript sequences, one comprising three exons from the 5' end of the *KNOX*-like gene and the other two exons from the 3' homeodomain region, potentially forming part of the same *KNOX*-like gene; this was confirmed using a *de novo* Trinity (Grabherr *et al*., 2011) transcriptome assembly containing a single transcript spanning the boundary between the two exon sequences.

RNA-Seq reads or transcriptome data can be used to resolve gaps between contigs as it captures information about the connections between exons in a single gene (Zhang et al., 2016). Introns that are absent from transcribed sequences represented by RNA-Seq reads may be sufficiently repetitive to complicate the process of assembling genomic sequencing data (Wang et al., 2008); this can lead to fragmented genes comprising predictions in multiple contigs that can be connected by scaffolding approaches that

leverage the information provided by exons anchored to distinct genomic fragments; as a result, evidence such as that present in the above *de novo* transcriptome data is sufficient to join the two contigs in the *Primula vulgaris* genome assembly, with a gap of unknown size between contigs and the fragmented intron sequence therein. The *P. vulgaris* genome thus contains eight expressed *PvKNOX* genes, with the predicted gene structures shown in Figure 3.4.

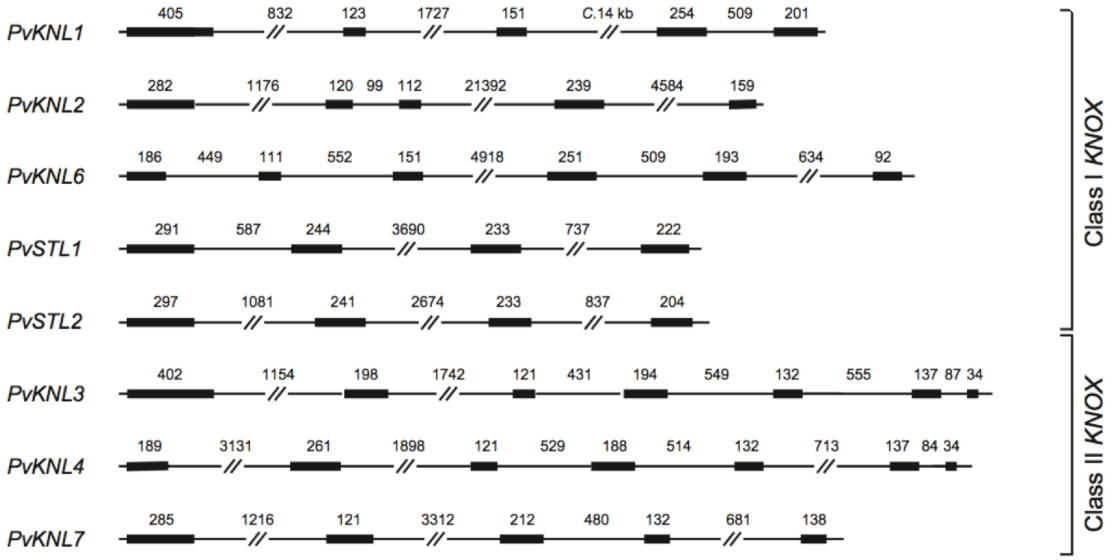


Figure 3.4 *Primula vulgaris Knotted*-like genes grouped based on similarity to Class I and Class II *A. thaliana KNOX* genes: predicted gene structures of the eight *Primula vulgaris Knotted*-like (*PvKNL*) and *Shoot meristemless*-like (*PvSTL*) gene models that show amino acid similarity to *Arabidopsis thaliana KNAT* and *STM KNOTTED*-like homeobox (*KNOX*) gene family members (thick lines = exons, thin lines = introns. Figure reproduced from Cocker et al. (2015).

3.4.2 Characterisation of the *PvKNOX* gene family

The *Oakleaf* phenotype is similar to that displayed by *Arabidopsis thaliana* plants where overexpressed *KNAT1* induces lobed leaves with ectopic meristems (Lincoln, 1994). The *Knotted* mutant, after which *Knotted*-like homeobox (*KNOX*) genes are named, was first discovered in maize. *KNAT1* and *KNAT2* were identified through the use of the maize *Knotted-1 (Kn1)* homeobox as a heterologous probe (Lincoln, 1994), whilst further low stringency screening using these two sequences identified further

KNAT genes in *Arabidopsis* (Long et al., 1996, Serikawa et al., 1996, Belles-Boix et al., 2006, Li et al., 2011a). The phylogenetic tree (Figure 3.5) was generated using the eight PvKNOX-family protein sequences, as well as the amino acid sequences for *A. thaliana* KNAT (Long et al., 1996, Serikawa et al., 1996, Belles-Boix et al., 2006, Li et al., 2011a), STM (Long *et al.,* 1996), and the *Zea Mays* KNOTTED-1 protein that was originally used to design the heterologous probe (based on the coding sequence), as described above. This analysis facilitated naming of the *PvKNOX* genes as *Knotted*-like (*PvKNL*) and *Shoot meristemless*-like (*PvSTL*) based on protein sequence similarity, and facilitated grouping of these genes as Class I and Class II, as defined for *KNOX* homeodomain gene families in Maize (Vollbrecht et al., 1991), *Arabidopsis* (Lincoln et al., 1994, Long et al., 1996, Serikawa et al., 1996, Belles-Boix et al., 2006, Li et al., 2011) and other species (Bharathan et al., 1999, Hay and Tsiantis, 2010) based on their expression and phylogenetic relationships (Kerstetter et al., 1994, Bharathan et al., 1999). Further low-stringency alignments with Exonerate (v2.2.0) confirm that *P. vulgaris* has no homologue of *AtKNAT5*, but it does have two *AtSTM*-like genes, resulting in five Class I and three Class II *P. vulgaris KNOX*-like genes in total.

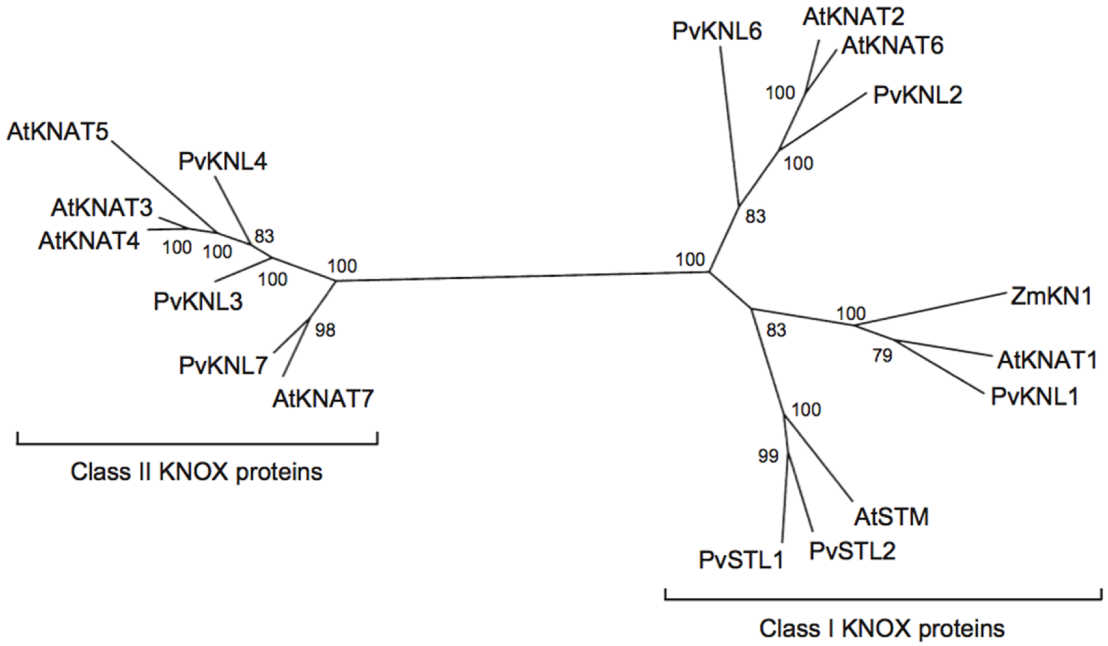


Figure 3.5 Bayesian strict-click phylogenetic tree based on multiple sequence alignments of *Zea Mays* KNOTTED1, *Arabidopsis thaliana* KNAT and STM, and the eight *Primula vulgaris* PvKNL and PvSTL amino acid sequences. Posterior probabilities for clades are shown as percentages. Figure reproduced from Cocker et al. (2015).

3.4.3 Expression of *PvKNOX* genes in *Oakleaf* and wild-type *P. vulgaris*

The normalised read counts for RNA-Seq reads mapped to the nine *PvKNOX*-family genes facilitated a direct comparison between expression levels in leaves and flowers of *Oakleaf* and wildtype plants (Figure 3.6). The criterion of constitutive over-expression of a *PvKNOX* gene identified by upregulation in both leaves and flowers of *Oakleaf* plants compared to wild-type was used to determine whether the *Oakleaf* phenotype is caused by overexpression of Class I *KNOX* genes, as is the case for the similar *knotted* phenotype in *A. thaliana* (Lincoln et al., 1994, Chuck et al., 1996, Hay and Tsiantis, 2010).

For RNA-Seq data analysis, an estimate of the false discovery rate (FDR) is often used to correct for multiple testing errors when considering thousands of genes simultaneously; the more tests performed, the more likely some of those tests will be significant just by chance (Bi and Liu, 2016). In practice, the true FDR of a new dataset is unknown (Li, 2012): with a low number of replicates there is a failure to control FDR under the model assumed by differential expression tools due inaccurate calculation of the uncorrected p-values (Schurch et al., 2016). It is therefore prudent to carry out RNA-Seq analyses with biological replicates for each condition to unambiguously identify differentially expressed genes. If the biological variance within a condition is unknown, some of the estimates for fold change will be imprecise. In contrast, parallel measurements of biologically distinct samples capture random biological variation in a design that incorporates RNA-Seq replicates of biological conditions (Blainey et al., 2014).

In the current study, despite lack of replicates for each condition, it was reasoned that upregulation in both *Oakleaf* leaves and flowers as an indication of constitutive upregulation was sufficient to counteract the possibility of false positives in differential expression if either of these measures were to be used in isolation; the above observations are noted as a possible caveat.

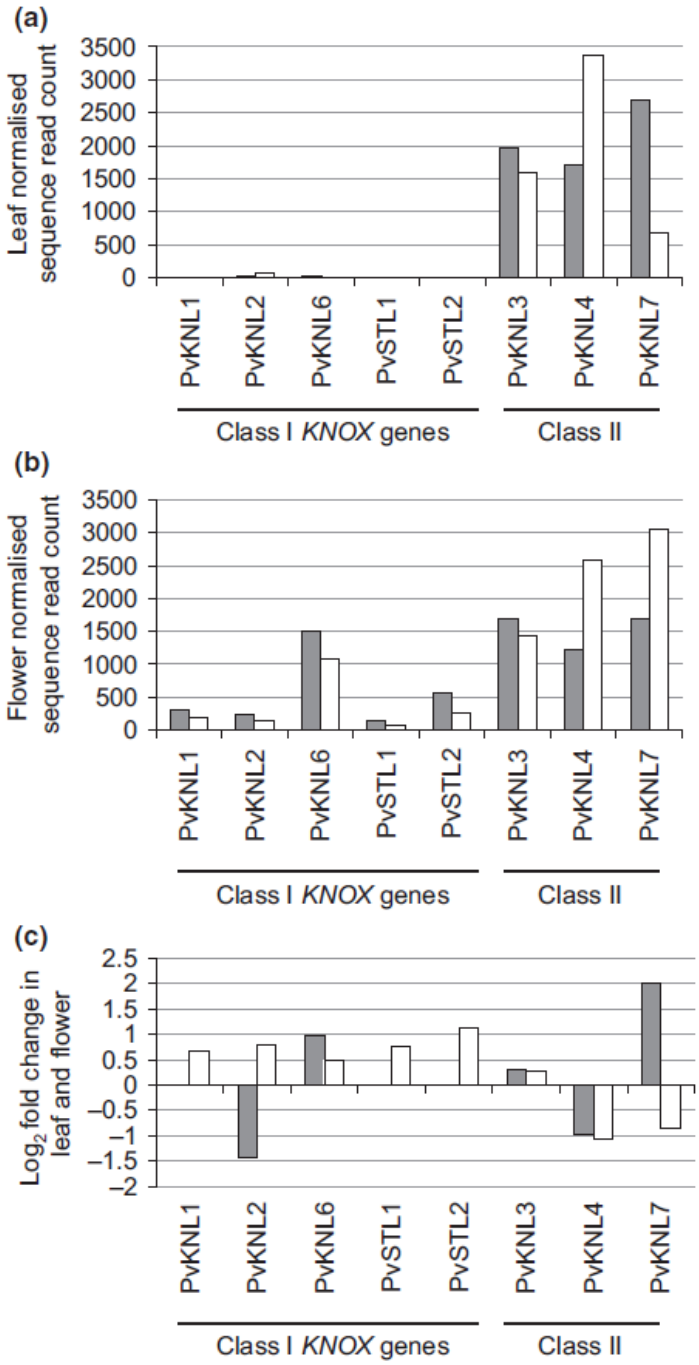


Figure 3.6 The normalised counts for RNA-Seq reads mapping to *PvKNOX* family genes for RNA isolated from *Oakleaf* leaves (grey bars) and wild-type leaves (white bars) (a); *Oakleaf* flowers (grey bars) and wild-type flowers (white bars) (b); and the $\log_2$ fold change in expression between *Oakleaf* and wild-type leaves (grey bars) and *Oakleaf* and wild-type flowers (white bars) (c). Figure reproduced from Cocker et al. (2015).

The low expression of Class I *PvKNOX* genes in the leaves of *Oakleaf* and wildtype *P. vulgaris* (Figure 3.6) suggests that these genes are not responsible for the *Oakleaf* phenotype; *PvKNL6* is upregulated in both leaves and flowers of *Oakleaf* but its low expression in *Oakleaf* leaves is not suggestive of constitutive upregulation. In contrast, the Class II *PvKNOX* genes show high expression in *P. vulgaris* leaves and flowers; *PvKNL3* shows minimal $\log_2$ fold-upregulation in both *Oakleaf* leaves (0.31) and flowers (0.26) as compared to wildtype. These data suggest that none of the *PvKNOX* family genes are promising candidates for *Oakleaf* based on an expectation of constitutive expression, as they do not show a high level of expression and upregulation in both *Oakleaf* leaves and flowers.

3.4.4 Sequence comparison of *PvKNOX* genes in wild-type and *Oakleaf*

SAMtools (v0.1.18) (Li et al., 2009a) was used to identify single nucleotide polymorphisms (SNPs) between *Oakleaf* RNA-Seq reads and the *P. vulgaris* LH_v1 genome assembly (Table 3.2). The *Oakleaf* plant that was used is heterozygous at the *Oakleaf* locus in a pin genetic background, SNPs that are heterozygous in *Oakleaf* and absent in the pin RNA-Seq reads are therefore potentially *Oakleaf* specific.

This analysis identified 10 *Oakleaf*-specific SNPs between seven of the *PvKNOX* genes in *Oakleaf* and wildtype: three are predicted to cause truncated proteins (in *PvKNL2* and *PvSTL1*) (Table 3.2), with the remaining seven analysed with SIFT to predict the potential impact on protein function (Ng and Henikoff, 2003). Of these, five SNPs were predicted to result in tolerated amino acid substitutions, with the remaining two, Leu^{335}-Ser (*PvKNL3*) and Gly^{6}-Glu (*PvKNL7*), predicted to result in non-tolerated amino acid substitutions that could affect protein function. In addition, profiles of RNA-Seq read coverage for *Oakleaf* and wildtype leaves and flower were analysed using GenomeView (http://genomeview.org/); no discernible difference was observed between *Oakleaf* and wildtype read profiles, suggesting the *PvKNOX* genes in *Oakleaf* do not have alternate splicing profiles that could result in a protein with dominant function due to lack of a critical regulatory domain for example.

Gene	**Position**	**Base Change**	**Amino acid change**	***Oakleaf* flower**		***Oakleaf* leaves**		**Comments**
				SNP reads	Total reads	SNP reads	Total reads	
PvKNL1	671	TCT - ACT	Ser35 – Thr	7	7	0	0	Homozygous
PvKNL1	1882	ACG - ATG	Thr161 - Met	3	3	0	0	Homozygous
PvKNL2	3534	GCT - GGT	Ala120 - Gly	3	3	No SNP	-	Homozygous
PvKNL2	25367	GGG - GCG	Gly233 - Ala	9	10	No SNP	-	Also found in pin
PvKNL2	30146	CAG - TAG	Gln298 - Stop	2	15	No SNP	-	Truncated protein
PvSTL1	8202	Frameshift - 2 bp	aa12: stop16	3	8	No SNP	-	Truncated protein
PvSTL1	8188	GAA - GCA	Glu17 – Ala	8	15	2	2	Substitution tolerated
PvSTL1	8068	Frameshift +2 bp	aa57: stop63	4	11	No SNP	-	Truncated protein
PvSTL1	2413	TCG - CCG	Ser271 - Pro	5	10	No SNP	-	Substitution tolerated
PvSTL1	2278	GTC - ATC	Val316 - Ile	8	13	No SNP	-	Substitution tolerated
PvSTL2	7432	GAA - CAA	Glu47 - Gln	37	37	0	0	Homozygous and in Pin
PvSTL2	7336	GCG - ACG	Ala79 - Thr	29	29	0	0	Homozygous and in Pin
PvSTL2	3225	ATT - ATG	Ile197 - Met	31	31	0	0	Homozygous and in Pin
PvKNL3	7545	CAG - CAT	Gln65 - His	67	98	50	63	Substitution tolerated
PvKNL3	2860	TTG - TCG	Leu335 - Ser	No SNP	No SNP	2	15	Substitution not tolerated
PvKNL3	2034	AGT - AGA	Ser395 - Arg	No SNP	No SNP	2	11	Also found in pin
PvKNL4	2083	AAC - AGC	Asn28 - Ser	17	69	23	66	Substitution tolerated
PvKNL7	8564	GGG - GAG	Gly6 - Glu	0	0	66	132	Substitution not tolerated

Table 3.2 Single nucleotide polymorphisms (SNPs) in *PvKNOX* family genes for *Oakleaf* RNA-Seq reads versus wildtype LH_v1 genome assembly. Position in genomic DNA (contig) is shown, with the nucleotide and associated change in amino acid indicated. Read counts for the total number of mapped reads and reads supporting the SNP are shown for *Oakleaf* leaves and flowers; where there is no SNP at a particular position, "No SNP" is shown. For predicted frameshifts, the amino acid at which the nucleotide change occurs and the new stop codon in the shifted reading frame is shown. The comments indicate whether the *Oakleaf* SNP is homozygous or heterozygous; where it is heterozygous and not present in pin, the impact of the SNP on the encoded protein is indicated; this was predicted (by PMG) using SIFT (Ng and Henikoff, 2003) (http://sift.bii.a-star.edu.sg/).

3.4.5 Differential expression between *Oakleaf* and wildtype *P. vulgaris*

The *P. vulgaris* draft transcriptome generated in this study contains 39,193 transcripts; normalised RNA-Seq read counts for each of the transcripts obtained using HTSeq and DESeq (Anders and Huber, 2010, Anders et al., 2014) were subjected to a log2 fold-change cut-off > 2 to reveal 507 genes potentially upregulated in both *Oakleaf* leaves and flowers versus wildtype (Figure 3.7). If there are no biological replicates, as in our analysis, then we do not know the within-group variance (Blainey et al., 2014): p-values can only be calculated with the use of pooled data from genes with similar expression strengths for the purpose of variance estimation (Anders and Huber, 2010), thereby reducing power; as a result, in this analysis we instead use a cut-off on the fold change in normalized expression to produce a set of differential expressed genes.

If the *Oakleaf* phenotype is a result of constitutive overexpression as predicted, then this geneset could contain a gene unrelated to *PvKNOX* that is responsible for the *Oakleaf* phenotype; none of the *PvKNOX* genes are in this set of differentially expressed genes as their fold-change in expression is below the two-fold cut-off. The *P. vulgaris* LH_v2 genome assembly and annotations were not available at the start of the current study, hence the use of a draft transcriptome with LH_v1 as a reference genome.

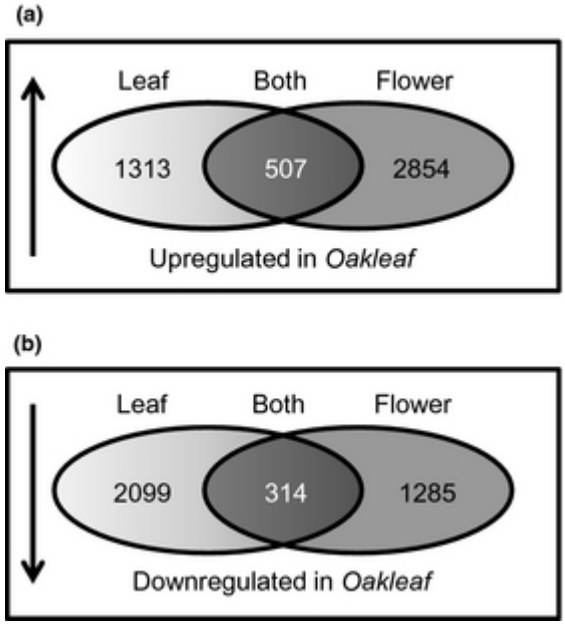


Figure 3.7 (a) Genes 2x log_2fold-upregulated upregulated and (b) 2x log_2fold-downregulated in *P. vulgaris Oakleaf* leaves and flowers versus wildtype *P. vulgaris* pin leaves and flowers. Genes that are up- or downregulated in both leaves and flowers of *Oakleaf* are indicated. Figure reproduced from Cocker et al. (2015).

KNOX proteins are transcriptional regulators, suggesting changes in their expression in *Oakleaf* would result in wider alterations in the expression of target genes in the downstream regulatory network. In the event that a *PvKNOX* gene is shown to be responsible for the *Oakleaf* phenotype, then the differentially expressed genes, including those that are both up- and downregulated in *Oakleaf* as compared to wildtype, are potential downstream targets that may be in the downstream regulatory network leading to the *Oakleaf* phenotype (Figure 3.7).

3.4.6 Annotation of the BAC contig assembly

The *de novo* annotation of BAC contigs flanking the *S* locus was carried out using GeneMark (http://exon.gatech.edu/GeneMark/). In total, 266 predicted genes or gene fragments were identified. BLASTX alignments identified matches in the NCBI "nr" and TAIR10 databases for 119 of the predicted protein sequences; after removal of potential duplicates indicated by gene models on the same contig or neighbouring

contigs that matched the same *Arabidopsis* gene, there were 82 *Arabidopsis* genes related to the predicted genes (Li et al., 2015).

The BAC assembly comprises BAC contigs positioned on the "left" and "right" of the *S* locus. In BAC *Contig S-right* (comprising BACs to the right of the *S* locus) *PvSLL1* was identified on S_locus_groupB_ctg13, *PvGLO* on S_locus_groupB_ctg58, and *PvSLP1* on S_locus_groupB_ctg36. *PvSLL2* was found on contig S_locus_groupA_ctg9 within *Contig S-left* (comprising BACs to the left of the *S* locus); see Li et al. (2015) for details. These results unequivocally confirm the order of the *S*-linked markers as *PvSLL2-S* locus*–PvSLL1–PvSLP1–PvGLO*.

3.4.7 *Oakleaf* in the BAC contig assembly

The *S* locus-linked marker *PvSLL2* is located 0.05 to 1.37 cM from the *S* locus within BAC *Contig S-left*, whilst crosses performed in Cocker et al. (2015) and Li et al. (2015a) give a map distance of between 0.56 and 2.55 cM for *Oakleaf* on the same side of the *S* locus as *PvSLL2*. Perhaps unsurprisingly then, none of the *PvKNOX* genes are located in BAC *Contig S-left* in which *PvSLL2* lies due to its relatively close proximity to the *S* locus. Furthermore, none of the genes upregulated in *Oakleaf* versus wildtype are in this contig. The functional annotation of predicted genes in this assembly does not reveal any obvious candidates for *Oakleaf.*

3.4.8 Linkage analysis of *Oakleaf* candidates

Four libraries of genomic paired-end sequencing reads for individual pin and thrum plants, and separate pools of pin and thrum progeny resulting from a cross between these individual pin and thrum parental plants, were mapped to the LH_v2 genome. SAMtools (v0.1.18) (Li et al., 2009a) was used to call SNPs between these libraries and LH_v2 contigs, and the results used to calculate the estimated genetic distance from the *S* locus based on the ratio of reference to alternate alleles for either the pin or thrum progeny; the lower genetic distance of the two at each position was incorporated into the mean distance tabulated (Table 3.3). The SAMtools SNP calling pipeline is designed for use with samples from diploid individuals (Li and Durbin, 2009, Raineri et al., 2012); sequenced pools of individuals offer a lower cost alternative to sequenced individuals but may be unsuitable for inferring linkage, it is difficult to distinguish between read

errors (0.1 – 1% for Illumina sequencing) and low-frequency alleles, based on each read being an independent draw from a large pool of chromosomes (Schlotterer et al., 2014). Individuals may be differentially represented due to unequal amounts DNA in the pool; a lack of multiple reads for each individual means sequencing errors are hard to resolve (Schlotterer et al., 2014). Furthermore, the issues above in combination with the low number of individuals present in the pools of 24 pin and 28 thrum progeny means that the calculated genetic distance may be imprecise, with discordant genetic distances estimated at each site; small pool sizes yield suboptimal results (Schlotterer et al., 2014).

In summary, the approach applied here offers only a crude estimate of the mapping distance. The depth (DP) and mapping quality (MQ) values were thus used to filter SNPs in the progeny pools in order to reduce coverage problems based on misaligned reads, the quality values associated with SAMtools (Li et al., 2009a) SNP calling or genotype quality (GQ) were not used, as they may be inappropriate based on the above. Nonetheless, ignoring the above scenarios and assuming an equal number of reads for each individual in the progeny pools, it was reasoned that the ratio of pin progeny alleles could provide a preliminary measure of the distance of each contig from the *S* locus. The estimated distance in centimorgans (cM) was determined for each contig by dividing the number of reference alleles at each site by the number of alternate alleles, then calculating the mean distance (±SE) across all sites (Table 3.3). The analysis indicates that seven of the eight *PvKNOX* genes either recombine very frequently with the *S* locus, or are otherwise situated on a different chromosome; there are no SNPs that pass the criteria described above for *PvKNL2*.

Gene	# SNPs	Distance (cM)	±SEM
PvKNL2	0	NA	NA
PvKNL1	35	31.92	1.63
PvKNL7	65	32.93	1.74
PvSTL2	12	24.42	1.42
PvSTL1	385	29.93	0.61
PvKNL3	18	27.86	2.80
PvKNL6	143	30.30	0.89
PvKNL4	443	41.22	0.86

Table 3.3 The mean genetic distance from the *S* locus (cM) is shown as predicted using SNPs present in pin and thrum read libraries mapped to LH_v2 contigs associated with the indicated *PvKNOX* genes. The standard error is indicated; problems with coverage and unresolved sequencing errors due to the use of pooled data (Schlotterer et al., 2014) may result in variability between the predicted distances at each site; the number of SNPs (# SNPs) is shown as the SEM may be artificially low in the event of a low number of SNPs being present in the contig.

In a preliminary analysis, the same strategy was used to test for linkage of contigs associated with the 507 genes upregulated in both *Oakleaf* leaves and flowers; using this approach the contigs associated with ten genes show mapping distances < 2.55 cM from the *S* locus, suggesting they may be potential candidates for *Oakleaf* should a *KNOX* gene not be responsible.

3.5 Discussion

The lobed leaves with occasional ectopic meristems that *P. vulgaris Oakleaf* mutants display is reminiscent of the overexpression phenotype of Class I *KNOX* genes in *Arabidopsis thaliana* (Lincoln et al., 1994, Chuck et al., 1996, Hay and Tsiantis, 2010). For this reason, we explored *KNOX*-like genes in *P. vulgaris* in an attempt to identify the cause of the *Oakleaf* phenotype. The current study reveals eight *KNOX*-like genes in the *P. vulgaris* genome; phylogenetic analyses facilitated classification into groups of five Class I and three Class II *KNOX*-like genes based on similarity to *A. thaliana* KNOX protein sequences, and phylogenetic analysis (Figure 3.5) (Kerstetter et al., 1994, Bharathan et al., 1999).

Differential expression studies carried out in the current study reveal that Class I *KNOX*-like genes in *P. vulgaris* are expressed at a low level in *Oakleaf* leaves and flowers, thus rendering them poor candidates for *Oakleaf*. In contrast, the Class II *PvKNOX* genes are highly expressed in both leaves and flowers of *Oakleaf* and wildtype. However, *PvKNL3* is the only gene upregulated in *Oakleaf* leaves and flowers, and the level of upregulation is minimal. In addition, studies in *A. thaliana* show that Class I and II *KNOX* genes are involved in different developmental processes; Class II *KNOX* family genes *KNAT3*, *KNAT4* and *KNAT5* are typically involved in root development (Truernit et al., 2006) and *KNAT7* in the formation of secondary cell walls (Li et al., 2011a, Li et al., 2012) whereas the Class I *KNOX* family genes have roles in shoot apical meristem identity (Hay and Tsiantis, 2010); this is consistent with the ectopic meristems that are sometimes produced on the veins of *Oakleaf* leaves (Cocker et al., 2015).

The above, as well as the abnormal lobed-leaf phenotype of *A. thaliana* plants with overexpression of *KNAT1* (Chuck et al., 1996), is the primary basis for the prediction that abnormal expression of a Class I *KNOX*-like gene might be the underlying cause of the *Oakleaf* phenotype. In comparison to the Class I *PvKNOX* genes, the comparable high-level expression of Class II genes in both leaves and flowers of *P. vulgaris* and *Oakleaf* is consistent with the broader expression of Class II *KNOX* genes in *A. thaliana*. This suggests the Class I and II *KNOX* genes in *P. vulgaris* may also be associated with different developmental processes due to their contrasting expression profiles (Serikawa et al., 1997, Bharathan et al., 1999, Truernit et al., 2006). The Class

II *KNOX* gene *PvKNL3* is therefore not considered a strong candidate for *Oakleaf* as it is a Class II *KNOX* gene that shows only minimal upregulation in leaves and flowers of *Oakleaf* plants. The Class I *KNOX* genes do not show strong upregulation in *Oakleaf*; the Class I *KNOX* gene *PvKNL6* does have higher read counts in both *Oakleaf* leaves and flowers, with higher fold-upregulation than *PvKNL3*, but the extremely low expression in both leaves and flowers is not consistent with the constitutive overexpression of *KNOX*-genes that is observed in *A. thaliana* (Lincoln et al., 1994, Chuck et al., 1996, Hay and Tsiantis, 2010). In conclusion, the *Oakleaf* phenotype is most likely not the result of overexpression of a *KNOX*-like gene in *Primula vulgaris*.

In lieu of a discernible difference in transcript expression, it was reasoned that mutation in a *PvKNOX* gene could lead to the *Oakleaf* phenotype through a dominant gain of function in protein activity. If there is an intronic mutation affecting a splice site in a *PvKNOX* gene then it would not be detected as a SNP in the *Oakleaf* RNA-Seq reads. However, it was reasoned that a splice site mutation might affect the coverage profile of *Oakleaf* RNA-Seq reads across the gene. For the eight *PvKNOX* genes, there was no discernible difference in the coverage profiles of *Oakleaf* and wildtype pin RNA-Seq reads, thus suggesting a splice site mutation was not resulting in a dominant gain of function in PvKNOX protein activity that could occur due to the absence of a critical regulatory domain as a result of protein truncation. The analysis of RNA-Seq reads mapped to *PvKNOX* genes revealed several *Oakleaf*-specific SNPs following removal of homozygous sites in *Oakleaf* and SNPs present in wild-type pin RNA-Seq reads. In the remaining heterozygous *Oakleaf* SNPs that were absent in pin, several were predicted to result in conservative amino acid changes and were discounted as the basis for *Oakleaf*. Three SNPs in *PvKNL2* and *PvSTL1* were predicted to cause truncation of the encoded protein, but were only observed in *Oakleaf* flower RNA-Seq reads, not leaves. Two SNPs in *PvKNL3* and *PvKNL7* would cause non-conservative amino acid substitutions, but were only observed in *Oakleaf* leaf RNA-Seq reads, not flowers. In conclusion, the absence of these SNPs in both leaf and flower *Oakleaf* RNA-Seq reads suggests that they are unlikely to be the cause of the *Oakleaf* phenotype. The differential expression and SNP analyses taken together suggest that mutation or overexpression of a *PvKNOX* gene is not the basis for *Oakleaf*.

In the studies associated with this chapter (Cocker et al., 2015), linkage of *Oakleaf* to the *S* locus was supported by the predominant co-segregation of *Oakleaf* with either the pin or thrum phenotype, with the observed number of recombinants between *Oakleaf* and the *S* locus being small in all crosses (Cocker et al., 2015). The crosses carried out with *Oakleaf* as the female parent pollinated with pollen from wildtype pin reveal *Oakleaf* as a single dominant locus, with chi-squared analysis supporting a 1:1 ratio ($P > 0.700$). The 1:1 ratio would be expected with a heterozygous parental *Oakleaf* plant, given the dominance of *Oakleaf* over wildtype. Furthermore, crosses between two heterozygous *Oakleaf* plants with thrum as either pollen donor or recipient reveal the 3:1 ratio of *Oakleaf* to wild-type progeny that would also be expected given a dominant *Oakleaf* locus ($P > 0.95$ and $P > 0.50$). However, one of the crosses with *Oakleaf* as the male parent revealed a significant ($P < 0.001$) deviation from the expected 1:1 ratio of *Oakleaf* to wild-type progeny. In this cross, 45 seedlings were lost before secondary leaf development. The first true leaves of *Oakleaf* plants show a consistent lobed phenotype that is similar in appearance to the leaves of Oak trees and shrubs (*Quercus*); this facilitates reliable identification of an *Oakleaf* plant (Cocker et al., 2015); seedlings that are lost before the first true leaves are produced cannot therefore be reliably classified as wildtype in the absence of *Oakleaf* characteristics prior to secondary leaf development.

The underrepresentation of *Oakleaf* progeny may be a statistical consequence of the low total number of progeny under consideration; however, if the 45 seedlings lost before secondary leaf development are included in the chi-squared analysis as wild-type, then the test would support a 1:1 ratio ($P > 0.300$) of *Oakleaf* to wild-type progeny. In addition, for three out of four crosses carried out between *Oakleaf* and wildtype plants, it appears that progeny losses before flowering are higher in wild-type than *Oakleaf*. The leaves of *Oakleaf* plants are thicker and firmer than wildtype (Cocker et al., 2015), and perhaps may be less susceptible to the 'damping off' that can occur in *Primula* seedlings due to bacterial or fungal infection before the secondary leaves emerge, thus suggesting the *Oakleaf* mutation may result in greater resilience to seedling loss under pathogen exposure or unfavourable conditions; this is a possible explanation for the distorted ratio of *Oakleaf* to wildtype progeny.

In *Arabidopsis*, *Antirrhinum* and tobacco, studies show that *asymmetric leaves 1* (*as1*) mutants of the *AS1* gene involved in downregulation of *KNOX* gene expression have

enhanced resistance to necrotrophic fungi, with phenotypes similar to *KNAT1* overexpression lines (Hay et al., 2002, Nurmberg et al., 2007). If *AS1* is responsible for the *Oakleaf* phenotype, then upregulation of the *PvKNOX* genes in *Oakleaf* might be expected, but this is not the case; perhaps overexpression of a downstream target of *AS1* or *PvKNOX* genes is the cause of *Oakleaf*. The identification of *Oakleaf* is of interest in order to determine the molecular basis of this potential resilience in *Oakleaf* seedlings; this may be of use in an agronomic setting.

RNA-Seq data assembled into a *de novo* assembly were used to join two fragments of *PvKNL1* that were located on two distinct genomic contigs; a single transcript spanning the two exons resolved the two fragments as a single gene. In this way, transcriptome data, particularly that spanning large introns, can effectively act as a long mate-pair (LMP) library with insert size "equivalent to intron length" (Zhang et al., 2016). This is presumably best approximated by the average intron size of the full complement of genes in the genome, analogous to the use of the average insert-size of an LMP library in conventional scaffolding approaches, without the complex and time-consuming process of generating such a library (Xue et al., 2013). In an analysis of intron size distributions in five high-quality assemblies, introns greater than 5 kb in length were present in 42.5%–83.9% of all gene loci (Xue et al., 2013). In addition, the concept of pervasive transcription indicates that the majority of the genome is transcribed to some degree (Clark et al., 2011), suggesting the above may act as a tool to scaffold large portions of the genome, particularly in transcribed and correspondingly gene-rich regions of functional interest (Xue et al., 2013). The use of RNA-Seq data to improve the *Primula vulgaris* genome assembly without the need for significant further funding is therefore an intriguing possibility for this non-model species given the availability of a comprehensive array of libraries representing a number of different tissues (Chapter 2); the current study provides a practical example for the potential utility of such an approach through the resolution of the *PvKNL1* gene structure.

Differential expression in *Oakleaf* and wildtype pin plants have revealed sets of up- and downregulated genes that may be downstream targets of *Oakleaf*. If one of the *PvKNOX* genes is shown to be the cause of the *Oakleaf* phenotype then these candidate genes will offer insight into the regulatory pathways leading to the development of the distinctive lobed leaves and attenuated flowers. If, on the other hand, the *PvKNOX* genes are not

responsible for *Oakleaf*, then these sets of genes may contain the key gene itself. Indeed, six of these genes lie in contigs that show preliminary mapping distances less than the highest estimated distance of *Oakleaf* from the *S* locus (2.55 cM). Class I *KNOX* genes in *Arabidopsis thaliana* are associated with downstream upregulation of *GA2* oxidase and *IPT7*, and downregulation of *GA20* oxidase, affecting cytokinin and gibberellin concentrations; genes involved in lignin synthesis are also downregulated (Hay and Tsiantis, 2010). There are no such genes present in the set of differentially expressed genes identified in *Oakleaf* based on the functional annotations performed; this perhaps suggests that the phenotype is caused by a different pathway. Indeed, the expression and mutation analyses presented here suggest that some of the eight *PvKNOX* genes are at best only very weak candidates for *Oakleaf*, despite the similarity of *Oakleaf* to the phenotypes seen in *Arabidopsis thaliana KNOX* overexpression lines; these lines do not however, present abnormalities in the petals despite the occasional appearance of ectopic meristems on the leaves as in *Oakleaf*. In addition, based on the preliminary linkage analysis carried out, it is likely that seven of the eight *PvKNOX* genes are not linked to the *S* locus; there is no evidence of linkage for the remaining gene. This analysis offers only a very crude estimate of the mapping distance from the *S* locus; the large distances predicted suggest that although the distance may be somewhat smaller due to the inherent imprecision of the method used, it is unlikely these genes are in tight linkage with the *S* locus. Future segregation analyses offering more robust estimates of the genetic distance will confirm whether any of the eight *PvKNOX* genes are linked to the *S* locus.

None of the eight *PvKNOX* genes are present in the BAC assembly generated in analyses associated with this study (Li et al., 2015a); this is perhaps not surprising as *PvSLL2* is estimated at < 1.37 cM from the *S* locus (Figure 3.2), with *Oakleaf* between 0.56 and 2.55 cM away. The *de novo* annotation of the BAC assembly confirms the positions of the *S* linked markers and reveals at least 86 new *S*-linked genes based on comparisons with genes in *Arabidopsis thaliana* (Li et al., 2015). The inspection of the functional annotations for the predicted genes in the BAC assembly reveals no obvious candidates for *Oakleaf*. The findings presented here suggest that *Oakleaf* is not caused by a *PvKNOX* gene. Future segregation analyses will confirm this, as well as reducing the number of candidates in the set of up- and downregulated genes identified. The molecular characterization of *Oakleaf* will contribute to resolving the orientation of the

BAC contig, and facilitate further BAC- or *in silico*-walking with genome contigs towards the *S* locus, to close the gap in the BAC assembly. The various genome assemblies presented in Chapter 2 perhaps offer the resources to accomplish this. This chapter describes efforts to identify *Oakleaf* and annotate the BAC assembly flanking the *S* locus, providing validation for the position of the *S* locus-linked markers and a platform for further analysis of the regions surrounding the *S* locus.

4

The structure of the *Primula vulgaris S* locus

4.1 Relevant publications

Li, J.*, **Cocker, J.M.*,** Wright, J., Webster, M.A., McMullan, M., Ayling, S., Swarbreck, D., Caccamo, M., Oosterhout, Cv., Gilmartin, P.M. (2016) Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nature Plants*, 2: 16188.

Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Dyer, S., Caccamo, M., Gilmartin, P.M. (2017) The *Primula vulgaris* genome (in preparation).

* These authors contributed equally

4.2 Introduction

The established genetic model for heterostyly in *Primula* predicts that the *S* locus comprises a cluster of at least three genes (Ernst, 1936c, Lewis and Jones, 1992). These genes control the development of morphological features associated with two forms of flower, the pin or thrum: *G* controls stigma height; *P* controls pollen size; and *A* controls anther positioning (Dowrick, 1956, Webster and Gilmartin, 2006). The pin form is thought to be homozygous recessive for the three genes (*gpa*/*gpa*) (*s*/*s*) and the thrum heterozygous (*GPA*/*gpa*) (*S*/*s*), such that presence of the dominant alleles for the three predicted genes results in plants presenting thrum flowers with a short style due to repression of cell elongation (*G*), elevated anthers through increased cell division in the corolla tube (*A*), and large pollen (*P*); pin flowers have the opposite morphology, thus promoting insect-mediated outcrossing based on the reciprocal positions of the reproductive structures (Webster and Gilmartin, 2006).

The *S* locus has remained elusive since Darwin first explained the importance of heterostyly in promoting outcrossing more than 150 years ago (Darwin, 1877); the independent origins of this system in distinct orders marks it as a remarkable example of convergent evolution, with similar genetic architectures predicted for the *S* locus in *P. vulgaris*, *Turnera subulata*, and *Fagopyrum esculentum* (Gilmartin and Li, 2010). It is predicted that the constituent genes are very tightly linked, so much so that they have been termed a 'supergene', a term used to describe loci controlling other complex morphologies in plants, animals and fungi (Pellow, 1928, Darlington and Mather, 1949, Schwander et al., 2014, Thompson and Jiggins, 2014).

It is widely accepted that extremely rare recombination events within the *S* locus supergene result in homostyles that present flowers with the anthers and stigma at the same height (Lewis, 1954, Dowrick, 1956). In some cases no recombinants have been observed in as many as 12,000 *Primula* plants, with reports suggesting that the *S* locus in *Turnera sublata* could be as small as 32 kb (Ernst, 1928b, De Winton and Haldane, 1935, Mather, 1950, Dowrick, 1956, Labonne et al., 2009), as such it was originally proposed that homostyles were the result of mutation, prior to recombination becoming the accepted explanation (Ernst, 1936b). Disruption of the architectural constraints imposed by heterostyly in homostyles is also associated with a breakdown in self-incompatiblity (SI); this is another key facet of the accepted genetic model, thought to result from recombination at the *S* locus disrupting linkage with the potentially *S*-linked SI components that normally serve to inhibit within-morph pollination (Lewis, 1954, Dowrick, 1956, Lewis and Jones, 1992, Barrett and Shore, 2008).

The genes controlling heterostyly are thought to be within a region of suppressed recombination (Mather, 1950), thus maintaining the collective supergene and the distinct pin and thrum phenotypes; *in situ* hybridisation studies related to Chapter 3 show that the *S* locus is located close to the centromere of the largest pericentric chromosome (Li et al., 2015a), a region characterised by reduced recombination (Mather, 1939, Choo, 1998). The attempts to isolate genes at the *P. vulgaris S* locus through the BAC walking strategy described in Chapter 3 (Li et al., 2011b, Li et al., 2015a) were halted in progress towards the *S* locus, with a gap remaining between the two BAC contigs flanking the *S* locus due to highly repetitive sequences, resulting in multiple BACs being identified in further screening of the BAC library (Li et al.,

2015a); one possibility is that this is a result of repeat-rich sequences near to the centromere (Kursel and Malik, 2016).

The minimum number of genes at the *S* locus (three: *G*, *P*, and *A*) is predicted through the observation of putative recombinants that disrupt linkage between the reciprocal morphological features that co-segregate with the two floral forms (Ernst, 1936a, Lewis and Jones, 1992), with the order *GPA* defined based on the least number of double recombinants being required to describe the number of recombinants seen (Ernst, 1936a, Lewis and Jones, 1992). The prediction that recombination is the mechanism that disrupts the supergene and produces homostyles underpins the last 60 years of research directed towards identifying the molecular and evolutionary basis of heterostyly. There is considerable debate on the predicted number of genes at the *S* locus. Dowrick (1956) speculated that the *S* locus supergene could comprise up to seven genes, with the various morphological and physiological traits specified by distinct genetic factors, but there is no support from recombinants for this prediction. Kurian and Richards (1997) concluded that there must be seven genes at the very minimum, whilst Richards (2003) proposed as many as nine.

The identification of the *S* locus and subsequent functional analyses will reveal the true number of genes controlling heterostyly, but since there is no reason to suggest that the heterostyly-determining supergene contains only protein-coding genes, it will also be important to carry out further annotation of the *S* locus region. This is exemplified by reports that microRNAs (miRNAs) play a key role in the development of floral reproductive structures through the targeted regulation of *auxin response factor 6* (*ARF6*) and *ARF8* expression; double mutants of *ARF6* and *ARF8* present short stamen filaments due to decreased cell expansion as compared to wildtype, whilst overexpression of MIR167a driven by the 35S promoter mimics this phenotype (Nagpal et al., 2005, Wu et al., 2006, Su et al., 2016). Because of the increased distance between the anthers and stigma, *arf6-2* and *arf8-3* plants self-pollinate inefficiently; epidermal cells of double mutant stamen filaments were about half as long as those of wild-type filaments (13 compared with 24 μm), indicating that decreased cell expansion caused the short filaments (Wu et al., 2006). The annotation of miRNAs at the *S* locus will therefore be important in providing a full picture of the potential regulatory modules contained therein.

It has been suggested that until the molecular basis of heterostyly is revealed, further speculation on the SI determinants and evolutionary steps leading to heterostyly will be severely limited (Barrett and Shore, 2008, McCubbin, 2008). The integration of the genetic map and BAC assembly associated data (Chapter 3) (Li et al., 2011b, Li et al., 2015a) with the array of read libraries, assemblies and associated genomic resources presented in Chapter 2, provides an opportunity to identify the *S* locus and its constituent genes based on a holistic approach that exploits both genetics and genomics, to offer insights into the genomic architecture of the region that were touched on by Mather (1950) and provide a blueprint for defining the basis of heterostyly in diverse angiosperm families for the first time since Darwin's studies 150 years ago (Darwin, 1877).

4.3 Methods

4.3.1 Read depth across the *S* locus

Four long-homostyle (LH_v2) contigs forming the 455,880 bp *S* locus and flanking regions identified in Li et al. (2016) were removed from the LH_v2 assembly and replaced with the contiguous 455,880 bp sequence comprising the full *S* locus and flanking regions, as assembled by JL (Li et al., 2016). Genomic sequencing reads from thrum, pin, long-homostyle and short-homostyle plants (Table 2.1) were aligned to the *Primula vulgaris* LH_v2 genome assembly containing the contiguous *S* locus sequence using BWA (v0.7.12) (Li and Durbin, 2009).

The SAMtools "depth" tool (v0.1.19) (Li et al., 2009a) was used to return the depth of coverage at each position for reads across the *S* locus and flanking regions for each read library. The read depth between each of the four libraries was normalised according to the average size of the four read libraries, and the depth of read coverage plotted across the *S* locus region in 5,000 bp windows using R (v3.2.0) (https://www.r-project.org/).

In addition, box plots were drawn to summarise read depth across the 278 kb *S* locus region for each of the four read libraries using R (v3.2.0); the *CFB* genes flanking the *S* locus were excluded and the predicted read depth calculated by taking the median of the total summed depth for all read libraries at each position and then determining the

expected read depth for each library by multiplying by the number of *S* alleles in that individual: the homozygous long-homostyle (median multiplied by two), heterozygous thrum and short-homostyle (median multiplied by one), and absence of the 278 kb region in pin (median multiplied by zero).

4.3.2 RNA-Seq differential expression analysis

RNA was isolated in biological replicates from 15-20 mm buds of four wild-type pin plants and four wild-type thrum plants for RNA-Seq with Illumina HiSeq2000 by JL, as described in Li et al. (2016); reads were screened for rRNA removal using SortMeRNA (Kopylova et al., 2012) and quality-trimmed with trim galore (Q20) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore).

The RNA-Seq reads were aligned to the long-homostyle (LH_v2) genome assembly with TopHat (Trapnell et al., 2012) (v2.0.13) and assembled using Cufflinks (Trapnell et al., 2012) (v2.2.1), guided by LH_v2 gene model annotations after manual curation of all *S* locus genes (Trapnell et al., 2012). Differential expression analysis between the four pin- and four thrum-replicate libraries was carried out using Cuffdiff (Trapnell et al., 2012); the number of fragments per kilobase of transcript per million fragments mapped +1 (FPKM + 1) ($\log_{10}$-transformed) is reported for genes at the *S* locus in Figure 4.7.

4.3.3 Analysis of thrum-specific genome regions

Individual pin and thrum plants used for genome sequencing and assembly of PP_ and TP_ assemblies (Table 2.2) were used (by JL) to generate segregating pin and thrum progeny, which were sequenced in separate pools of DNA from 24 pin and 28 thrum plants (LIB1730 and LIB1731, respectively; Table 2.1); the thrum progeny pool was not used in this analysis. RNA-Seq reads from the four pin and four thrum replicate libraries, as described in the differential expression analysis above, were aligned to the thrum parent genome assembly (TP_v1) using TopHat (v2.0.13) (Trapnell et al., 2012); transcripts were assembled and merged with Cufflinks and Cuffmerge (Trapnell et al., 2012) (v2.2.1) and differential expression carried out with Cuffdiff (Trapnell et al., 2012).

Transcripts showing thrum-specific expression or near to thrum-specific expression (cut-off < 0.1 FPKM for pin flower) were identified from the differential expression results; pin-progeny reads were then aligned to the TP_v1 assembly using BWA (v0.6.2) (Li and Durbin, 2009) and the per-base depth of read coverage for contigs containing the transcripts calculated using the SAMtools (Li et al., 2009a) (v0.1.18) "depth" tool. The per-base depth of read coverage was then used to calculate the average depth and breadth of pin-progeny read coverage across each transcript region in the genomic contigs identified by a thrum specific transcript.

The transcripts were classified into two groups using the *k*-means algorithm implemented in the *scikit-learn* package for Python, with the breadth of pin-progeny read coverage and $\log_{10}$-transformed depth of pin-progeny read coverage across each transcript region as input variables (n_clusters=2); matplotlib was used for plotting in Python (Hunter, 2007, Pedregosa et al., 2011). Gene identities of thrum genome-specific transcripts were determined by alignments to the LH_v2 assembly and gene models using Exonerate v2.2.0 (Slater and Birney, 2005).

4.3.4 Detection of recombination in *S* locus flanking regions

Genomic paired-end reads from pin and thrum parental plants (Table 2.1) were aligned to the long homostyle (LH_v2) genome with BWA (v0.6.2) (Li and Durbin, 2009). SAMtools (v0.1.18) (Li et al., 2009a) was used to remove PCR-duplicates (over-amplified fragments) with the "rmdup" tool, and for variant calling between the two read libraries and LH_v2. The genotype (GT) sub-field in the resulting Variant Call Format (VCF) files was used to determine the genotype for pin and thrum at each nucleotide position; two analyses were then carried out: firstly, a phased analysis using only heterozygous sites in thrum (figures not shown; see Li et al. (2016)), and secondly, using heterozygous sites in thrum as well as homozygous sites in thrum where at least one of the alleles in pin was different to the nucleotide in thrum at that site. Sites were excluded with depth (DP) < 10, genotype quality < 30, or mapping quality (MQ) < 20 for heterozygous thrum sites (first and second analysis), or in either pin or thrum for homozygous thrum sites (second analysis).

The cumulative binomial probability was calculated for the *S* locus left- and right-flanking sequences using a sliding window of 5,000 bp and an overlap (step size) of 1,000 bp to test whether the observed frequency of variant sites in each window was significantly lower than that expected given the total number of variant sites in each flanking sequence (Equation 4.1); this was performed using variant sites in (i) to (iii) below.

$$P(X \leq m) = \sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k}$$

Equation 4.1 The cumulative binomial formula used to calculate the probability of observing *m* or fewer variant sites *(k)* in a window of size *n*, given an overall average frequency (probability) of variant sites (*p*) for the (left or right) flanking sequence under consideration, where *n* is the size of the window reduced by the number of sites excluded based on quality cut-offs or presence of ambiguous bases (Ns) (see below), and *p* is as stated in Equation 4.2.

$$p = \frac{t}{l-n}$$

Equation 4.2 The probability (*p*) of observing *t* variant sites in a flanking sequence of overall length (*l)*, where *n* is the number of sites excluded in the overall sequence based on quality cut-offs or presence of ambiguous bases (Ns), as applied in the cumulative binomial formula presented in Equation 4.1.

In cases where ambiguous bases ("N"s) were present, the total size of the window, or flanking sequence as a whole, was reduced by the number of Ns in that window or flanking sequence, respectively, with windows comprising solely of Ns being excluded

from the analysis; sites excluded from the genotyping analysis above based on depth and quality cut-offs were omitted in the same manner.

Three analyses were carried out in this way for both left- and right-flanking sequences: (i) including all variant sites, (ii) with variant sites in coding sequences excluded, (iii) with variants sites in genic regions (including predicted introns, exons, 3'- and 5'-untranslated regions) excluded, based on LH_v2 gene annotations (above). The $-\log_{10}$(cumulative binomial probability) and total number of SNPs in each window were plotted in R (v3.2.0), with $-\log_{10}(p=0.05)$ and $-\log_{10}(p=0.05)$ with Bonferroni correction indicated; the latter based on the total number of windows analysed in each flanking region. Probabilities falling below the Bonferroni corrected critical value can be taken as evidence of recent genetic exchange between the two sequences, based on the method used in HybridCheck (Ward et al., 2015).

4.3.5 Annotation of microRNAs (miRNAs) in the *Primula vulgaris S* locus

MicroRNAs were predicted in the *Primula vulgaris S* locus using the following procedure: first, regions with homology to known miRNAs were identified, then miRNAs were predicted using further filtering and statistical analysis with NOVOMIR (Teune and Steger, 2010). MicroRNA sequences from the miRBase (release 22) library of known miRNA sequences (Griffiths-Jones et al., 2006) were mapped to the *Primula vulgaris S* locus using BLASTN (e-value=10) (Camacho et al., 2009). To generate sequences more likely to fold into a hairpin structure, BLASTN search results were filtered to retain forward-orientation hits with less than 5 mismatches, and reverse-complement hits with more than 6 mismatches, sequences with both a forward and reverse hit were retained; this follows the procedure for annotation of miRNAs in the bread wheat (*Triticum aestivum*) genome (Mayer et al., 2014). The distance between forward and reverse hits was fixed at 3-1200 nucleotides; from the remaining hits, precursor sequences were generated using the sequence spanning 13 nucleotides before the 5' hit to 13 nucleotides after the 3' hit, to account for the pri-extension region of the hairpin (Mayer et al., 2014). NOVOMIR (Teune and Steger, 2010) was used (default options) to predict miRNAs from the generated precursors; overlapping precursors were considered as one miRNA locus.

4.3.6 Similarity searches of *S* locus genes to the *P. vulgaris* genome

The coding sequences (CDSs) for the five protein-coding genes at the *P. vulgaris S* locus were aligned to the *P. vulgaris* LH_v2 genome assembly using TBLASTX (Camacho et al., 2009). High Scoring Pairs (HSPs) occurring on the same contig for each CDS were collated together and the average percent (%) identity (PID) calculated and plotted with R (v3.2.0) (https://www.r-project.org/).

4.3.7 Repeat analyses of the *P. vulgaris S* locus

The repeat library constructed using the *P. vulgaris* LH_v2 genome assembly (Chapter 2) was used to find the proportion of TEs in the 278 kb *S* locus region using RepeatMasker (http://www.repeatmasker.org/). The RepeatMasker output file generated for the *P. vulgaris* LH_v2 assembly (Chapter 2) was used to calculate the percentage of repeats for all contigs in the genome. R (v3.2.0) (https://www.r-project.org/) was used to plot the TE density.

4.3.8 Analysis of intron sizes in the *Primula vulgaris S* locus

The GFF file of predicted genes in the *P. vulgaris* LH_v2 assembly was used to determine intron sizes for all genes in the genome. The density of intron sizes across the genome was plotted using R (v3.2.0) (https://www.r-project.org/).

4.4 Results

4.4.1 Alignment of thrum-specific BAC 70F11 generates 455 kb assembly

In the studies associated with Chapter 3 (Li et al., 2015a) and the BAC-walking strategy presented in Li et al. (2011b) a BAC was identified (by JL) using *GLO*^{T}, a thrum-expressed allele of the *S* locus-linked *P. vulgaris GLOBOSA* (Li et al., 2010) as a probe. BAC 70F11 could not be positioned relative to BACs in the map generated around the *S* locus (Li et al., 2015a). The alignment of 70F11 to contigs in the LH_v2, TP_v2 and PP_v1 genome assemblies joins two LH_v2 contigs together, with a further contig from the TP_v2 assembly extending this assemblage to include two more LH_v2 contigs; thus forming a contiguous 455,811 bp sequence (Figure 4.1) as presented in Li et al. (2016). The PP_v2 assembly contains a contig that spans the boundaries of a 278 kb region that appears to be absent from the pin genome (Figure 4.1), as such this region was earmarked as the potential *S* locus.

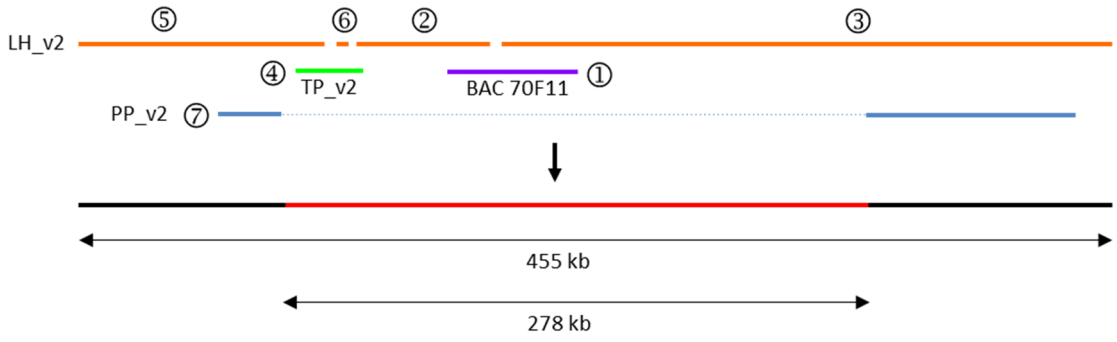


Figure 4.1 The 455,811 bp assembly initiated by BAC 70F11. BAC 70F11 (1) (purple) aligns with long homostyle (LH_v2) contigs (2) and (3) (orange); thrum (TP_v2) contig (4) (green) aligns with LH_v2 contigs (5) and (6) (orange) to generate the complete 455 kb assembly (black and red). PP_v2 (pin) contig (7) blue aligns to LH_v2 contigs (5) and (3) but does not align to the central 278 kb region (red), thus highlighting the 278 kb thrum-specific region as the potential *S* locus.

4.4.2 Predicted genes in the 455 kb assembly

Five predicted LH_v2 genes are present in the 278 kb region identified above (Figure 4.1), supported by RNA-Seq data (see Chapter 2 for details of gene predictions). Prior to differential expression analyses (see below), these gene models were manually curated by JL based on PCR analyses and manual alignments (Li et al., 2016), with gene models correspondingly modified (by JMC) in the GFF (General Feature Format) file comprising the full complement of *P. vulgaris* genes and their predicted positions in the LH_v2 assembly. Interestingly, the 278 kb region is immediately flanked by direct repeats of around 3 kb in length; these sequences show similarity to *Cyclin-like F box* (*CFB*) genes, and are themselves ~98% similar according to alignments with BLASTN (v2.5.0+) (Zhang et al., 2000). There are seven additional predicted genes in the region to the left of the central 278 kb, and eight to the right.

4.4.3 Genomic reads aligned to the 455 kb assembly

Pin and thrum genomic reads were aligned to the LH_v2 assembly and the read depth plotted across the 445 kb assembly (Figure 4.2). The analysis reveals a read depth of ~60 in regions flanking the 278 kb region, whilst in the central 278 kb the depth is about half this for the thrum reads. The pin genomic reads largely do not map to this central region suggesting it is absent from the pin genome.

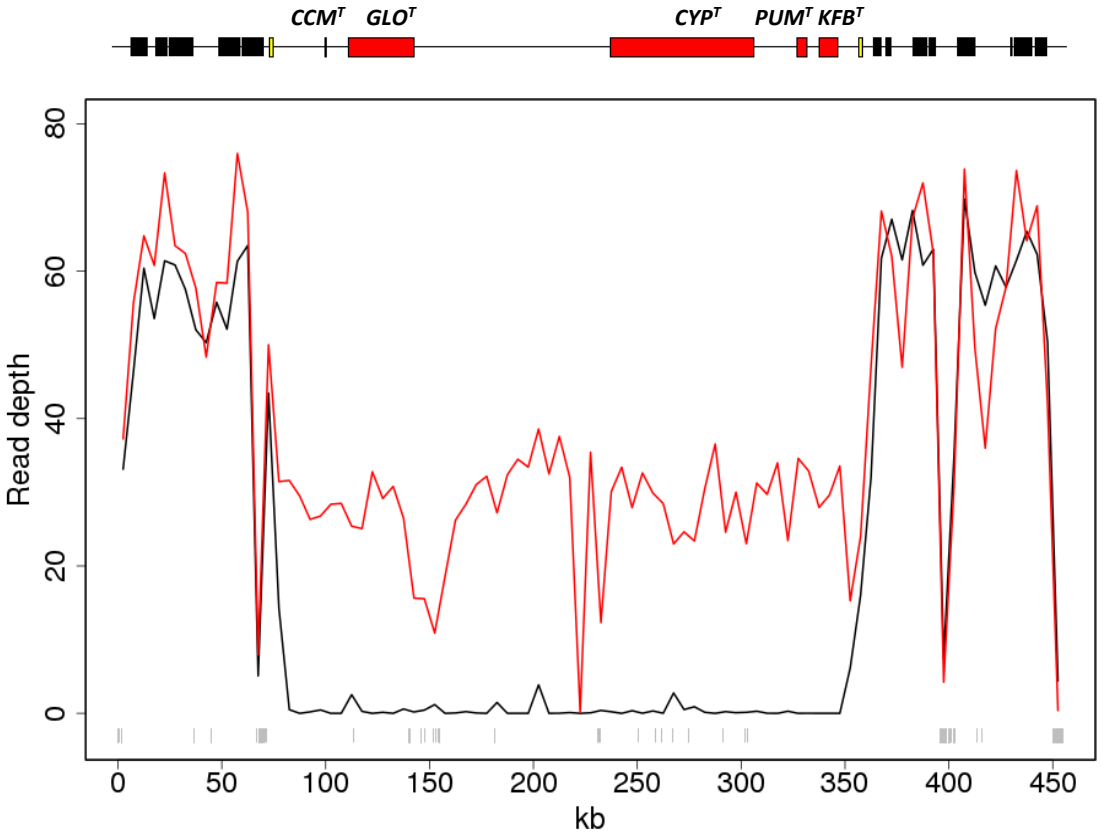


Figure 4.2 Read depth of genomic paired-end reads across the 455 kb assembly in 5 kb non-overlapping windows, normalised according to library size, pin = black, thrum = red. Genes within the 278 kb central region are shown (top of figure) in red, with the two yellow boxes representing duplicated *CFB* loci that flank the 278 kb sequence. Predicted genes in the flanking regions are also shown (black). Grey vertical lines near the x-axis represent ambiguous bases ("N"s) in the assembled sequence.

Further alignments of short and long homostyle genomic sequencing reads (Figure 4.3) show that the inbred homozygous long homostyle (*S**/*S**) has twice the coverage of the thrum in the 278 kb region, whilst the short homostyle (*S**/*s*) has roughly the same coverage as thrum (with S^* denoting a disrupted *S* haplotype present in either the long or short homostyle). These analyses show that the thrum plant is hemizygous for the 278 kb region, and confirm that it is absent from the pin (*s*/*s*) genome (PP_v2) rather than being the result of a misassembly or lack of genome coverage. This is supported by the apparent absence of the *s* haplotype in both short homostyle and thrum genomes (both are hemizygous but behave as heterozygotes), with the putative *S* haplotype present in two copies in the homozygous long homostyle.

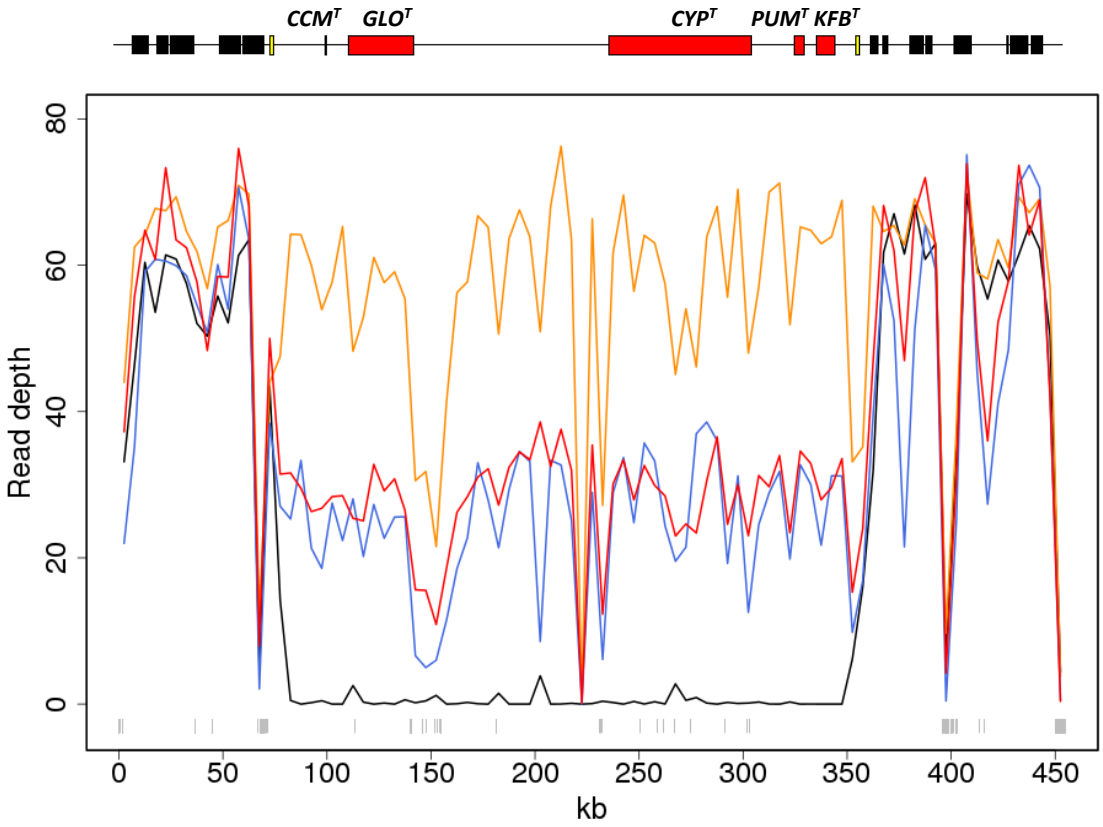


Figure 4.3 Read depth of genomic paired-end reads across the 455 kb assembly in 5 kb non-overlapping windows, normalised according to library size, pin = black, thrum = red, long homostyle = orange, short homostyle = blue. Genes within the 278 kb central region are shown (top of figure) in red, with the two yellow boxes representing duplicated *CFB* loci that flank the 278 kb sequence. Predicted genes in the flanking regions are also shown (black). Grey vertical lines near the x-axis represent ambiguous bases ("N"s) in the assembled sequence.

To further evaluate read depth across the 278 kb region, the observed read depth was compared to predicted read depth based on the median of the overall depth for all mapped read libraries (Figure 4.4). The pin (2n=0) has approximately zero mean(±SEM) read depth (0.11(±0.40)) suggesting the minimal number of aligned reads are the result of mapping errors. The majority of troughs seen in the read coverage profiles across the 278 kb region (Figure 4.2 and Figure 4.3) are the result of ambiguous bases ("N"s); the remainder are presumably the result of misaligned reads due to repetitive sequences, with the long homostyle most affected due to regions of low coverage being comparatively lower than the predicted coverage for this 2n=2 sample.

The long homostyle mean(±SEM) read depth (45.52(±1.32)) is approximately double that of the thrum (2n=1) (24.98(±0.79) and short homostyle (2n=1) (23.95(±1.12)), thus confirming that the long homostyle is homozygous, and thrum and short homostyle hemizygous, for the presumptive 278 kb *S* locus. This is in contrast to previous predictions that the thrum is heterozygous at the *S* locus (Bateson and Gregory, 1905, Dowrick, 1956)

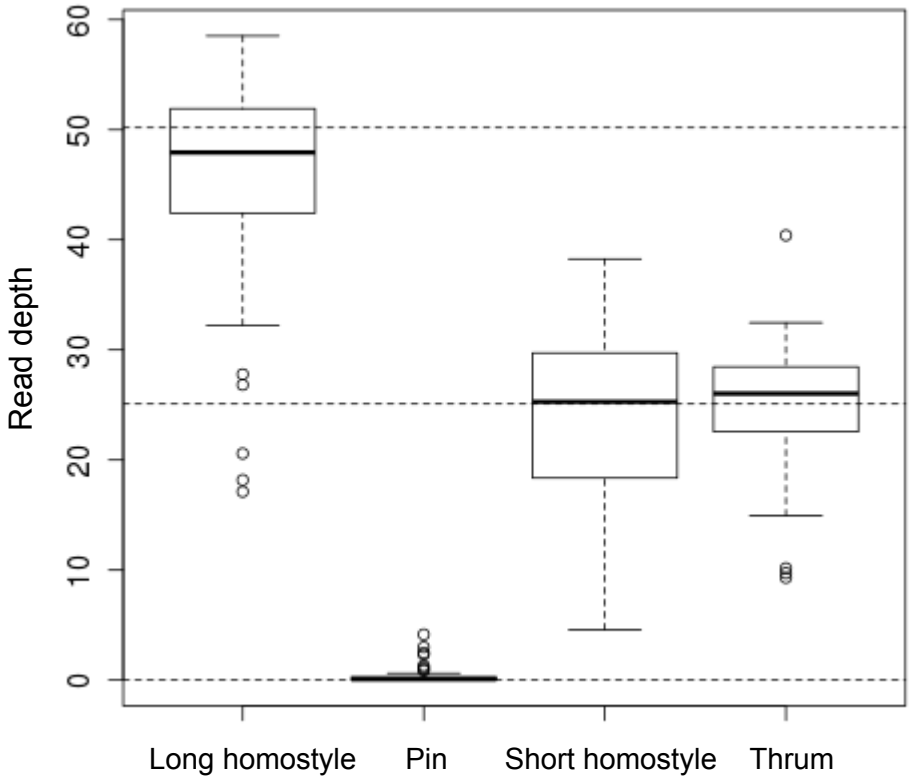


Figure 4.4 Box plots showing the normalised genomic read depth in 5 kb windows across the 278 kb region, with horizontal dotted lines in black representing the expected read depth for the long homostyle (2n=2), thrum and short homostyle (2n=1), and pin (2n=0), where 2n=x is the median of the summed read depths for all read libraries at each site divided by x.

Analyses (by JL) using PCR to amplify across the boundaries of the 278 kb sequence with primers designed for the sequences flanking this region show that the 278 kb region is completely absent in pins (Li et al., 2016). The pin (*s*) haplotype contains only

one *CFB* gene, which shows greater similarity to the left-flanking *CFB* gene (~99%) than the right-flanking *CFB* gene (~97%), thus the right-flanking *CFB* gene appears to be thrum-specific, consistent with its positioning in Figure 4.2 and Figure 4.3.

4.4.4 Functional evaluation of genes in the 278 kb thrum-specific region

The five *S* locus genes in the 278 kb region were named CCM^T, GLO^T, CYP^T, PUM^T and KFB^T (displayed from left to right as red boxes in Figure 4.2) based on their similarity to genes from other species (Figure 4.6). GLO^T was originally characterised as a thrum-expressed allele of *P. vulgaris GLOBOSA (GLO)* (Li et al., 2010), a B-function MADS-box gene responsible for floral development that is positioned in the BAC assembly discussed in Chapter 2 (Li et al., 2015a). *GLO* is linked to but not at the *S* locus. In this study GLO^T is shown to be a separate locus on a distinct contig, and is 83% similar in nucleotide sequence to *GLO*.

CYP^T shows similarity to *Arabidopsis CYP72B1*, which encodes a cytrochrome P450 with brassinolide 26-hydroxylase activity (Turk et al., 2003); brassinosteroids such as brassinolide play an important role in the regulation of cell elongation, whilst specific hydroxylation at the C-26 position caused by upregulation of *CYP72B1* has been shown to mimic brassinolide deficiency (Fu et al., 2012, Meaney, 2005, Zhu et al., 2013).

CCM^T (*Conserved Cysteine Motif*) encodes a protein with a novel C-terminal domain conserved across monocots and dicots (see discussion). PUM^T shows similarity to Pumilio, an RNA-binding protein (Abbasi et al., 2011) and KFB^T shows similarity to the *Arabidopsis* Kiss-Me-Deadly Kelch-repeat F Box protein, which downregulates cytokinin activity (Kim et al., 2013b); cytokinins are plant hormones with a well-known role in stimulating cell division, or cytokinesis (Del Bianco et al., 2013). The genes to the left and right of the thrum-specific region were named *S* locus *Flanking Gene Left* (SFG^L) and *Right* (SFG^R), with the duplicated sequences immediately flanking the thrum-specific region designated CFB^{TL} and CFB^{TR} based on sequence similarity to *Cyclin-like F Box* genes; as mentioned above, the pin has only one of these genes, named CFB^P.

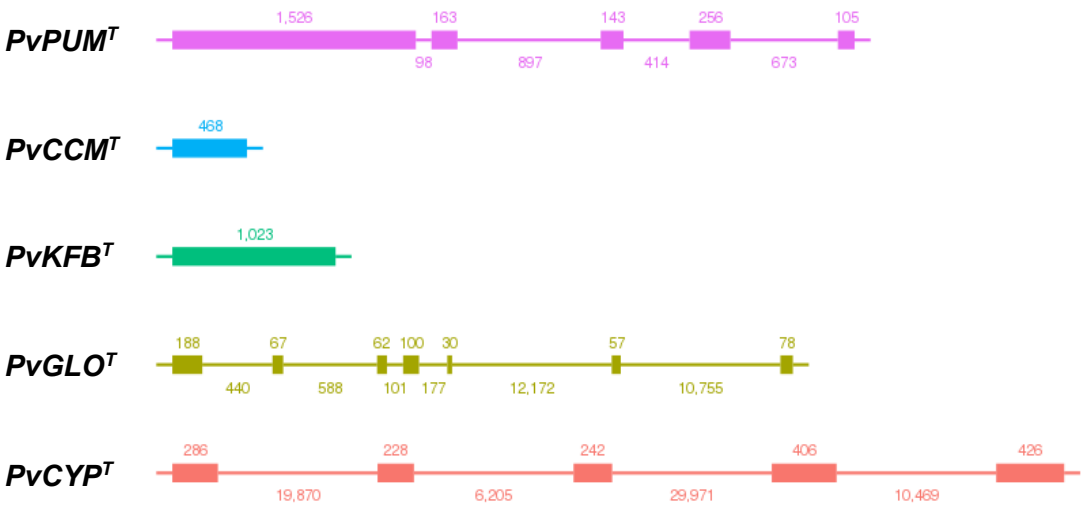


Figure 4.5 The gene structures of the five genes at the *P. vulgaris S* locus; exons = thick lines, introns = thin lines, introns are to scale except those greater than 1 kb which are displayed as 1 kb.

The short homostyle genome assembly (SH_v1) reveals a 2.5 kb retro-transposon insertion in exon 2 of GLO^T that would cause truncation of the encoded protein compared to the coding sequence for GLO^T present in LH_v2. CYP^T has a single base insertion in exon 3 in the LH_v2 assembly compared to the thrum assembly (TP_v2), resulting in a premature stop codon. CYP^T cDNA from an independent long homostyle from the Chiltern Hills (Crosby, 1949) provided a second CYP^T mutant allele with a G–C transversion in exon 2 that results in an Asp 126 His substitution (Li et al., 2016).

In the context of the established *GPA* model, which describes at least three genes at the *S* locus that control the development of heterostyly (Ernst, 1936a, Lewis and Jones, 1992), these findings earmark GLO^T as the *A* function gene, since the short homostyle has anthers with reduced elevation. CYP^T is the *G* function: it predicted to encode a brassinolide 26-hydroxylase, suggesting it might inhibit cell elongation that is normally promoted by brassinolides (Fu et al., 2012, Meaney, 2005, Zhu et al., 2013). Increased cell elongation in the style dictates stigma height in the pin form, whilst in contrast to the short style and high anthers of the thrum, the long homostyle in which CYP^T is disrupted has elevated anthers and stigma. Correspondingly, PCR analysis (by JL) showed that GLO^T and CYP^T are not expressed in the short and long homostyle,

respectively (Li et al., 2016): these plants are self-fertile, suggesting the SI determinants act downstream.

4.4.5 Linkage of the 278 kb region to the *S* locus

The thrum-specificity of the 278 kb sequence is supported by the multiple coverage profiles of thrum, short homostyle and pin genomic reads (Figure 4.3). These data suggest that this region might be linked to the thrum allele of the *S* locus.

The approximate distance from the *S* locus (cM) for contigs flanking the 455 kb assembly was calculated using the crude pooled-progeny method used to test for linkage of *Oakleaf* candidates to the *S* locus (Chapter 3); due to the thrum-specificity of the *S* locus, SNPs between pin and thrum would be present only in the regions flanking the *S* locus that lie beyond the duplicate *CFB* genes. The same procedure was carried out for genes previously identified as linked to the *S* locus: namely *PvSLL1*, *PvSLL2*, and *PvGLO* (Li et al., 2015a), alongside the unlinked *PvDEF* that is not on the *S* locus chromosome (Li et al., 2008).

Following the finding that the putative *S* locus is thrum-specific, this method provided the first indication though segregation analysis that contigs in the 455 kb assembly were linked to the *S* locus (Table 4.1). The method appears to be reasonably capable of identifying linkage given the presence of a sufficient number of usable SNPs in the contig, therefore supporting the conclusion that seven of the eight *PvKNOX* genes (Chapter 3) are unlikely to be linked to the *S* locus (Table 3.3). The distance from the *S* locus for *PvSLL1* is underestimated, with previous studies suggesting a genetic distance of < 0.57 cM; this is most likely due to the low number of SNPs analysed.

In general, however, it appears that distances are over-estimated using this method, perhaps due to the inherent problems with unresolvable sequencing errors and coverage issues when using pools of individuals in a mapping approach (Raineri et al., 2012, Schlotterer et al., 2014). There are no SNPs associated with *PvGLO* and the unlinked *PvDEF*, suggesting a lack of SNPs is not indicative of a sequence being either linked or present on a different chromosome (unlinked). The method provides a preliminary measure of whether a sequence is linked to the *S* locus.

Contig	# SNPs	Distance (cM)	±SEM	Li et al. 2015 (cM)
PvSLL2 contig	50	4.68	0.54	0.05 to 1.37
LH_v2 Contig 5 (left-flanking)	169	0.68	0.08	-
PvSLL1 contig	2	0.00	0.00	0.05 to 0.57
LH_v2 Contig 3 (right-flanking)	70	1.45	0.38	-
PvGLO contig	0	NA	NA	0.39 to 1.64
PvDEF	0	NA	NA	-

Table 4.1 The mean genetic mapping distance from the *S* locus (cM) for SNP positions across LH_v2 contigs associated with genes previously shown to be linked to the *S* locus (Li et al., 2015a) and LH_v2 contigs flanking the *S* locus, numbered according to Figure 4.1. The standard error of the mean (SEM) is indicated for each distance; problems with coverage and unresolved sequencing errors due to the use of pooled data (Schlotterer et al., 2014) may result in variability between the predicted distances at each site. The number of SNPs (# SNPs) is shown as the SEM may be artificially low in the event of a low number of SNPs being present in the contig. The genetic distance determined for each of genes in Li et al. (2015) is also shown where available.

In *P. vulgaris* there are two *GLOBOSA*-like MADS-box genes: GLO^T is at the *S* locus, and *GLO* close by in the BAC map flanking the *S* locus (Chapter 3). Further analyses (by JL) (Li et al., 2016) used *GLO* (present in both pin and thrum) (Li et al., 2010) and GLO^T specific primers with DNA from a 2075 progeny three-point cross. This cross placed *Oakleaf* (Cocker et al., 2015) (< 1.7 cM) and *Hose in Hose* (Li et al., 2010) (< 1.6 cM) on either side of the *S* locus (Li et al., 2015a). This analysis did not reveal evidence of linkage disruption between GLO^T and the thrum phenotype in DNA from the recombinant or double-recombinant progeny that comprise recombinants between *Oakleaf* and/or *Hose in Hose* and *S*. This places the 278 kb region between *Oakleaf* and *Hose in Hose*, and suggests that it might lie in the gap in the BAC assembly that exists due to no BACs being identified that link the 455 kb region to BAC contigs *S*-left and *S*-right (Li et al., 2015a) (Chapter 3).

PCR analyses (by JL) show that GLO^T is completely absent in a pool of 200 pin plants from a natural population of *Primula vulgaris*; none of these plants showed

recombination, therefore confirming that the 278 kb region containing the five-gene cluster is tightly linked to the thrum allele of the *S* locus (< 0.2 cM) (Li et al., 2016).

4.4.6 Expression of genes at the *S* locus

The expression of the five genes at the *S* locus was analysed using RNA-Seq expression data comprising four replicate RNA-Seq libraries from pin and thrum flowers aligned to the LH_v2 genome assembly, as presented in Figure 2.9 (Chapter 2). These data reveal that the five genes in the central *S* locus region (GLO^T, CYP^T, PUM^T, KFB^T and CCM^T) are expressed only in thrum-flowers (Figure 4.7).

In the regions flanking the 278 kb region, the predicted genes are expressed in both pin and thrum flowers, except SFG^R6 which has extremely low expression in thrum flowers; SFG^L1 is expressed at a low level in both pin and thrum. The duplicate flanking gene CFB^{TL} is expressed at a low level in both pin and thrum, whilst CFB^{TR} shows no expression (Figure 4.7).

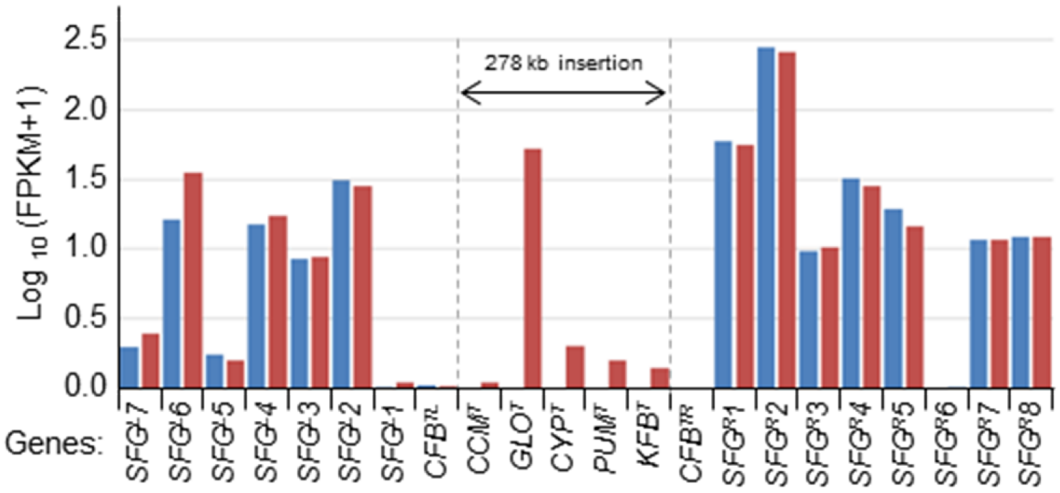


Figure 4.6 The $\log_{10}$(FPKM+1) expression in thrum flowers (red bars) and pin flowers (blue bars) for the five genes at the 278 kb thrum-specific *S* locus, and for the $SFG1\text{-}8^R$ (*S* locus *flanking gene-right*) and $SFG1\text{-}8^L$ (*S* locus *flanking gene-left*) genes predicted in the regions flanking the 278 kb sequence that are present in both pin and thrum genomes.

The level of expression for some of the genes at the *S* locus appears to be lower than it is for the majority of genes surrounding the 278 kb region; the data show that CYP^T for example is expressed to a lower level than GLO^T. One reason for this might be that CYP^T, if it proves to be the gene controlling style length, may only need to be expressed at a low level and act within a small developmental time frame on a minimal number of style cells, whilst GLO^T may need to regulate downstream genes controlling a broader range of cells at different time points given that the control of anther height in *Primula* is dictated by cell division in the entire corolla rather than just the anther filaments as in some heterostylous species (Webster and Gilmartin, 2006, Cohen, 2010). Once the *S* locus genes of those species have been identified it will be interesting to note whether the expression of the *A* function gene controlling anther height is comparatively lower. The relatively low expression of CYP^T and KFB^T could also be because they are phytohormone-related; these chemical messengers are produced in very low concentrations (Wani et al., 2016).

It might otherwise be that the RNA samples were taken at a developmental time point where the expression of these genes is relatively low, that some of the genes are expressed only in specific tissues, or it may be down to a dosage (or genomic copy number) related effect that results from the hemizygosity of the region, analogous to that affecting mammalian X chromosome genes present in only one copy in (XY) males, counteracted through X chromosome-inactivation in (XX) females (Avner and Heard, 2001, Graves and Disteche, 2007).

Excluding GLO^T, the vast majority of the genes differentially expressed between pin and thrum flowers have expression levels higher than the *S* locus genes themselves; these genes are potential downstream targets of the *S* locus (Figure 4.7). It therefore seems possible that some of the *S* locus genes are expressed at a relatively low level as they act as genetic switches to effectuate a cascade of more highly expressed genes in the downstream regulatory network; functioning as master regulators for the development of heterostyly-associated morphological features (Chan and Kyba, 2013).

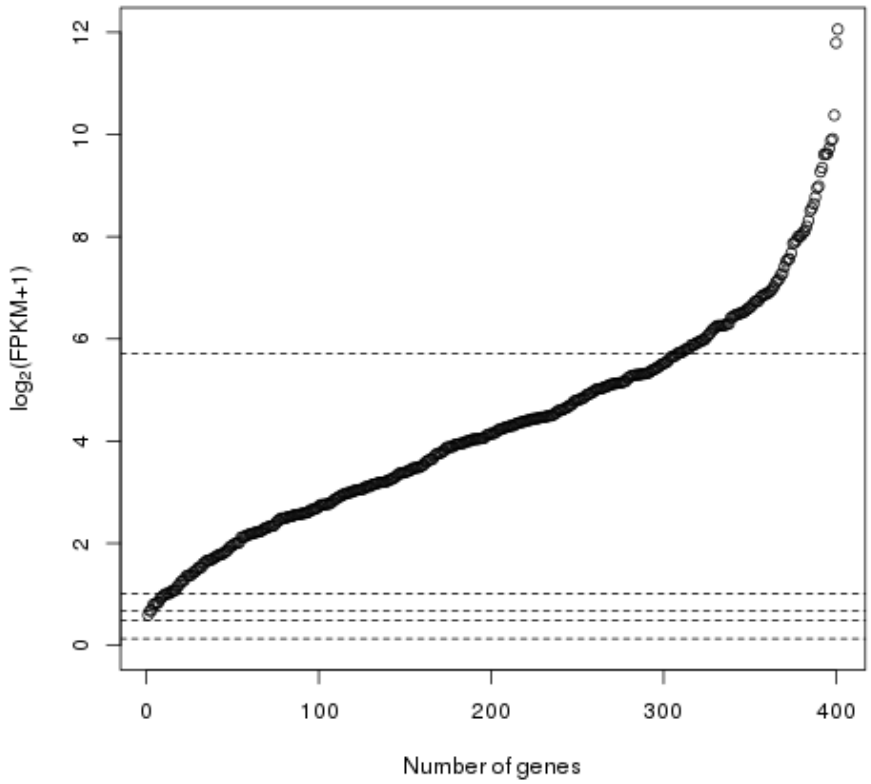


Figure 4.7 The $\log_2$(FPKM+1) expression of genes differentially expressed between pin and thrum flowers (see Chapter 2); expression is shown for either pin or thrum, whichever is higher. The dotted lines indicate expression levels in thrum for the five genes at the *S* locus, from top to bottom: GLO^T, CYP^T, PUM^T, KFB^T, CCM^T.

4.4.7 Identification of thrum-specific regions in the *P. vulgaris* genome

To establish whether the 278 kb region is the only thrum-specific region in the *P. vulgaris* genome, genomic reads from a pool of pin plants (LIB1730; Table 2.1) were mapped to the thrum (TP_v1) genome assembly to determine the depth and breadth of coverage across predicted TP_v1 transcripts expressed only in thrum flowers. The pool of pin plants used in this analysis were progeny from a cross between the thrum plant used for the TP_v1 assembly, and the pin plant used for the pin assemblies (PP_v1 and PP_v2); therefore minimising the number of SNPs between the pin pool read library (LIB1730; Table 2.1) and TP_v1, such that regions of low coverage were not erroneously identified.

If the 278 kb region is not the only thrum-specific region in the genome, then a hypothetical scenario such as that shown in Figure 4.8 may mean that the *S* locus comprises two thrum-specific regions, with an intervening region in both pin and thrum genomes.

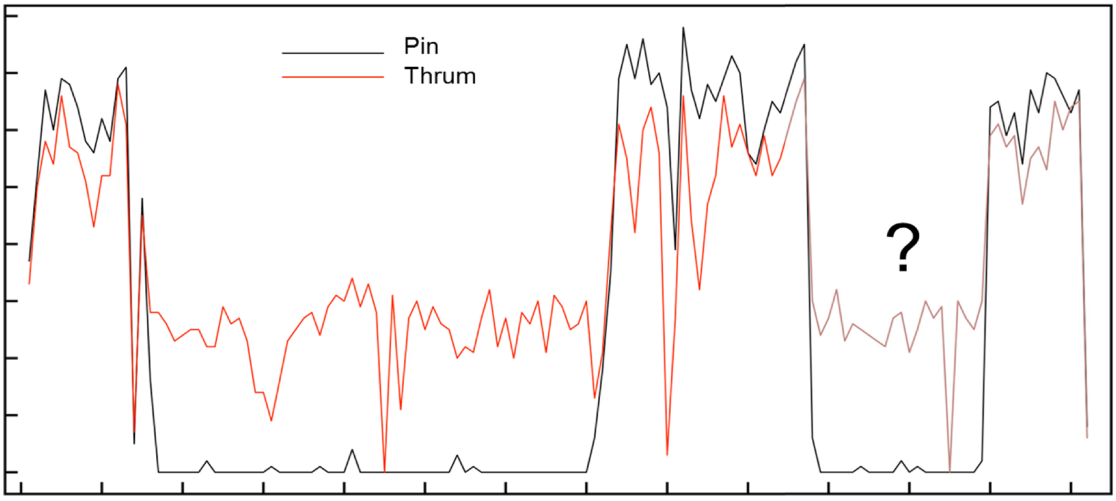


Figure 4.8 Diagram of a hypothetical composite *S* locus comprising two thrum-specific regions with a low depth and breadth of pin genomic read coverage: the identified 278 kb region (left), and a second hypothetical thrum-specific region (marked "?"), with an intervening region in both pin and thrum genomes that is characterised by a high depth of pin (and thrum) genomic read coverage (pin reads = black, thrum reads = red, as indicated).

The short-read alignment strategy employed means that some reads may be erroneously mapped to thrum-specific contigs (Qu et al., 2009), as such one would expect two distinct groups of transcripts with different coverage profiles associated with them: transcripts with a high depth and breadth of associated pin read coverage that are present in the pin genome, and transcripts with a distinctly lower depth and breadth of pin read coverage that are absent from the pin genome, where the low coverage comprises pin reads that are erroneously mapped, as can be seen for pin reads across the 278 kb thrum-specific region (Figure 4.2).

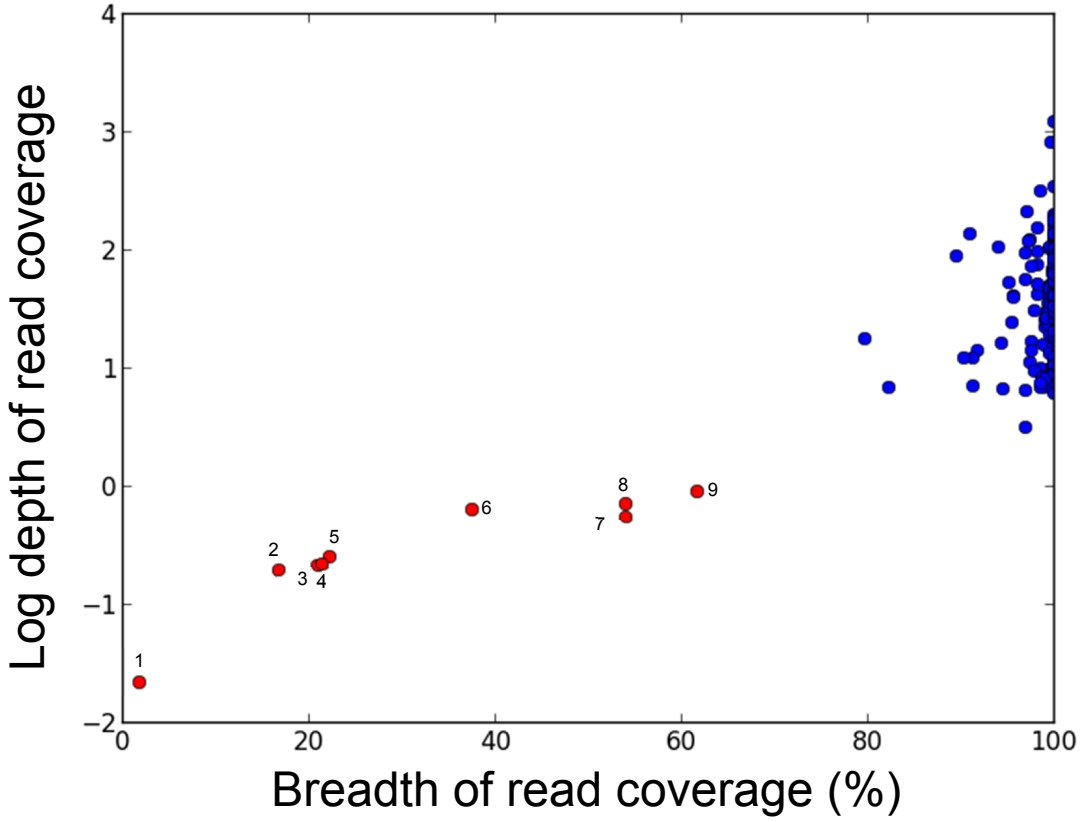


Figure 4.9 The breadth (%) and $\log_{10}$-transformed depth of pin pool genomic read coverage across predicted transcripts in the thrum genome (TP_v1) that are expressed only in thrum flowers (FPKM < 0.1 in pin flowers), with red (thrum-specific; absent in pin genome) and blue (present in pin genome) groups of data points assigned based on clustering with *k*-means; transcripts in the red group with low depth and breadth of pin read coverage represent the following genes: KFB^T (1), GLO^T (2), CYP^T (3), GLO^T (4), KFB^T (5), PUM^T (6), GLO^T (7), CYP^T (8), KFB^T (9) (Table 4.2).

The analysis revealed one group of transcripts with a high associated depth and breadth of pin pool read coverage, and another comprising nine transcript regions with a distinctly lower associated depth and breadth of pin read coverage, as ascertained by analysis with *k*-means clustering (Figure 4.9). The transcript sequences absent from the pin genome are represented by the data points with a low pin read coverage; these transcripts align to four of the five thrum-specific genes from the 278 kb region: GLO^T, CYP^T, PUM^T and KFB^T (Table 4.2). CCM^T is expressed at a low level (Figure 4.7) and is the only gene in the 278 kb region not represented.

The thrum TP_v1 genome assembly was used in this analysis as opposed to LH_v2 as it was previously predicted that long homostyles were the result of recombination between genes at the *S* locus (Lewis, 1954, Dowrick, 1956). The LH_v2 genome cannot be used in an analysis to rule out the presence of additional thrum-specific regions, as under the assumption of recombination, part of the *S* locus in the long homostyle would be of the pin genotype, and thus present in the pin genome. This suggests that a second thrum-specific region, if it exists, might not be identified in the long homostyle due to potential recombination in an intervening pin-thrum genomic region that lies between the two thrum-specific regions. The identification of three contigs for GLO^T and KFB^T, and two for CYP^T, is due to the use of the non-scaffolded thrum TP_v1 genome assembly with smaller average contig size than LH_v2, as well as the large size of GLO^T and CYP^T. The non-scaffolded thrum assembly was used to avoid suggestions that contigs might be incorrectly joined together through scaffolding; this could potentially result in amalgamated gene structures, part thrum-specific and part not, resulting in thrum-specific transcripts not being highlighted as having an overall low coverage of pin reads associated with them.

Transcript no.	TP_v1 contig ID	Coverage			*S* locus gene
		Breadth	Depth	Log depth	
1	TP_v1_3432270	1.73	0.02	-1.67	KFB^T
2	TP_v1_3559018	16.76	0.19	-0.72	GLO^T
3	TP_v1_3534674	20.90	0.21	-0.68	CYP^T
4	TP_v1_3579940	21.48	0.22	-0.67	GLO^T
5	TP_v1_3103680	22.30	0.25	-0.60	KFB^T
6	TP_v1_3432270	37.64	0.62	-0.20	PUM^T
7	TP_v1_3526291	54.01	0.54	-0.27	CYP^T
8	TP_v1_3554783	53.98	0.70	-0.15	GLO^T
9	TP_v1_3103680	61.84	0.90	-0.05	KFB^T

Table 4.2 Transcripts identified in the thrum TP_v1 genome as thrum-genome specific, and the *S* locus genes to which they align. The transcript numbers refer to those shown in Figure 4.9. The depth, $\log_{10}$(depth) and breadth of genomic pin pool read coverage (Figure 4.9) is shown for each transcript.

The data presented here identify the 278 kb region as the only thrum-specific region in the *Primula vulgaris* genome containing flower-expressed genes; there are no additional flower-expressed transcripts that are specific to the thrum genome as compared to the 278 kb region identified in the long homostyle (LH_v2).

4.4.8 Recombination in regions flanking the *S* locus

The signal of recombination was analysed in the regions flanking the *S* locus to determine whether these sequences and any regions beyond them could be part of the *S* locus. The analyses above show that the 278 kb thrum-specific region identified is the only thrum-specific region in the *Primula vulgaris* genome, but it remains a possibility that the *S* locus could be part of a wider non-recombining region; a hybrid between a thrum-specific region and a region present in both pin and thrum genomes that contains genes with pin- and thrum-specific alleles. In such a region, recombination between pin and thrum must be suppressed in order to maintain the integrity and functionality of the *S* locus.

If the regions flanking the *S* locus are freely recombining then these regions cannot form part of the *S* locus. Recombination would disrupt linkage between the genetic components of the supergene. The 278 kb thrum-specific region is linked to the *S* locus. Thus, any region that is part of the *S* locus must include the 278 kb sequence that co-segregates with the thrum phenotype.

The recombination studies performed analysed the distribution of SNPs across the flanking sequence regions, comparing alleles in a pin and thrum plant, to identify any signatures of recombination marked by regions of significantly reduced polymorphism. The first analysis (Figure 4.10) considers all sites with SNPs between the four alleles present in an individual pin and thrum plant. This includes sites heterozygous in the thrum plant, which represent a SNP between the pin and thrum allele present in thrum at that position (for example, where the supporting mapped reads contain nucleotides A and T at that position), and secondly, sites homozygous in thrum, with SNPs identified between either of the pin alleles present in the pin plant and the nucleotide present in thrum at that site (for example, pin = AA, thrum = TT represents a SNP between thrum and pin); sites with a SNP between pin and thrum as determined by the above scenarios are considered in Figure 4.10.

The above analysis shows that regions of significantly reduced polymorphism between pin and thrum are present in the flanking regions; this is the hallmark of recent recombination between pin and thrum, where the sequences are homogenised through active genetic exchange. The absence of SNPs is an indication that the sequence under consideration has been exchanged recently relative to the divergence seen elsewhere (López-Pérez et al., 2014). The homogeneous regions with a significantly reduced number of SNPs (Figure 4.10) include both genic (e.g. *R5; SFG5*^{R}) and intergenic regions (e.g. between *R1* and *R2; SFG1*^{R} and *SFG2*^{R}). This suggests that the increased sequence similarity is not just the result of strong purifying selection against non-synonymous mutations.

The next analysis (Figure 4.11) was performed as above, but excludes exons to account for conserved sequences that might show reduced polymorphism between thrum and pin due to sequence constraints. To extend this further, genic regions including exons and introns were omitted in a third analysis (Figure 4.12). Both of these analyses show regions of reduced polymorphism, therefore demonstrating that it is recombination that

is homogenising the sequences, not selection. The same analyses using only heterozygous sites in thrum yielded similar results (figures not shown) (Li et al., 2016).

In conclusion, this analysis shows that the 278 kb thrum-specific region comprises the entire *S* locus; sequences beyond this thrum-linked region are actively undergoing recombination, and thus cannot be a part of the *S* locus that determines pin and thrum floral phenotypes.

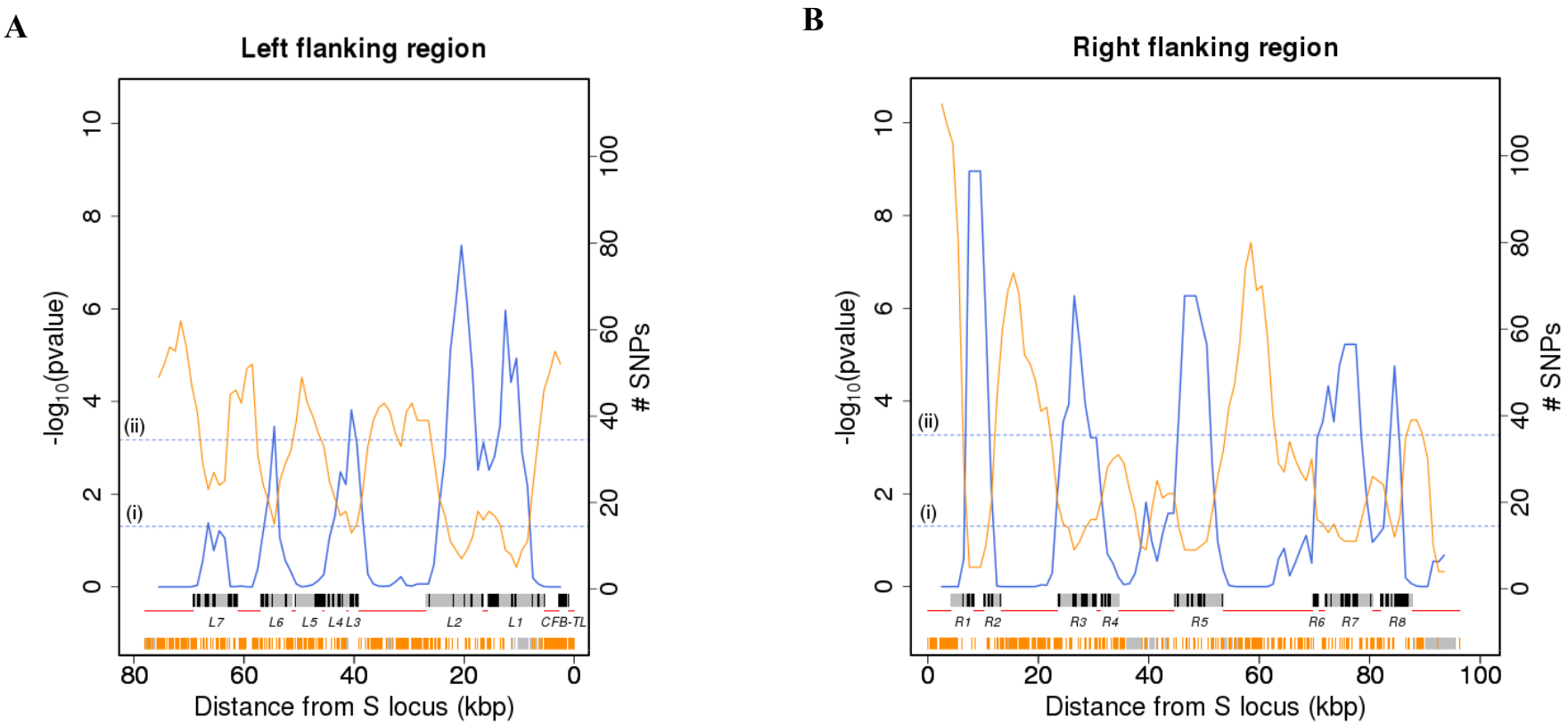


Figure 4.10 The number of SNPs (# SNPs) between pin and homozygous sites in thrum, and heterozygous sites in thrum, within 5 kb sliding windows (orange line) across the sequences to the right (A) and left (B) of the 278 kb thrum-specific region. The cumulative binomial probability ($-\log_{10}$(pvalue)) of observing the number of SNPs shown (or fewer) given the frequency of SNPs in the flanking sequence as a whole (blue line); horizontal dotted lines indicate critical values (i) $-\log_{10}(p=0.05)$ (uncorrected) and (ii) $\log_{10}(p=0.05)$ with Bonferroni correction for multiple comparisons based on the number of windows analysed. The locations of individual SNPs are shown by orange vertical bars; unresolved bases represented by Ns in the sequence, and sites excluded based on depth and quality cut-offs for SNP calling, were omitted and are indicated by vertical grey bars alongside the orange SNP bars. The labels *R1-8* and *L1-7* refer to the eight *SFG1-8*^{R} (*S* locus *flanking gene-right*) and seven *SFG1-8*^{L} (*S* locus *flanking gene-left*) genes predicted in these regions of the LH_v2 genome, the left flanking duplicated *CFB*^{TL} locus (*CFB-TL)* is also shown (see Figure 4.2), introns (grey bars), exons (black bars), with intergenic regions shown by red lines; in some cases indicated gene features (grey/black) and SNPs/omitted sites (orange/grey) in close proximity cannot be distinguished due to insufficient resolution.

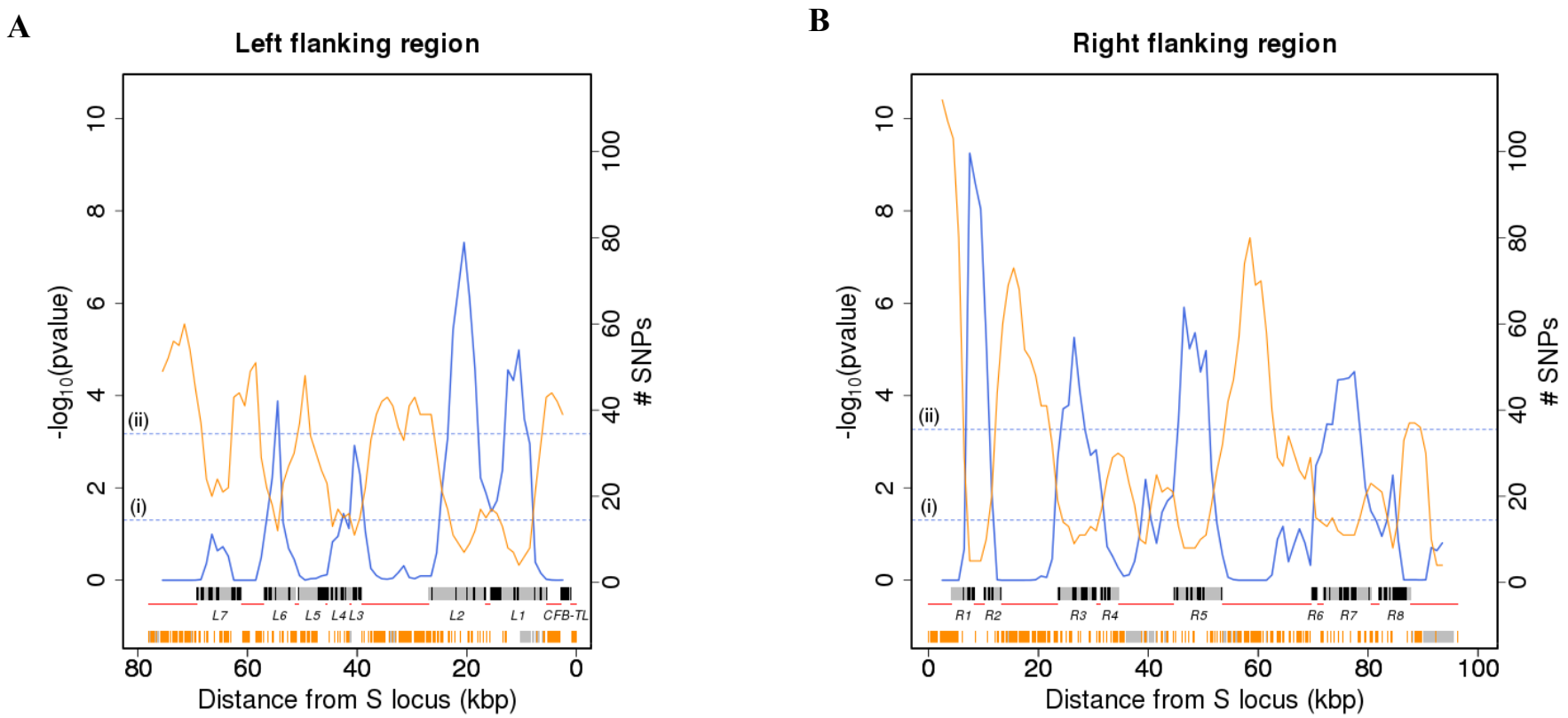


Figure 4.11 Analysis of recombination in *S* locus flanking sequences with coding sequences (exons) excluded. The number of SNPs (# SNPs) between pin and homozygous sites in thrum, and heterozygous sites in thrum, within 5 kb sliding windows (orange line) across the sequences to the right (A) and left (B) of the 278 kb thrum-specific region, alongside the cumulative binomial probability ($-\log_{10}$(pvalue)) of observing the number of SNPs shown (or fewer) (blue line); (i) and (ii) indicate uncorrected and Bonferroni corrected critical values, respectively. Presented as in Figure 4.10 but with coding sequences (black on grey bars) and associated SNPs excluded.

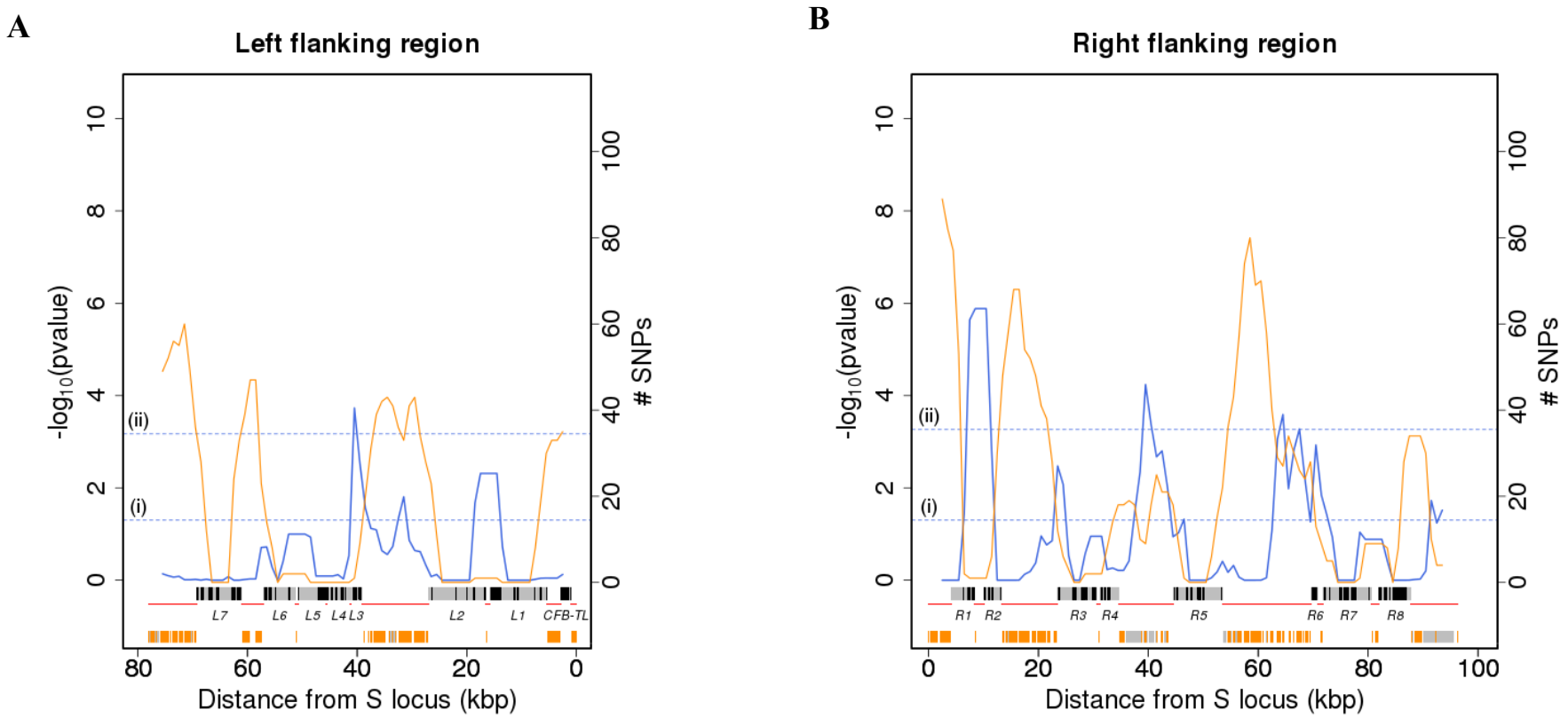


Figure 4.12 Analysis of recombination in *S* locus flanking sequences, excluding genic regions (UTRs, coding sequences (exons) and introns). The number of SNPs (# SNPs) between pin and homozygous sites in thrum, and heterozygous sites in thrum, within 5 kb sliding windows (orange line) across the sequences to the right (A) and left (B) of the 278 kb thrum-specific region, alongside the cumulative binomial probability (-$\log_{10}$(pvalue)) of observing the number of SNPs shown (or fewer) (blue line); (i) and (ii) indicate uncorrected and Bonferroni corrected critical values, respectively. Presented as in Figure 4.10 but with genic sequences (black and grey bars) and associated SNPs excluded.

4.4.9 The annotation of miRNAs in the *Primula vulgaris S* locus

MicroRNAs (miRNAs) have been implicated in male and female reproductive structure development through the targeted regulation of genes involved in pollen, pistil and stamen development (Nagpal et al., 2005, Wu et al., 2006, Li et al., 2015b, Su et al., 2016). Having determined that the whole *S* locus lies within a 278 kb thrum-specific region, further annotation was carried out to locate putative miRNA sequences in this region.

The analysis identified potential mature miRNAs at 32 predicted overlapping pre-miRNA loci within the *P. vulgaris S* locus region. The pre-miRNAs were initially predicted based on similarity to known miRNAs in the miRBase database (Griffiths-Jones et al., 2006) as well as the presence of both miRNA and miRNA* sequences. This was followed by assessment of other pre-defined criteria using Novomir, including detection of a stem-loop structure as the lowest energy folding form in the pre-miRNA (Teune and Steger, 2010). The strategy used is similar to that implemented in miRNA annotation of the draft bread wheat (*Triticum aestivum*) genome (Mayer et al., 2014). The predicted targets in the LH_v2 gene annotations for the mature miRNAs predicted by Novomir were identified using psRobot; this software leverages the near perfect sequence complementary to target sequences that miRNAs exhibit (Wu et al., 2012). Six of the predicted mature miRNAs had predicted targets (Table 4.3).

miRNA ID	Target gene ID	Expression of target gene (FPKM)		Fold change FPKM	Functional annotation of target gene
		Pin flowers	Thrum flowers		
PvS_mir1	PvLHv1_203940	1.674	0.754	1.151	Ethylene-responsive transcription factor 1; IPR016177 (DNA-binding domain)
PvS_mir2	PvLHv1_040730	1.465	1.135	0.367	Pentatricopeptide repeat-containing protein; IPR002885 (Pentatricopeptide repeat); IPR011990 (Tetratricopeptide-like helical domain)
PvS_mir3	PvLHv1_087070	0.161	0.141	0.198	Ethylene-responsive transcription factor 1; IPR016177 (DNA-binding domain)
PvS_mir4	PvLHv1_236220	14.647	16.613	0.182	Unknown protein
PvS_mir5	PvLHv1_242980	8.025	8.516	0.086	Dof zinc finger protein; IPR003851 (Zinc finger, Dof-type)
PvS_mir6	PvLHv1_110120	36.340	36.519	0.007	DUF21 domain-containing protein; IPR000644 (CBS domain); IPR002550 (Domain of unknown function DUF21; GO:0005886 (duf21 domain-containing protein at1g47330-like)

Table 4.3 Putative microRNAs (miRNAs) in the *Primula vulgaris S* locus, and their predicted target genes in the *Primula vulgaris* LH_v2 genome annotations. Expression (FPKM) in pin and thrum flowers, and functional descriptions based on RNA-Seq analyses and functional annotations presented in Chapter 2 are shown for the target genes.

4.4.10 Similarity of genes at the *P. vulgaris S* locus to other genomic regions

Three of the transcript sequences associated with genes at the *S* locus (*CYP*^{T}, *PUM*^{T} and *KFB*^{T}) map via TBLASTX alignments to a second sequence region (on distinct contigs) with relatively low similarity (< 60% identity (ID)) in comparison to *GLO*^{T} (~70% ID) and *CCM*^{T} (~85% ID). The putative transposon sequences located at the *S* locus all align to many similar (> 80% ID) sequences in the genome, suggesting high copy numbers, and providing further evidence that they are TE-related due to repetitive sequence compositions (Li et al., 2016). The percent identity (pid) in all cases (Figure 4.13) is lower than might be anticipated due to lack of precision in the TBLASTX alignments in determining intron and exon boundaries.

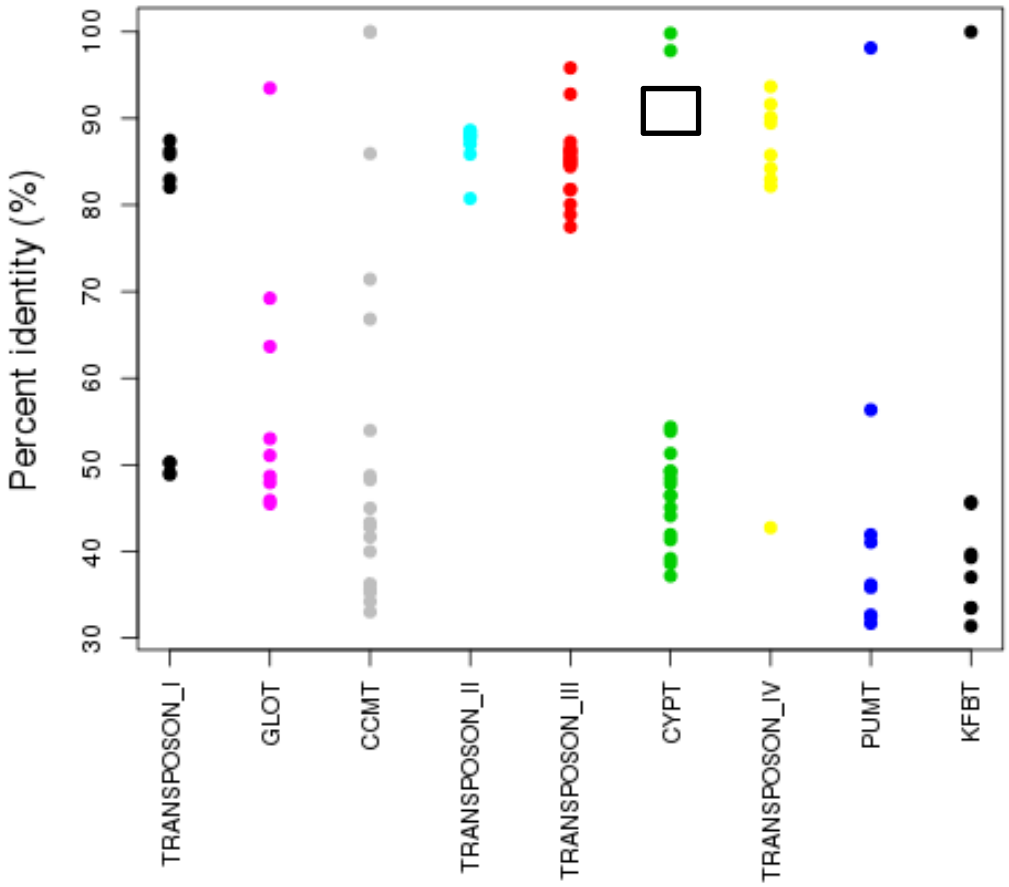


Figure 4.13 HSPs (High Scoring Pairs) identified via TBLASTX alignments of predicted transcripts for genes at the *Primula vulgaris S* locus to distinct contigs in the LH_v2 genome, are grouped into singular data points; gene names are indicated. Four putative transposon-like gene annotations located at the *S* locus are also shown (TRANSPOSON_I-IV). The alignments are against the whole genome, as such for each column the data point with the highest PID is a mapping against the locus associated with itself. The two uppermost green data points for *PvCYP*^{T} represent the two predicted transcript sequences that together form this gene (outlined by a black square), with associated alignments grouped into one column (coloured data points).

4.4.11 Repetitiveness of the *Primula S* locus

The 278 kb *S* locus contains a high proportion of repetitive sequences (64.07%) (top 5%) compared to both contig sequences in the long-homostyle (LH_v2) genome as a whole (37.03%) (Table **2.6**) (Figure 4.14) and contigs ±20% the length (278 kb) of the *S* locus (Figure 4.15).

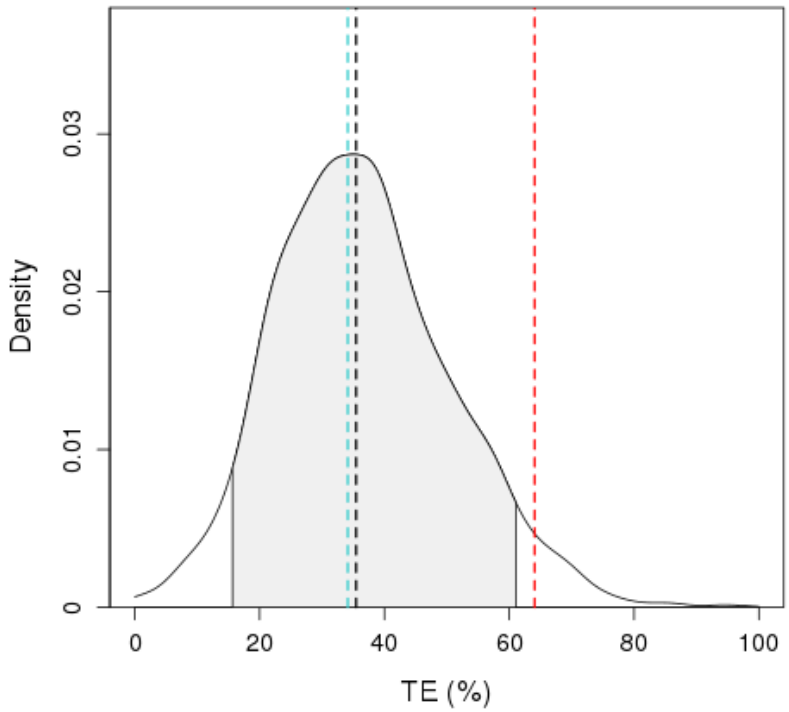


Figure 4.14 Density plot of Transposable Element (TE) (%) for contigs (>10kb) in the *P. vulgaris* LH_v2 genome assembly (n=2,409); dashed black line = median (35.47%); dashed red line = repeat percentage of the *P. vulgaris S* locus (64.07%); dashed teal line = repeat percentage (34.13%) of concatenated left and right *P. vulgaris S* locus flanking regions (174 kb); shaded area (grey) = 95% quantile.

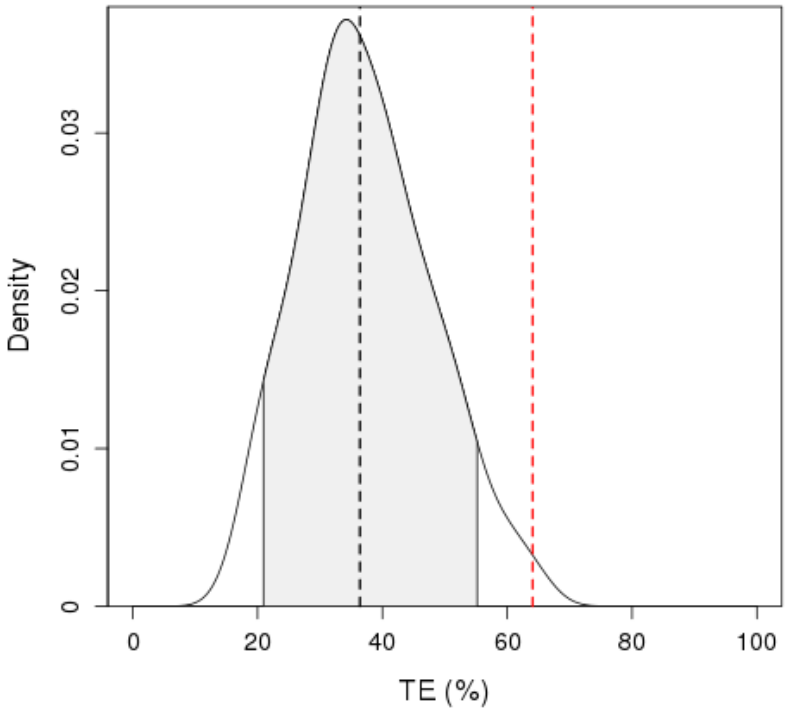


Figure 4.15 as in Figure 4.14 for contigs ±20% the length (278 kb) of the *P. vulgaris S* locus (n=218); median (dashed black line) = 36.39%.

4.4.12 Intron sizes at the *Primula S* locus

Two of the five predicted genes at the *S* locus (*GLO*^{T} and *CYP*^{T}) have particularly large introns (>10 kb and 30 kb, respectively) (Figure 4.16).

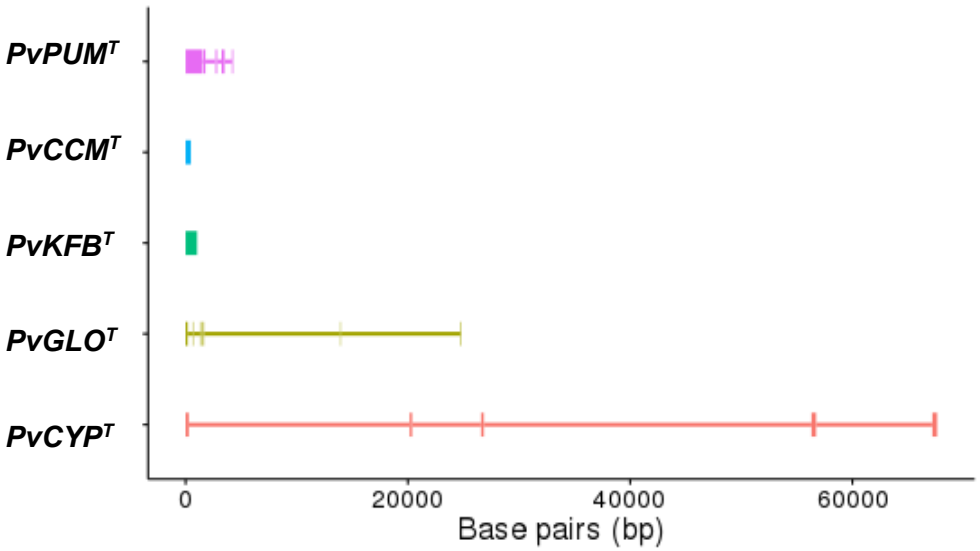


Figure 4.16 The gene structures of the five genes at the *P. vulgaris S* locus with fully expanded introns, exons = thick lines, introns = thin lines.

Interestingly, when considering intron sizes across the genome (Figure 4.17) the six introns over 5 kb are in the top 5% of intron lengths.

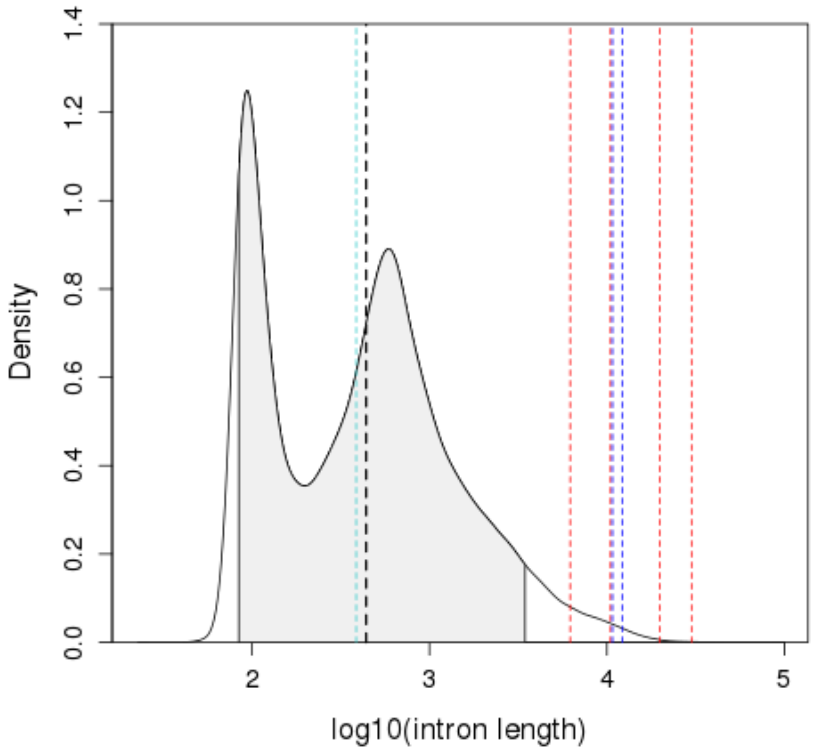


Figure 4.17 Density plot of $\log_{10}$-transformed lengths (bp) of *P. vulgaris* introns (n=133,334); dashed lines = median $\log_{10}$(intron length) (black), median $\log_{10}$(intron length) in regions flanking the left and right of the *S* locus (teal), and $\log_{10}$(length) of *P. vulgaris S* locus introns > 5 kb: GLO^T (blue) (12172 bp, 10755 bp), CYP^T (red) (19870 bp, 6205 bp, 29971 bp, 10469 bp); the difference in intron lengths between this group and those genome-wide is significant ($p < 2.789 \times 10^{-5}$) (Wilcoxon Rank Sum).

4.5 Discussion

The analyses presented in this chapter reveal the *Primula vulgaris S* locus as a 278 kb thrum-specific region, containing a cluster of five genes expressed in thrum flowers. This is the only region in the *Primula vulgaris* genome that is specific to thrums; furthermore, sequences beyond the 278 kb *S* locus are not in recombinational isolation as they contain significant stretches of similarity between pin and thrum, and thus cannot form part of the *S* locus. This is the first account for any species of the *S* locus supergene that controls heterostyly.

The *Primula S* locus has long been defined as a supergene (Dowrick, 1956, Mather, 1950), but it is unlike those subsequently characterised for butterfly mimicry, and avian and insect social behaviour (Thomas et al., 2008, Joron et al., 2011, Wang et al., 2013). The supergene controlling butterfly mimicry for example comprises a single gene, splice variants of which bring about specific wing type-associated polymorphisms that allow the female sex to mimic a toxic species (Kunte et al., 2014). In contrast, the supergene that controls heterostyly in *Primula* comprises multiple genes, and is therefore arguably closer to a "supergene" under the original definition: "a group of genes acting as a mechanical unit in particular allelic combinations" (Darlington and Mather, 1949, Thompson and Jiggins, 2014), except in this case the pin (*s*) haplotype is a null allele.

The polymorphisms associated with the supergenes involved in the control of butterfly mimicry, and avian and insect social behaviour are maintained in recombinational isolation by chromosomal inversions, such that localised reductions in recombination result in high levels of genetic divergence as compared to collinear chromosomal regions (Thomas et al., 2008, Joron et al., 2011, Wang et al., 2013, Thompson and Jiggins, 2014). Mather (1950) noted that the breakdown of heterostyly in *Primula sinensis* appeared to be different to that in *Primula viscosa* (Ernst, 1936c) with no abnormal allelomorphs of the supergene being observed despite the breeding of over 10,000 plants. Ernst (1936c) documented numerous self-compatible homostyle plants with "anomalous combinations" of the stigmatic- and anther-associated characters. This apparent lack of disruption in the *P. sinensis* supergene led Mather (1950) to conclude that perhaps occasional recombination could occur in this species, despite lack of observed homostyles, or that a chromosomal inversion might otherwise result in the constituent genes being "tied up once and for all into a super-gene".

This chapter reveals that the integrity and tight linkage of the cluster of genes is not maintained through suppression of recombination by a proposed mechanism such as chromosomal inversion or proximity to the centromere (Mather, 1950, Thompson and Jiggins, 2014), but by the hemizygosity of the region in thrums, thereby precluding recombination and ensuring the dominance of the *S* haplotype that was defined by Bateson and Gregory (1905). It was predicted that homostyles were the result of recombination within the *S* locus supergene (Lewis, 1954, Dowrick, 1956), but the 278 kb region that defines the *S* locus is completely absent in pins, *s* is a null haplotype; as such there is simply nothing to recombine with, homostyles are not the result of recombination. Furthermore, predictions on the order and number of genes at the *S* locus based on the observation of recombinants have no foundation (Lewis, 1954, Dowrick, 1956, Lewis and Jones, 1992); the *S* locus contains five thrum-specific genes that are expressed in thrum flowers, not three as predicted.

These findings show that Ernst (1928a) was correct in his original conclusion that homostyles were the result of mutation, and explains the observed rarity of putative recombinants in *Primula* populations by Mather (1950) and others. CYP^T and GLO^T contain mutations in the long and short homostyle, respectively, which earmarks these genes as candidates for the *G* and *A* function genes. The transposon insertion in the second exon of GLO^T also points towards an example of how "anomalous" transitions from recessive to dominant alleles at the *S* locus might have occurred, as documented by Ernst (1957) and Lewis and Jones (1992). The majority of these documented transitions result from *GPa*/*gpa* and *gpA*/*gpa* plants that we would now classify as mutants based on the current study (*GPa*/- and *gpA*/-). The recovery of the thrum haplotype in such plants could perhaps be explained by retro-transposon excision if the *GPa* haplotype was a result of retro-transposon insertion in GLO^T for example. In conclusion, despite the suggestion by Mather (1950) that chromosomal inversion might be a potential mechanism preventing recombination in the *S* locus supergene, it is in fact the hemizygosity of the 278 kb region in thrums that means the linkage and co-segregation of genes in the *S* haplotype is guaranteed.

The finding that the *S* locus is hemizygous in thrums (*GPA* / -) is a curious insight indeed, not least because it has been widely accepted for so many years that thrums are heterozygous (*S*/*s*), and pins homozygous (*s*/*s*) (Bateson and Gregory, 1905). It appears

that this is the first autosomal hemizygous region to be identified in a "wildtype" organism on the molecular level; hemizygosity can otherwise result from chromosome aberrations associated with conditions such as leukaemia (Kees et al., 2005), and microdeletions that bring about phenotypic abnormalities through the unmasking of recessive deleterious alleles (Halder et al., 2010) but it seems that a "wildtype" organism with a hemizygous region that determines the fundamental morphology of the organism must be an uncommon occurrence at the very least.

Studies suggest that the phenomenon of apomixis may also be under the control of a hemizygous locus; apomixis is a form of asexual reproduction in plants that results in embryo formation without the need for fertilization, the progeny are genetically identical to the maternal parent, which means apomixis has potential applications in the breeding and propagation of crop plants (Ozias-Akins et al., 1998, Goel et al., 2003). The form of apomixis that occurs in the grass-family genus *Pennisetum* is termed apospory; this taxon contains the sexually-reproducing and agronomically-important pearl millet (*Pennisetum glaucum*) (Ozias-Akins et al., 1998). In crosses between pearl millet and *Pennisetum squamulatum*, an obligate-aposporous wild relative, markers linked to the apospory trait were shown to be hemizygous in the apomicitic progeny that can bypass the need for fertilization (Ozias-Akins et al., 1998). Such apomixis-linked markers have also been found in *Paspalum*, with no apparent allele on chromosomal homologs transmitted to the sexual offspring (Labombarda et al., 2002). Further studies using fluorescence *in situ* hybridization (FISH) probes comprising BAC clones of 80-100 kb confirm that genomic regions linked to the apomixis-determining region are present on only a single chromosome in apomictic *Pennisetum squamulatum* and *Cenchrus ciliaris* plants, and show that it lies in close proximity to the centromere, much like the *P. vulgaris S* locus (Goel et al., 2003). This suggests that the determining region is either highly polymorphic, or completely absent in sexually reproducing individuals, and has led to the apomixis-determining region being defined as the apomixis-specific genomic region (ASGR) (Goel et al., 2003). Despite this, the key ASGR genes have yet to be uncovered.

The *k*-means clustering analysis employed in the current study was used to distinguish between the distinct pin read coverage profiles associated with transcripts present in both pin and thrum genomes, and those present in thrum-specific regions that are completely absent from pin. This study successfully uncovered the 278 kb thrum-specific region as a

group of transcripts with a distinctly low associated coverage of pin genomic reads, thus highlighting this clustering approach as a possible method for the discovery of hemizygous regions in other genomes. If the ASGR is indeed hemizygous in apomictic individuals (Goel et al., 2003), then given sufficient genomic resources, the methods developed and presented in this chapter, such as the analysis of read-depth profiles and the *k* means strategy described above, could be useful tools in the isolation and characterisation of such a region; likewise, the genes responsible for heterostyly in other families could be identified in this manner, should their functional roles be underpinned by hemizygous genomic regions as in *Primula*.

The hemizygosity of the *S* locus in thrums invites comparison with the XY chromosome system in mammals, where the Y-chromosome specific region that comprises 95% of the chromosome's length undergoes no X-Y recombination and is surrounded on both sides by pseudo-autosomal regions where X-Y recombination is frequent (Skaletsky et al., 2003). It was proposed that this system evolved from an ordinary pair of autosomes, with the Y chromosome seen as a profoundly degenerate remnant of the X chromosome containing perhaps a single gene involved in sex determination (Skaletsky et al., 2003); this hypothetical inducer of testes formation was subsequently termed the testis-determining factor (TDF) (Harley et al., 2003, Skaletsky et al., 2003). The Y chromosome has since been characterised as repeat-rich and contains very few protein-coding genes, including the sex-determining gene *SRY* (sex-determining region Y) which was identified as the theoretical TDF through analysing the genomes of sex-reversed patients where small fragments of the Y chromosome had translocated to the X chromosome (Sinclair et al., 1990, Waters et al., 2007). This male-specific gene is the only gene required for testes development based on male XX mice transgenic for *SRY* (Koopman et al., 1991), and is dominant due to human males being the heterogametic (XY) sex, much like the dominant genes of the *S* locus in the hemizygous thrum (*S*/*s*) (Skaletsky et al., 2003, Waters et al., 2007). Interestingly, Darwin drew comparisons with the male and female sexes to first explain the morphologies of the pin and thrum flowers in *Primula*, suggesting the pin was more feminine like with its long style and short anthers with small pollen grains (Darwin, 1877). He first proposed that the two forms of *Primula* flower were in the process of evolving towards dioecy, but after observing that the alleged female pin-form did not produce more seeds than the thrum, he ultimately concluded that the benefit of the

arrangement was in promoting inter-crossing between distinct plants. He was of course correct in this final deduction, but it is perhaps fitting that the thrum is hemizygous for thrum-specific genes, much like a large portion of the Y chromosome in male humans, or vice versa in species where females are the heterogametic sex.

The differentiation of the Y chromosome proceeded after acquisition of the *SRY* gene, resulting in a non-recombining region and the degeneration of the Y chromosome through accumulation of mutations and deletions (Waters et al., 2007). This led to a prediction that the human Y chromosome could go extinct within 10 million years (Aitken and Marshall Graves, 2002). The above is confounded by the following: complete absence of recombination is thought to result in genetic degeneration due to increased mutational load in hemizygous regions of the Y chromosome. Likewise, the thrum is unable to undergo recombinational repair as homozygous thrums are inviable in most *Primula* species (Kurian and Richards, 1997). If there is no ability to recombine to produce favourable combinations of alleles, then highly fit genotypes cannot be recovered; like asexual organisms, each hemizygous thrum individual will inherit a full load of mildly deleterious mutations from its predecessor, effectively a clone of the parental *S* locus. Due to genetic drift the class of individuals with a minimal quantity of mildly deleterious mutations could be lost, leading towards ever more mildly deleterious mutations "hitchhiking" on beneficial mutations, and gradual deterioration. Increased genetic drift due to the effective population size of the *S* allele being a quarter of that for unlinked regions, or the Y chromosome being a quarter of that for autosomes, will increase this effect (Xu et al., 2011, Johnson and Lachance, 2012).

However, it has been shown that 25% of the male-specific euchromatic DNA on the human Y chromosome appears in eight palindromic repeats with over 99.9% sequence identity, encompassing many testis-specific genes. These repeats are maintained by intra-chromosomal gene conversion and facilitate the restoration of deleterious mutations by replacement with mutation-free gene copies (Rozen et al., 2003, Bachtrog, 2013). The results presented in this chapter suggest that the *S* locus region is unlike the Y chromosome in this respect; the *P. vulgaris* genome does not contain any almost-identical palindromic repeats that encompass the constituent genes of the *S* locus region, which leads one to question how its integrity is maintained in lieu of intra-chromosomal gene conversion. The *S* locus cannot otherwise undergo repair via recombination as there is no region with which

to recombine in hemizygous thrums. The repair of DNA lesions for example, is essential to maintain genome integrity (Branzei and Foiani, 2008). Double-stranded breaks (DSBs) in DNA that are associated with replication forks are typically repaired by homologous recombination (HR) in mammalian cells (Arnaudeau et al., 2001); an alternative method of repair is the non-homologous end-joining (NHEJ) pathway that results in ligation of double-stranded ends (Ray and Langer, 2002). It has been suggested that NHEJ and HR may compete with each other for the repair of DSBs (Sonoda et al., 2006, Lieber, 2010). The majority of exogenously induced DSBs are repaired by NHEJ in mammals, but HR-deficient mutants are associated with an increased occurrence of chromosomal breaks, suggesting HR may be essential for the repair of a different type of DSB, such as those resulting from stalled replication forks during the DNA replication (S-phase) of the cell cycle (Sonoda et al., 2006).

The human Y chromosome has many repeat sequences with diverse origins (Graves, 2006). This chapter reveals the *S* locus in *Primula vulgaris* as a particularly repetitive region in comparison to the rest of the *P. vulgaris* genome, and contains genes with confirmed intron sizes of up to 30 kb; notwithstanding any errors in the genome-wide gene annotations, this represents one of the largest intron sizes in the *P. vulgaris* genome and is larger than the longest introns in the tomato, rice and *Arabidopsis* gene databases (https://solgenomics.net/; http://rice.plantbiology.msu.edu/; https://www.arabidopsis.org/). Perhaps then, the error-prone repair of DSBs through the preclusion of repair by homologous recombination is a possible source for these large introns due to associated insertions. HR is generally a high-fidelity repair mechanism (Brenneman et al., 2002). In the NHEJ pathway, ligation of double-stranded ends can occur irrespective of whether the ends are from the same genomic region, and so frequently results in deletions, mispairing and translocations (Brenneman et al., 2002, Sonoda et al., 2006, Branzei and Foiani, 2008). Indeed, evidence suggests that novel introns could result from error-prone repair of DSBs through NHEJ (Li et al., 2009c, Farlow et al., 2011). The *P. vulgaris S* locus is apparently unable to undergo repair by homologous recombination, and this could be one way in which it has seemingly accumulated such large introns and an abundance of repeat sequences.

Despite this, the integrity of the *S* locus genes and thrum haplotype has been maintained such that heterostyly persists. If a deleterious mutation occurs in a heterostyly-associated

gene such as *CYP*^{T} or *GLO*^{T}, then a homostyle phenotype is produced. It may be that inbreeding depression does indeed infer a sufficient selective disadvantage on homostylous plants such that mutations leading to loss of heterostyly are purged from the population; this would explain the rarity of homostyles (Dowrick, 1956). If a homostyle population is established and maintained, however, then if *S** denotes a short or long homostyle haplotype of the *S* locus with mutated *GLO*^{T} or *CYP*^{T}, in a compatible cross between a homostyle plant homozygous at the *S* locus (*S**/*S**) and a pin plant (*s*/*s*), the result is a homostyle (*S**/*s*) where *s* is a null allele, whilst crossing with a thrum would result in *S*/*S**, therefore allowing recombination between *S* and *S** to take place. In this way, perhaps the 278 kb *S* haplotype could be repaired through recombination with the *S** haplotype such that the *S* locus is maintained.

If the above scenario occurs, then perhaps this is inefficient considering the presence of such large introns in *CYP*^{T} and *GLO*^{T}. On the other hand, it may explain the rare persistence of homostyles due to the repair of the *S* locus instead of deterioration; even without hemizygosity it has been predicted that they should spread and replace pins and thrums, so their rarity is surprising, particularly considering an apparent colonisation advantage due to self-fertility for the homostylous species *P. scotica* and *P. magellanica* from the Orkney and Falkland Islands (Charlesworth and Charlesworth, 1979a, Guggisberg et al., 2009). There is thought to be an increased likelihood of homostyle selfing such that outcrossing as the male parent occurs more frequently. The long homostyle as the male parent is dominant in compatible crosses with pin, whereas in the compatible cross with thrum the short homostyle is not (see introduction). This is thought to result in a deficiency of short homostyles as they cannot spread in a heterostylous population as efficiently; recessive mutations spread much more slowly than dominant ones (Charlesworth and Charlesworth, 1979a, Ganders, 1979). However, in comparison to long homostyles, short homostyles are extremely rare indeed; this phenotype has seemingly been fixed only once in *Primula* (Richards, 2014). Perhaps the findings in this chapter further support that rarity; in crosses with a homostyle as male parent, recombination could take place with thrum to disrupt the short homostyle haplotype. If the long homostyle is only compatible with pin, however, then recombination cannot occur due to excision of the *S* locus region.

The above suggestions are highly speculative and will require further thought; if nothing else they serve to reinforce the idea that the hemizygosity of the *S* locus in thrums will

provoke a great deal of discussion regarding its maintenance. To begin to investigate the maintenance of the *S* locus and the possibility of heterostyly being maintained by homostyles, the sequence capture approach (Chapter 1) could be used isolate the *S* locus from an array of species and distinct *P. vulgaris* populations. Polymorphisms between the *S* locus sequences, and between that and the flanking regions can be compared, to see if sequences across different populations are highly dissimilar, and determine the extent of degeneration to infer whether repair by homologous recombination might take place.

The annotation of the 278 kb *S* locus region was extended to include six putative miRNAs that have predicted targets within the *Primula vulgaris* LH_v2 geneset; two of these predicted gene targets are functionally annotated as ethylene-responsive transcription factors through similarity to *Ethylene response factor 1* (*ERF1*), and contain DNA binding domains functionally related to the AP2/ERF (APETALA 2/ethylene-responsive element binding factor) domain. This suggests they may have roles in mediating ethylene-related responses (Mizoi et al., 2012). AP2/ERF genes and the gaseous phytohormone ethylene (ET) have roles in the regulation of plant growth and development, as well as a variety of stress responses (Ohta et al., 2000, Saleh and Pagés, 2003, Xu et al., 2013, Wani et al., 2016). The rice ERF protein Submergence 1A (Sub1A) for example, is responsive to ethylene and negatively regulates cell elongation and carbohydrate consumption by controlling the expression of other AP2/ERF family transcription factors (Mizoi et al., 2012). These two putative miRNAs are therefore interesting candidates for the control of heterostyly-associated morphological features.

The elucidation of the supergene controlling butterfly mimicry as a single gene (Kunte et al., 2014) is in stark contrast to the five-gene *S* locus region identified in the current study (Kunte et al., 2014). The *doublesex* gene is differentially spliced in male and female butterflies, but not between female wing-pattern variants (Kunte et al., 2014). It has therefore been suggested that the intricate developmental control of the specific wing patterns associated with butterfly mimicry in the female may be the result of the several mutations found in the protein-coding sequences (Kunte et al., 2014), but perhaps more interestingly it has also been suggested that polymorphisms within non-genic regions of some butterfly species might act as *cis*-regulatory loci in the control of the wing-patterns (Supple et al., 2013, Wallbank et al., 2016). The miRNAs annotated here may prove to be involved in the specification of heterostyly-associated features. Perhaps the region

controlling butterfly mimicry might also harbour small RNAs of potential functional relevance, thus the single gene of the butterfly-mimicry supergene may not act alone.

The first step in validating the miRNA targets could be to show anti-correlation of the expression level of the target mRNA with the accumulation pattern of the corresponding regulatory miRNA (Wang et al., 2004). In the case of direct-cleavage of the target as opposed to downregulation, cleavage occurs towards the centre of the base-pairing interaction with the miRNA and results in a fragment characterised by the presence of a 5' phosphate group; the precise expected product size can be detected using a modified RNA ligase-mediated 5' rapid amplification of cDNA ends (5' RACE) procedure to confirm the miRNA target (Wang, 2004). For one of the miRNAs with a predicted *ERF*-like gene target, the expression of the target gene is downregulated in thrum by more than two-fold, therefore designating this particular miRNA as an interesting candidate for controlling an aspect of floral heteromorphy.

The study of regulatory networks downstream of the key genes at the *S* locus through identification of their regulatory targets will be an important task in defining the functional role of the *S* locus, and could be accelerated through the comparative analysis of RNA-Seq data from homostyle plants by analysing genes that are differentially expressed in comparison to pin and thrum when CYP^T and GLO^T are disrupted. The thrum-specific CCM^T gene identified at the *Primula S* locus contains an uncharacterised motif that is present in both mono and dicotyledonous angiosperms. This motif is present in *Petunia x hybrida* PIG93, which is a partner of the PSK8 protein that is associated with brassinosteroid-signalling (Verhoef et al., 2013). Brassinosteroids were first discovered when organic extracts from the pollen of western rape were seen to promote stem elongation and cell division when applied to stems (Grove et al., 1979). This proposes a link to Brassinosteroid signalling for CCM^T that is tentative but intuitive, given that the gene is located at the developmentally important *S* locus, and suggests that the thrum-flower expressed CCM^T could have a role in the downstream regulation of cell elongation in the development of heterostyly-associated morphological features. If this is the case, then considering the wide conservation of this motif throughout the angiosperms, it seems it could have an important but as yet unidentified role in plant development, and perhaps suggests that there remains an untapped opportunity to uncover important motifs through the comparative analysis of angiosperm genomes.

CCM^{T} (*Conserved Cysteine Motif)* could on the other hand have a role in the SI response: in Brassica (cabbage) the pollen SI determinant is a small cysteine-rich protein expressed exclusively in anthers; expression analyses of specific floral tissues and of the short and long homostyle might be a good starting place for deducing the genes responsible for SI. Indeed, that the homostyle plants with mutation in *GLO*^{T} or *CYP*^{T} are self-fertile suggests that the SI components act downstream of these master regulators, and therefore could be located anywhere in the genome. One speculative scenario might entail the female and male SI determinants for pin being located outside of the *S* locus, and thus present in thrum also. However, in thrum *GLO*^{T} and *CYP*^{T} could inhibit these factors at the same time as activating the thrum SI determinants elsewhere in the genome; or these thrum factors could be present at the *S* locus itself. If *GLO*^{T} is mutated for example, then the thrum pollen factor is no longer expressed and the short homostyle plant is self-fertile. In the above scenario this would mean that the short homostyle pollen is now incompatible with pin styles: perhaps homostyle crosses with pin could be used to distinguish whether this is case. The thrum and pin SI systems could well be different, and may have evolved as a by-product of interaction between pollen and pistil in response to selection for reduction in self-fertilization; speculation on different points of inhibition (see Section 1.2) is therefore feasible.

The discovery of the heterostyly-determining genes is itself an important finding, but also highlights the possibility of further studies, such as research into how the *S* locus is maintained given its interesting hemizygous architecture. These studies could serve as a blueprint for the genetic and downstream functional analysis of heterostyly in over 28 angiosperm families, as well as other pollination syndromes and functional regions underpinning biodiversity and food security. The established evolutionary-genetic models for the origin and breakdown of heterostyly are based on the assumption that thrums are heterozygous at the *S* locus (Charlesworth and Charlesworth, 1979a, Charlesworth and Charlesworth, 1979b, Lloyd and Webb, 1992a, Lloyd and Webb, 1992b); the studies presented in this chapter show that thrums are hemizygous for the *S* locus region. In addition to functional analysis of the key genes, a primary focus following these findings will be to elucidate the ancestral steps leading to the establishment of heterostyly, through comparative genomic and phylogenetic analyses of the identified genes, with a view towards establishing new data-driven evolutionary models.

5

Evolution and cross-species comparative analyses of the *S* locus

5.1 Relevant publications

Li, J.*, **Cocker, J.M.*,** Wright, J., Webster, M.A., McMullan, M., Ayling, S., Swarbreck, D., Caccamo, M., Oosterhout, Cv., Gilmartin, P.M. (2016) Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nature Plants*, 2: 16188.

Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Ayling, S., Caccamo, M., Gilmartin, P.M. (2017) The *Primula vulgaris* genome (in preparation).

* These authors contributed equally

5.2 Introduction

Primula vulgaris plants present pin or thrum flowers with reciprocally positioned reproductive structures that serve to physically promote insect-mediated outcrossing. The previous chapter described the *S* locus, which controls the development of the distinct heterostylous morphologies, as a 278 kb genomic region that is completely absent in pins; thrums are hemizygous for this region, suggesting self-fertile homostyles must result from mutation of the genes controlling the development of the two forms of flower, and not recombination between them as previously assumed (Dowrick, 1956, Lewis and Jones, 1992)

There are two major competing models for the evolution of heterostyly in *Primula* (Behnke et al., 2012). In the first, Charlesworth and Charlesworth (1979b), proposed that

the ancestral form was most likely a long homostyle, and that di-allelic self-incompatibility (SI) evolved prior to the onset of heterostyly. In the second, Lloyd and Webb (1992a) argue based on characteristic features of heterostylous groups that the most common basal morphology is that of approach-herkogamy; a pin-like architecture where the stigma is positioned above the anthers. This progenitor is thought to have given rise to reciprocal-herkogamy, which in *Primula* was followed by the evolution of di-allelic SI. Piper and Charlesworth (1986) presented some evidence for reduction of self-pollen on pin and thrum stigmas as compared to homostyles, which suggests that instead of SI being the first ancestral step, separation of the anthers and stigma in response to selection for reduced self-pollination is possible.

The findings in the previous chapter contradict the established models for the evolution of heterostyly (Charlesworth and Charlesworth, 1979b, Lloyd and Webb, 1992b) in that: (i) the genes controlling SI are apparently regulated by the genes at the *S* locus themselves: the self-incompatibility system which reinforces the constraints that heterostyly imposes is disrupted through mutation of those genes (GLO^T and CYP^T) in short and long homostyle (see Chapter 4), not recombination, suggesting the SI genes are downstream of the determining genes; they could be anywhere in the genome and are not necessarily linked to the *S* locus (Dowrick, 1956). In contrast to the Charlesworth and Charlesworth (1979b) model, SI presumably evolved after the establishment of heterostyly, perhaps as a consequence of coadaptation between the pollen and pistil that interact more frequently due to the structural constraints that heterostyly imposes; (ii) contrary to the Lloyd and Webb (1992) model it seems unlikely that an approach herkogamous (pin-like) flower with the stigma above the anthers could be the ancestral form as a greater evolutionary step is required to make the jump from pin to thrum based on the complete absence of the 278 kb *S* locus region in pins. In addition, although CCM^T and GLO^T have homologues outside of the *S* locus, there is an absence of similar genes for the three remaining genes elsewhere in the genome (Fig. 4.13) and, perhaps more convincingly, no large regions of similarity for the *S* locus as a whole. The most parsimonious ancestral step would therefore be a transition from thrum to pin, thus suggesting that whatever the most basal form in the Primulaceae sister clade is, whether it be a homostylous or herkogamous ancestor, thrum would most likely establish before or at the same time as pin. In light of the molecular and genomic data, it is pertinent to again begin analysing the ancestral steps leading to

heterostyly, such that an updated model for the evolution of this breeding system can be established.

The first step in such an approach might be to focus on the genes at the *S* locus itself and determine how the thrum haplotype came to be. *GLO*^{T} is an apparent duplication of the B-function MADS-box gene *GLOBOSA*, with which it shares 83% nucleotide sequence identity and 82% amino acid similarity. MADS-box genes such as *GLO* are expressed in the second (petal) and third (stamen) whorls of the flower (Tröbner et al., 1992) suggesting a gene that results from a duplication of *GLO* would be expressed in the same domain, and could subsequently be subject to neofunctionalization. This, coupled with the transposon insertion found in the short homostyle, as described in the previous chapter, suggests that *GLO*^{T} regulates anther elevation in the thrum form. In support of *GLO*^{T} haven taken on a novel function, the mutation of *GLO*^{T} is not complemented by ectopic expression of *GLO* in the short homostyle that is in a *Hose in Hose* background (Chapter 3) (Li et al., 2010). The above presents a unique opportunity to analyse the timing of the duplication event leading to the distinct *GLO*^{T} locus within the *S* locus supergene, thereby providing an estimated date for the origin of heterostyly.

If the *S* locus is hemizygous in *Primula vulgaris* thrum plants, one might expect closely related heterostylous species in the Primulaceae to share the same characteristic. In this chapter, an analysis of *Primula veris* is undertaken to shed light on that possibility, through genomic and expression studies. In the event that *Primula veris* and other species contain orthologues of the genes identified at the *Primula vulgaris S* locus, then providing this determining region is absent in the pin form, it supports the role of a thrum hemizygous region in the development of the distinct floral morphologies in *Primula*. This would further contradict previous predictions that homostyles are the result of recombination (Dowrick, 1956, Lewis and Jones, 1992, Wedderburn and Richards, 1992).

The Androsace were predicted as the first family to exhibit floral heteromorphy within this taxon (Mast et al., 2001). Estimates for the divergence of the Primulaceae from the Androsace are 32 (20-51) million years ago (MYA) (Magallón et al., 2015), 39 (21-59) MYA (de Vos et al., 2014), 44 (33-54) MYA with fossil priors being set with a log normal distribution (Bell et al., 2010), and 40 (30-51) MYA with fossils modelled as exponential priors (Bell et al., 2010). If the age estimate for the *GLO*-*GLO*^{T} duplication precedes the date at which families containing heterostylous species diverged from their sister clade,

then this combined with the above evidence of similar genomic regions being present throughout the Primulaceae, would suggest a single origin for heterostyly in this clade, and confirm that the homostylous species within these groups evolved from herkogamous ancestors (Wedderburn and Richards, 1992). The analyses described in this chapter will form the first steps towards elucidating the ancestral steps leading to the evolution of heterostyly.

5.3 Methods

5.3.1 *Primula veris S* locus gene model curation

Exonerate v2.2.0 (https://www.ebi.ac.uk/~guy/exonerate/) was used to align protein coding sequences for the *P. vulgaris S* locus gene models described in Chapter 4 (Li et al., 2016) against the *Primula veris* thrum VT_v1 genome assembled in Chapter 2 (

Table **2.4**) and the *Primula veris* genome assembled by Nowak et al. (2015). This, as well as PCR-based analysis performed as described in Li et al. (2016) (by JL), and alignments against the *Primula veris* thrum VT_v1 genome assembly using BLASTN (by JL) (Camacho et al., 2009), facilitated manual curation (by JMC) of the published *Primula veris* GFF file of predicted gene models in the assembly (Nowak et al., 2015) to correct (GLO^T, CYP^T, KFB^T) or add (PUM^T, CCM^T) orthologues of the *S* locus gene models identified in Li et al. (2016).

5.3.2 *P. vulgaris* and *P. veris S* locus gene model visualization

The GFF file coordinates for the introns and exons of the *Primula vulgaris S* locus genes (Chapter 4) (Li et al., 2016) and the manually curated *S* locus genes identified in the *Primula veris* assembly (Nowak et al., 2015) were used to plot the gene structures with annotated intron and exon lengths using R (v3.2.0) (https://www.r-project.org/) (introns over 1 kb plotted as 1 kb). The gene structures with intron lengths fully expanded were also plotted alongside each other in a second figure using R (v3.2.0) (https://www.r-project.org/) to facilitate comparison of intron sizes.

5.3.3 *Primula veris S* locus gene expression analysis

RNA-Seq reads for *Primula veris* pin and thrum flowers were obtained from Nowak et al. (2015); reads were aligned to the published *Primula veris* genome (Nowak et al., 2015) with TopHat (v2.0.11) (Kim et al., 2013a) and assembled with Cufflinks (v2.1.1) (Trapnell et al., 2012) guided by the curated GFF file of predicted genes. Differential expression was carried out using Cuffdiff, and expression (FPKM) values for pin and thrum flowers extracted for the five *S* locus gene models. PCR-based analysis, performed by JL using the method described in Li et al. (2016), identified the presence or absence of expression for the five genes in *Primula veris* thrum and pin plants.

5.3.4 *S* locus genomic read coverage for *P. veris* and *P. vulgaris*

Four long-homostyle (LH_v2) contigs forming the 455,880 bp *S* locus and flanking regions identified in Li et al. (2016) (Chapter 2) were removed from the assembly and replaced with the contiguous 455,880 bp *S* locus sequence. Genomic sequencing reads from the *Primula veris* thrum plant used to generate the VT_v1 assembly, and the *Primula v*ulgaris thrum and pin plant used to generate TP_VX and PP_VX assemblies (Table 2.1), were aligned to the *Primula vulgaris* LH_v2 genome assembly containing the contiguous *S* locus sequence (Chapter 4) using BWA (v0.7.12) (Li and Durbin, 2009). The SAMtools depth tool (v0.1.19) was used to return depth of coverage for reads across the *S* locus and flanking regions; the depth of read coverage was plotted across the *S* locus region in 5,000 bp windows using R (v3.2.0) (https://www.r-project.org/).

5.3.5 Bayesian relaxed-clock phylogenetic analysis

Multiple sequence alignment (MSA) of full-length nucleotide coding sequences for *DEFICIENS* (*DEF*), *GLOBOSA* (*GLO*) and *GLO^T* was carried out with MUSCLE in MEGA6 (Tamura et al., 2013) using sequences from the species listed in Table 5.1. *GLOBOSA* and *DEFICIENS* have been previously identified in *Primula vulgaris* (Li et al., 2010) whilst sequences for the remaining five *Primula* species (Table 5.1) were isolated (by JL) for the purposes of this study. These sequences were combined with additional full-length angiosperm *GLO* and *DEF* coding sequences available via NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) to estimate the date of *GLO*-*GLO^T* duplication.

Species	Order	Family	Major lineage	Gene name	Clade (*DEF/GLO*)	Accession no. (GenBank)
Antirrhinum majus	Lamiales	Plantaginaceae	Asterids	*AmGLO* *AmDEF*	*GLO* *DEF*	AB516403.1 X52023.1
Arabidopsis thaliana	Brassicales	Brassicaceae	Rosids	*AlPI* *AlAP3*	*GLO* *DEF*	NM_122031.3 NM_115294.5
Arabidopsis lyrata	Brassicales	Brassicaceae	Rosids	*AtPI* *AtAP3*	*GLO* *DEF*	XM_002871885.1 XM_002877924.1
Petunia hybrida	Solanales	Solanaceae	Asterids	*PhFBP1* *PhPMADS1*	*GLO* *DEF*	M91190.1 X69946.1
Primula denticulata	Ericales	Primulaceae	Asterids	*PdGLO* *PdGLO^T*	*GLO* *GLO*	KT257671 KT257675
Primula elatior	Ericales	Primulaceae	Asterids	*PeGLO* *PeGLO^T*	*GLO* *GLO*	KT257670 KT257677
Primula farinosa	Ericales	Primulaceae	Asterids	*PfGLO* *PfGLO^T*	*GLO* *GLO*	KT257673 KT257678
Primula veris	Ericales	Primulaceae	Asterids	*PveGLO* *PveGLO^T*	*GLO* *GLO*	KT257669 KT257674
Primula vialii	Ericales	Primulaceae	Asterids	*PviGLO* *PviGLO^T*	*GLO* *GLO*	KT257672 KT257676
Primula vulgaris	Ericales	Primulaceae	Asterids	*PvGLO* *PvGLO^T* *PvDEF*	*GLO* *GLO* *DEF*	DQ381428.1 KT257666 DQ381427.1

Table 5.1 *DEFICIENS* (*DEF*), *GLOBOSA* (*GLO*) and *GLO^T* angiosperm coding-sequences used in the multiple sequence alignment for the Bayesian relaxed-clock phylogenetic analysis.

Bayesian age estimation was implemented in BEAST (v2.1.2) (Bouckaert et al., 2014) with a Yule tree prior and an uncorrelated lognormal relaxed clock. The GTR + I + Γ substitution model was selected based on the AIC result from jModelTest (v2.1.7) (Darriba et al., 2012) with two gamma categories and an estimated proportion of invariant sites (initial value, 0.11); the estimate option was selected for the shape, rates and frequencies (initial values, default).

Normal distribution priors with mean (±SD) based on age estimates from previous studies were used as calibration points for the divergence of *DEF-GLO* = 274.75 (±37.24) MYA (Aoki et al., 2004, Kim et al., 2004, Hernández-Hernández et al., 2007), and the most recent common ancestors of *Arabidopsis thaliana* - *A. lyrata* = 12.95 (±3.01) MYA (Beilstein et al., 2010); Lamiales - Solanales = 90.25 (±7.45) MYA (Bell et al., 2010, Magallón et al., 2015); Rosids - Asterids = 118.75 (±4.71) MYA (Bell et al., 2010, Magallón et al., 2015) and the Asterids = 110.00 (±5.47) MYA (Bell et al., 2010, Magallón et al., 2015) (Table 5.2). Monophyly was enforced for the nodes used for calibration and the *Primula GLO-GLO*^{T} clade in order to reduce uncertainty in the analysis (Bouckaert et al., 2014).

Divergence	**Age ranges (MYA)**	**Reference(s)**	**Mean age applied (SD)**
Arabidopsis thaliana and *A.lyrata*	8.0 - 17.9	Beilstein et al. (2010)	12.950 (3.009)
DEFICIENS (DEF) and *GLOBOSA (GLO)*	213.5 - 336.0	Aoki et al. (2004), Kim et al. (2004), Hernández-Hernández et al. (2007)	274.750 (37.237)
Asterids (Ericales, Solanales and Lamiales)	101.0 - 119.0	Bell et al. (2010), Magallón et al. (2015)	110.000 (5.472)
Lamiales and Solanales	78.0 - 102.5	Bell et al. (2010), Magallón et al. (2015)	90.250 (7.447)
Rosids and Asterids (Brassicales, Ericales, Solanales and Lamiales)	111.0 - 126.5	Bell et al. (2010), Magallón et al. (2015)	118.750 (4.712)

Table 5.2 Divergence dates used as secondary calibrations for the Bayesian relaxed-clock phylogenetic analysis, as applied in BEAST (v2.1.2) (with normal distribution priors). The mean age (MYA) of the divergence as listed in the referenced studies is shown to three decimal places, alongside the standard deviation (SD), which was set to encompass the range of ages from the original studies. Divergence times are those generated using lognormal distributions for the fossil priors from Bell et al. (2010), and the uncorrelated lognormal (UCLN) time-tree for Magallón et al. (2015).

Nine independent Markov Chain Monte Carlo runs with 1 x 10^8 generations and a sample frequency of 5,000 were combined using LogCombiner (v1.7.5) (Drummond et al., 2012) (10% burn-in) and assessed in Tracer (v1.6) (http://tree.bio.ed.ac.uk/software/tracer/). The maximum clade credibility tree (fig. 5) was generated with TreeAnnotator (Drummond et al., 2012) (v1.7.5) and visualised in FigTree (v1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/).

Tracer (v1.6) (http://tree.bio.ed.ac.uk/software/tracer/) was used to assess the effective sample size (ESS) of all estimated parameters, as well as mixing and convergence of the Markov Chain Monte Carlo (MCMC) to stationarity. The mean (5–95% Highest Posterior Density) coefficient of variation of combined runs was 0.35 (0.14-0.59), which indicates rate heterogeneity among branches and supports the selection of a relaxed clock (Drummond and Bouckaert, 2015). Trace plots presented as figures were drawn in R (v3.3.1) (https://www.r-project.org/) using data exported from Tracer (v1.6) (http://tree.bio.ed.ac.uk/software/tracer). The "Sample from prior" option was selected in a further independent run to produce a prior distribution of node times for the purposes of validating the interaction between individual calibrations. This analysis used identical model parameters as the above runs, but sequence data is not considered.

5.3.6 Selection of sequences and parameters

The use of B-function MADS-box genes only, as opposed to genes in the MADS-box family of transcription factors as a whole, was chosen due to known rate-heterogeneity between this group and other MADS-box gene family members (Nam et al., 2003, Kim et al., 2004). The inclusion of *DEFICIENS* as well as *GLOBOSA* orthologues in this analysis facilitates the use of a calibration point incorporating previous *DEF*-*GLO* divergence estimates (Aoki et al., 2004, Kim et al., 2004, Hernández-Hernández et al., 2007). This is preferable as it has been demonstrated that the use of calibrations at deeper nodes results in more precise date estimates (Duchêne et al., 2014, Ho and Duchêne, 2014). Furthermore, use of multiple calibration points better accounts for rate heterogeneity between branches, and reduces propagation throughout the tree of potential errors in single calibrations. The age of individual nodes must fall within the interval bounded by both its ancestral and descendant nodes, and the distance between nodes of unknown age and calibration points

is reduced, thus resulting in improved divergence time estimates (Rutschmann et al., 2007, Conroy and van Tuinen, 2003, Duchêne et al., 2014, Ho and Duchêne, 2014).

The substitution model used in this analysis (GTR+I+G) was that selected by jModelTest2 (Darriba et al., 2012) using the full-length coding sequence alignment. This model is the same as used by Hernández-Hernández et al. (2007) in their dating of the *DEF/GLO* divergence (290 MYA) and also by others in large-scale estimation of divergence dates for diverse angiosperm species (Bell et al., 2010, Magallón et al., 2015). The normal distribution is used for the secondary calibrations applied in the current study as a means of taking into account uncertainty in divergence estimates presented in the original studies; this is appropriate as the standard deviation can be used to reflect the non-directional uncertainty on the original date estimate (Ho, 2007).

5.3.7 Inspection of alignment for sequence saturation

DAMBE (v6.3.3) (Xia, 2013) was used (default option; fully resolved sites only) to inspect the above alignment of B-function MADS-box genes (i) and *Primula GLO* and *GLOT* sequences aligned with MUSCLE in MEGA6 (Tamura et al., 2013) (ii) for saturation of nucleotide substitutions; the index of substitution saturation (Iss) (i=0.3870, or 0.4351 with 0.11 proportion of invariant sites (as above), ii=0.1187) was significantly lower than the critical value (Iss.c) (i=0.7243, ii=0.7318) ($p < 0.0001$) indicating low saturation. PAML (v4.9) (yn00) (Yang, 2007) was used to calculate the mean number of synonymous substitutions per synonymous site (Ks) for i=1.5593 and ii=0.443109. The estimated number of transitions and transversions versus genetic distance (GTR) for pairwise comparisons of all coding sequences in (i) and *Primula GLO* and *GLOT* (ii) was plotted with DAMBE (v6.3.3) (Xia, 2013) for further assessment of sequence saturation.

5.4 Results

5.4.1 Analysis of read depth across the 455 kb *S* locus region

Primula veris thrum genomic reads used for the assembly of VT_v1 (Table 2.2) were aligned to the 278 kb *Primula vulgaris S* locus and flanking regions (Figure 5.1). This analysis reveals that *P. veris* short sequencing reads mapping to the 278 kb central *S* locus

region are hemizygous in *Primula veris* thrum, with a distinct drop in coverage between the flanking regions and central region. There is some dissimilarity in *S* locus sequence content as shown at positions with a lower read coverage for *P. veris*.

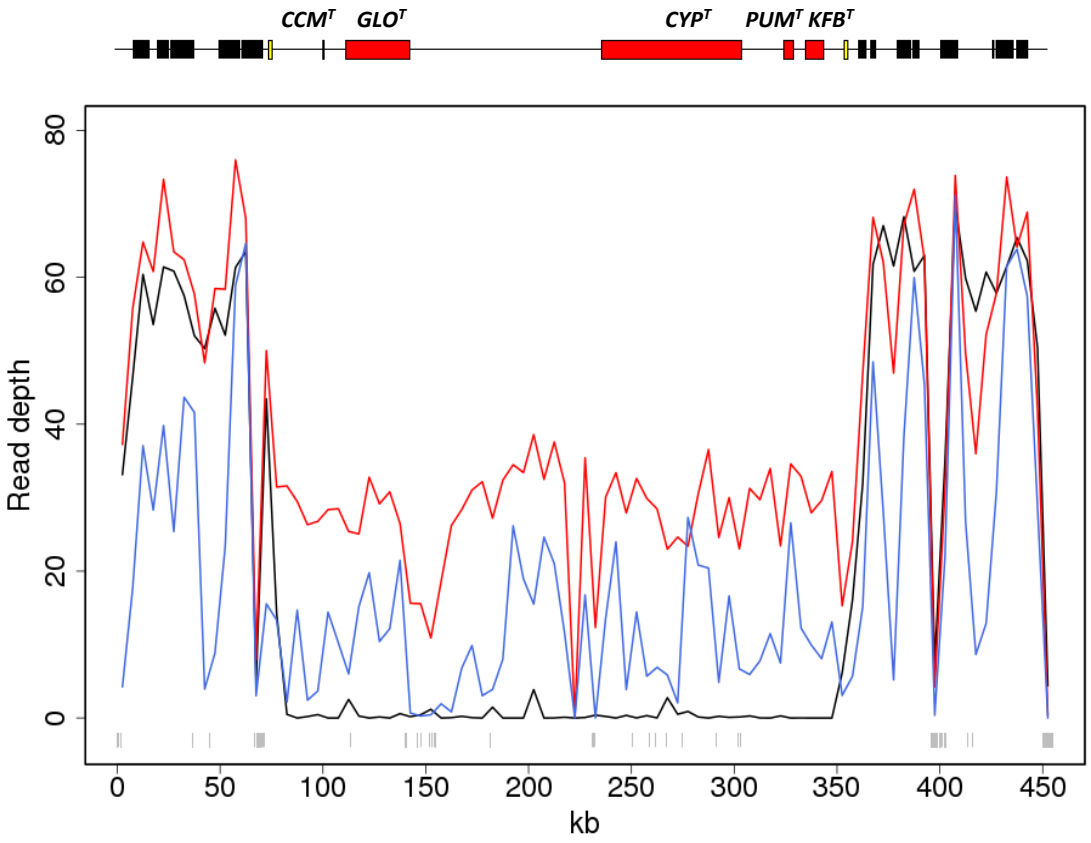


Figure 5.1 Read depth of genomic paired-end reads across the 455 kb *Primula vulgaris S* locus assembly region in 5 kb non-overlapping windows, normalised according to library size, blue = *P. veris* thrum, red = *P. vulgaris* thrum, black = *P. vulgaris* pin. *P. vulgaris* genes within the 278 kb central region are shown (top of figure) in red, with the two yellow boxes representing duplicated *CFB* loci that flank the 278 kb sequence. Predicted genes in the flanking regions are also shown (black). Grey vertical lines near the x-axis represent ambiguous bases ("N"s) in the assembled sequence.

The assembly of the complete 278 kb genomic region in *P. veris* was not possible using either VT_v1 or the Nowak et al. (2015); alignment of contigs from either assembly to this region reveals a fragmented array of associated sequences in comparison to the LH_v2 assembly derived from the homozygous long homostyle. This further demonstrates the utility of the highly-homozygous *P. vulgaris* long homostyle LH_v2 reference genome

(Chapter 2) in resolving the assembly of the *S* locus; a region in close proximity to the centromere that is characterised by a relatively high genomic repeat content (Chapter 4) (Li et al., 2015).

5.4.2 Identification of *S* locus genes in *Primula veris*

The manual curation of alignments and PCR-data facilitated the annotation or correction of gene models in the *P. veris* genome assembly (Figure 5.2). In addition, *GLO* (*GLOBOSA*) was isolated from *P. veris* (by JL) using PCR-based analyses. That the original predicted geneset of the published *P. veris* genome assembly (Nowak et al., 2015) lacks some of the *S* locus genes or parts of them is further evidence that the gene annotations in this assembly are incomplete; the pin and thrum floral RNA-Seq data without replicates was not sufficient to capture the expression of all five *S* locus genes in *P. veris*.

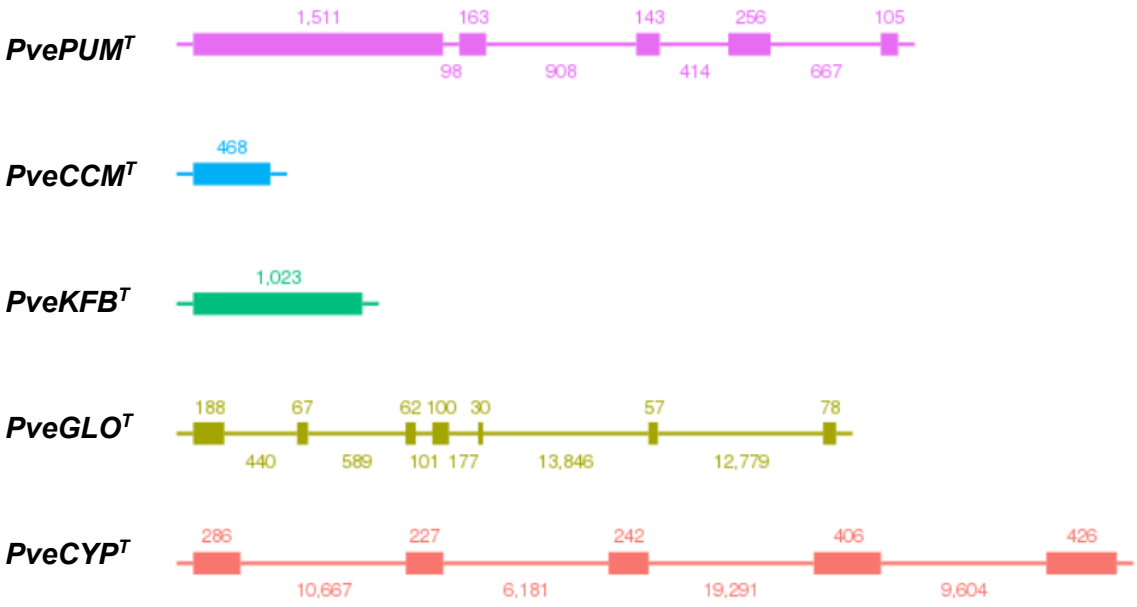


Figure 5.2 The gene structures of the five *Primula veris* (*Pve*) orthologues of genes identified in the *Primula vulgaris S* locus (Figure 4.5); exons = thick lines, introns = thin lines, introns are to scale except those greater than 1 kb which are truncated to 1 kb in the display.

If considering the *P. veris* genes alongside the *P. vulgaris* genes with introns fully expanded (Figure 5.3), then it seems that the *P. veris S* locus also contains particularly large introns. It should be noted that the introns in *P. veris* could be somewhat larger or smaller than predicted by the long mate-pair (LMP) reads used to join contigs into scaffolds. In *P. vulgaris* the ambiguous bases ("N"s) were replaced (by JL) to fill in the gaps and produce a contiguous sequence using BLASTN alignments to various *Primula* assemblies, as well as results from PCR-based analysis (Camacho et al., 2009).

However, even when accounting for a reasonable margin for error; it appears that the *Primula veris S* locus genes are also likely to contain relatively large introns. In a BLASTN (Camacho et al., 2009) alignment of the largest intron in CYP^T for both *Primula veris* and *Primula vulgaris*, there appears to be large (> 2 kb) fragmented regions of > 95% similarity between the two sequences. This perhaps suggests that large portions of this intron have been maintained, but that the region has been subject to large insertions due its hemizygosity in thrums. In future studies, the first step will be to resolve the ambiguous bases in the large *P. veris* intron, and to compare the *S* locus sequences with those in other heterostylous *Primula* species aided by sequence capture approaches for example.

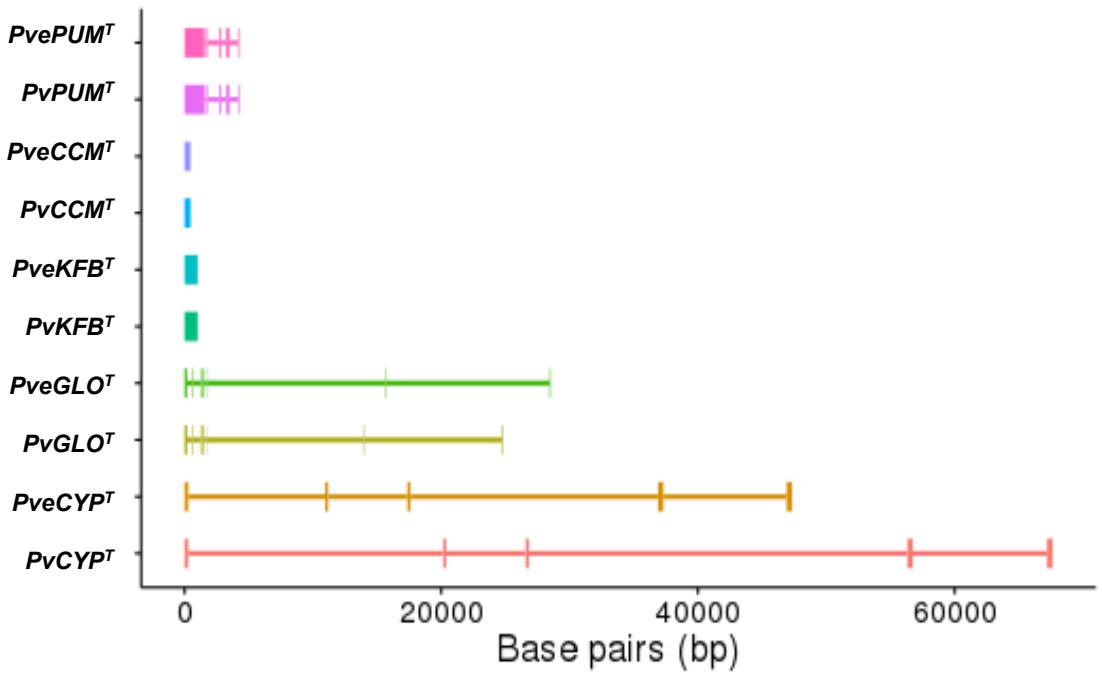


Figure 5.3 The *Primula vulgaris* (*Pv*) *S* locus genes and their *Primula veris* (*Pve*) orthologues with introns sizes fully expanded; exons = thick lines, introns = thin lines.

5.4.3 Expression of *Primula veris S* locus genes

To provide a preliminary analysis of expression for the genes at the *S* locus in *P. veris*, the *Primula veris* thrum and pin flower RNA-Seq reads from Nowak et al. (2015) were mapped back against the *P. veris* annotations; the gene models were manually curated to include the added or corrected coding sequences of the five *Primula vulgaris S* locus genes (Figure 5.4).

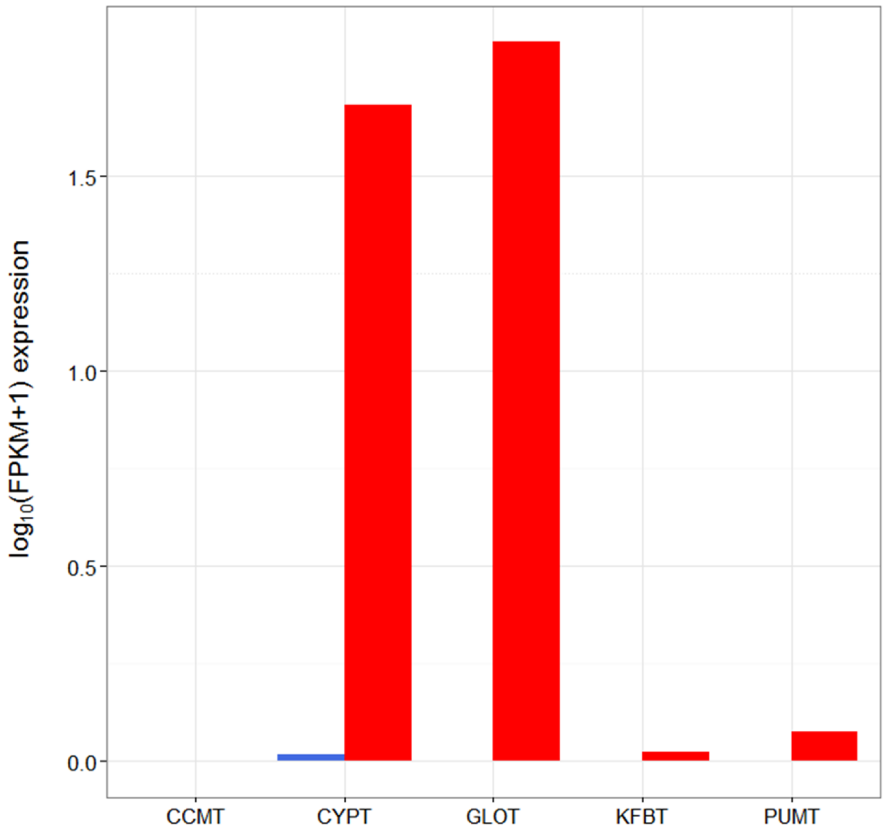


Figure 5.4 The $\log_{10}$(FPKM+1) expression of the five *S* locus genes in *P. veris* as determined by Cuffdiff (Trapnell et al., 2012) using *P. veris* pin (blue) and thrum (red) RNA-Seq reads from Nowak et al. (2015).

It is revealed in Figure 5.4 that four of the five *S* locus genes identified in *P. vulgaris* are also expressed in *P. veris*. Of these, three show thrum-specific expression as expected; and one (CYP^T) has an expression level associated with an extremely low number of pin flower RNA-Seq reads mapping to it. CCM^T shows no expression in either pin or thrum flowers

based on this dataset; this gene was not annotated in the published *P. veris* genome assembly, suggesting there was little or no transcript support from the RNA-Seq data, hence RNA-Seq reads relating to CCM^T are not found in this dataset. It should also be noted that this analysis, and indeed expression analyses by Nowak et al. (2015) are compromised by the same issues as the *Oakleaf* expression analysis, in that the published *P. veris* RNA-Seq data is unreplicated (see section 3.4.3). However, RT-PCR analyses (by JL) (Jinhong Li, personal communication) show that all genes are expressed in *P. veris* thrum, with no expression in pin.

Taken together, these data suggest that the five *S* locus genes in *P. veris* are expressed in a thrum-specific manner. CCM^T is expressed in thrum flowers as shown by PCR-based analysis, but its expression is not captured in the RNA-Seq read library. This could be due to the RNA-Seq reads being derived from flowers at a developmental stage where CCM^T is not expressed, for example. The RNA-Seq data (Figure 5.4) show that KFB^T expression in *P. veris* is very low, only 25% higher than pin-flower reads that have erroneously mapped to CYP^T. RT-PCR analyses also suggest that KFB^T expression is lower in *P. veris* than *P. vulgaris* (Jinhong Li, personal communication).

Further analysis of KFB^T will be required to ascertain its expression: it may be expressed to a low level in *P. veris* thrum flowers. This may be due to differences in floral morphology between *P. veris* and *P. vulgaris*: ancillary features may differ as they are not necessarily essential and may have evolved due to coadaptation between pollen and pistil to improve cross-fertilization or prevent self-fertilization in subtle ways (McCubbin, 2008). If this is the case, then assessment of these characters may reveal the potential role of KFB^T. This may be accomplished using an electron-microscopy approach as applied to *P. vulgaris* in Webster and Gilmartin (2006); this study identified a novel polymorphism that results in an increased cell width above the point of anther attachment in thrums (McCubbin, 2008). This suggests that the trait is not easy to detect by eye, or perhaps that it is simply not present in some species; for this reason "recombinants" (now known to be mutants; see Chapter 4) with disruption of ancillary traits are perhaps difficult to observe. KFB^T shows similarity to the *Arabidopsis* Kiss-Me-Deadly Kelch-repeat F Box protein, which regulates cytokinin activity; so perhaps KFB^T could play a role manipulating cell division above the point of anther attachment.

5.4.4 Phylogenetic analysis of *GLO*-*GLO*^{T} divergence

The *P. vulgaris S* locus gene *GLO*^{T} is an apparent duplication of the B-function MADS-box gene *GLOBOSA*, situated at distinct loci on separate LH_v2 contigs, and originally identified on distinct BAC clones (Chapter 4). For the purposes of the phylogenetic analysis that follows, *GLO* and *GLO*^{T} coding sequences were also isolated from four other heterostylous *Primula* species (by JL and OVCK) using PCR. Including *P. veris* and *P. vulgaris*, *GLO*^{T} is therefore present in six *Primula* species in total, suggesting the *GLO*-*GLO*^{T} duplication could be widespread throughout the Primulaceae.

The B-function MADS-box genes *DEFICIENS* (*APETALA3*) and *GLOBOSA* (*PISTILLATA*) are well studied and have therefore been isolated and characterised in a number of angiosperm species (Sommer et al., 1990, Trobner et al., 1992, Goto and Meyerowitz, 1994, Jack et al., 1994, Angenent et al., 1995, Krizek and Meyerowitz, 1996, Li et al., 2010). This, coupled with the availability of estimates for a broad range of angiosperm divergence dates (Bell et al., 2010, Magallón et al., 2015), provides an opportunity to date the landmark *GLO*-*GLO*^{T} duplication event associated with the assembly of the *S* locus supergene.

The alignment of full-length *DEFICIENS* (*DEF*), *GLOBSOA* (*GLO*) and *GLO*^{T} coding sequences for angiosperm species, including *Primula*, was used to produce a Bayesian relaxed-clock phylogenetic tree (Figure 5.5) with divergence dates estimated based on a combination of secondary calibrations (Table 5.2).

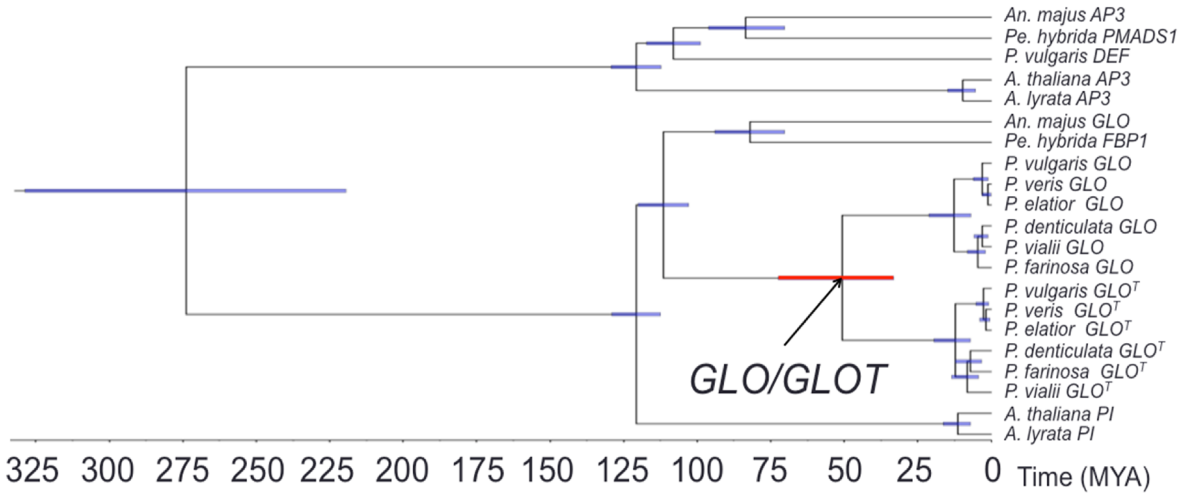


Figure 5.5 Phylogenetic analysis of coding sequences for B-function MADS-box genes from *Antirrhinum* (*An.*), *Petunia* (*Pe.*), *Arabidopsis* (*A.*) and *Primula* (*P.*) species sequences (Table 5.1). The evolutionary timescale is presented in millions of years (MYA), with the 95% Highest Posterior Density (HPD) intervals for estimated divergence dates at each branch point shown by thick blue lines; the estimate for the date of the *GLO-GLO*^{T} duplication is shown in red.

The phylogenetic analysis reveals a mean (5–95% Highest Posterior Density) age estimate of 51.7 (33.1-72.1) MYA for the divergence of *GLO* and *GLO*^{T}. The Androsace and Primulaceae families are groups containing heterostylous species within the Asterids clade. This age estimate for the duplication of *GLO-GLO*^{T} predates previous estimates for the mean divergence of the Androsace and Primulaceae of 32-44 MYA (Bell et al., 2010, de Vos et al., 2014, Magallón et al., 2015), suggesting a single origin for heterostyly in this clade with *GLO*^{T} representing a landmark evolutionary event in the assembly of the *S* locus supergene. This suggests homostyles evolved from distylous ancestors, thus supporting previous claims to that end and refuting the idea of ancestral "primary" homostyles (Mast et al., 2006).

5.4.5 Saturation analysis of B-function MADS-box genes

For the estimation of divergence times between nucleotide or amino acid sequences, a "molecular clock" model is required that describes the relationship between genetic distance and time. The genetic distances represented by branches of phylogenies from isochronous data, that is data sampled at one time point, requires the input of external information about divergence times or the evolutionary rate to infer this relationship, and thus generate a so-called "time tree" where branch lengths are proportional to time (Rambaut et al., 2016). This means evolutionary change can be compared to historical events; in our case this was necessary in order to estimate the divergence date of *GLO* and GLO^{T} sequences, to test whether this gene duplication predates the emergence of heterostyly.

The statistical relationship between genetic distance and time can be inferred using a number of statistical approaches, such as Bayesian inference, maximum-likelihood, and heuristic methods (Rambaut et al., 2016). It is possible for molecular phylogenies to be inferred over a timescale of months or years given that the sequences being analysed have undergone measurable amounts of sequence substitution, as is the case with rapidly evolving viruses (Rambaut et al., 2016). In contrast, the reliability of results from molecular phylogenies of more distantly related sequence data can depend on whether the phylogenetic information between the sequences being analysed has been lost due to substitution saturation; that is, substitutions that have been superimposed at the same site over time, masking historical signal (Hirt et al., 1999, Xia et al., 2003). This is a problem that plagues trees with deep branches, resulting in an underestimation of branch lengths; undermining the relationship between this and the time elapsed since divergence of the sequences in question (Xia et al., 2003, Philippe et al., 2011). Confirming the presence of a phylogenetic signal such that a statistical relationship between genetic divergence and time can be established is particularly important in Bayesian inference approaches such as BEAST; phylogenetic inference will proceed even when the alignments being analysed contain little or no temporal information, giving the impression of a well-supported timescale even when the data offers no basis for such conclusions to be drawn (Rambaut etal., 2016).

The effect of substitution saturation is lessened with increased sequence length (Xia et al., 2003), thus our exclusive use of full-coding sequences made possible due to the analysis of

well-studied B-function MADS-box genes would seemingly go some way to mitigating the effects of saturation. In protein coding genes, the third codon position is the most variable (Xia, 1998) but it is often undesirable to exclude it from analyses as it better conforms to neutral evolution, therefore arguably resulting in more accurate time estimates (Xia et al., 2003). This is due to substitutions at this site being mostly synonymous (Yang and Nielsen, 1998, Stewart et al., 2008), resulting in no change in amino acid or protein function due to the degeneracy of the amino acid code. Furthermore, without use of the third codon position there may be a lack of substitutions to evaluate between some sequences (Xia et al., 2003), which could be an issue when considering the closely related *GLO* and *GLO*^{T} for example (Xia et al., 2003). For reasons of sequence saturation, deep phylogenies often employ the use of amino acid sequences; amino acid sequences saturate less rapidly as there is an increased state space due to 20 possible amino acids, and only four possible nucleotides (ATGC) (Philippe et al., 2011). Xia et al. (2003) developed an entropy-based solution termed the "index of substitution saturation" that helps to determine whether the exclusion of codon positions or use of amino acid sequences is desirable; the method compares the observed entropy (information content) to the expected entropy under full substitution saturation, to test whether the entropy of the aligned sequences is significantly lower than the critical value at which sequences are expected to begin failing to recover the correct tree due to saturation (Xia et al., 2003).

The index of substitution saturation value (Xia et al., 2003, Xia, 2013) for *Primula GLO* and *GLO*^{T} sequences (0.1187) was significantly lower than the Iss critical value (0.7318, $p < 0.0001$). For all B-function MADS -box sequences in the alignment, the index of 0.4351 was also significantly lower than the critical value (0.7243) ($p < 0.0001$), indicating low saturation between the sequences used to build the phylogeny of B-function MADS-box genes. In addition, it is important to note that despite significant debate over the application of the neutral theory of evolution in the development of the "molecular clock" concept, a suite of relaxed-clock methods have been developed to account for substitution rate variation between branches of a tree; such models are implemented in BEAST, with the lognormal relaxed-clock model being applied to the current study (Kimura, 1984, Bromham and Penny, 2003, Rambaut et al., 2016). Furthermore, the use of the gamma parameter in the site substitution model means that rate heterogeneity is permitted across sites (e.g. across the M, K, I and C domains in the MADS-box genes used in this analysis), going some way towards alleviating the effects of within-sequence rate variation (Song et

al., 2016): allowing a proportion of invariant sites in addition to this means the gamma model has to explain less rate variation in the remaining sites (Drummond and Bouckaert, 2015). On the downside, use of a greater number of parameters adds uncertainty to an analysis, which means the effective sample size (ESS) of parameters and convergence of an analysis to stationarity must be evaluated (Drummond and Bouckaert, 2015).

5.4.6 Validation of the Bayesian phylogenetic analysis

In software for the Bayesian inference of phylogenetic relationships such as BEAST (Bouckaert et al., 2014), a Markov Chain Monte Carlo (MCMC) algorithm iterates over instantiations of many parameters (e.g. node age calibrations, branch lengths, substitution rate parameters, etc.) with initial values specified as priors in a pre-defined model. The convoluted parameter landscape is explored through an iterative process by changing the value of perhaps a single parameter in each step (Drummond and Bouckaert, 2015). The eventual result is a so-called posterior distribution comprising numerous distinct instantiations of the various parameters that together represent the probability of the parameters given the evidence, which in this case is data in the form of a multiple sequence alignment of B-function MADS-box coding sequences. The resulting set of phylogenetic trees forms a posterior distribution describing the uncertainty in the evolutionary relationships between the aligned sequences (Drummond and Bouckaert, 2015, Lanfear et al., 2016).

In using multiple calibrations as described above more precise age estimates might be expected (Duchêne et al., 2014, Ho and Duchêne, 2014, Schenk, 2016). However, the upper and lower bounds of the calibrations may interact with each other in unpredictable ways, such that they are in conflict with each other (Drummond and Bouckaert, 2015, Schenk, 2016). To ascertain whether this is an issue, the node times in the full prior distribution without the alignment data can be scrutinised (Ho and Duchêne, 2014, Drummond and Bouckaert, 2015). In this study, the full prior distribution in Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/) reveals that the age bounds of the individual parameters used for calibration are not dissimilar to the priors specified in the input file. This suggests that the joint prior accurately reflects our prior information on divergence dates for the angiosperm species and B-function MADS-box genes (Table 5.2).

In some cases, the changes in each step are sufficiently small that samples taken from the MCMC are not independent of each other; as such the number of samples can be larger than the Effective Sample Size (ESS). It is good practice to check the ESS of the posterior distribution for the various parameters to ensure the parameter space has been effectively explored; a good level of "mixing" should be apparent through visual inspection of the trace plot, with the chain moving quickly through the parameter space, converging on a stationary distribution that exhibits no long-term trends or large sustained changes in value (Nylander et al., 2008, Drummond and Bouckaert, 2015, Lanfear et al., 2016). It has been reported that ESS > 200 is a commonly used cut-off to determine accurate inference of the posterior distribution in Bayesian-based phylogenetic inference (Bouckaert et al., 2014, Lanfear et al., 2016). In the current analysis, the ESS for all parameters was above this value. In addition, multiple independent runs of the MCMC are often undertaken such that an assessment of consistent mixing and convergence to stationarity can be made (Drummond and Bouckaert, 2015). In this analysis, nine independent runs were combined and shown to be effectively sampling from the same distribution; converging on comparable estimates for the *GLO*-*GLO*^{T} divergence, as depicted by comparable posterior distributions for this parameter (Figure 5.6).

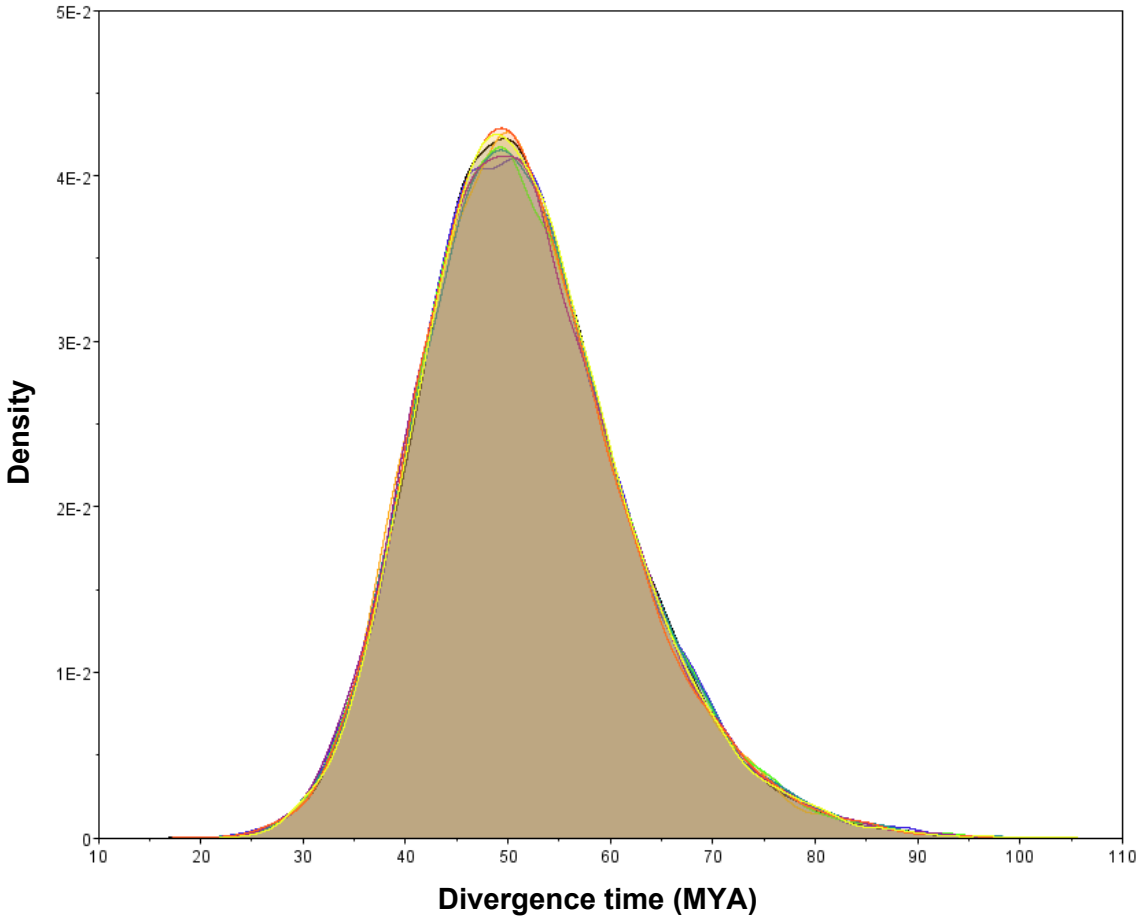


Figure 5.6 Density plot of divergence time (MYA) at the *GLO*-*GLO*^{T} node for the nine independent MCMC chains (shown by multiple colours).

Furthermore, in Figure 5.7 the trace plot of the posterior distribution is not wildly trending up or down over the course of the iterations, with consecutive samples being no more similar to each other than distant samples; this suggests the MCMC is effectively exploring the parameter space (Lanfear et al., 2016).

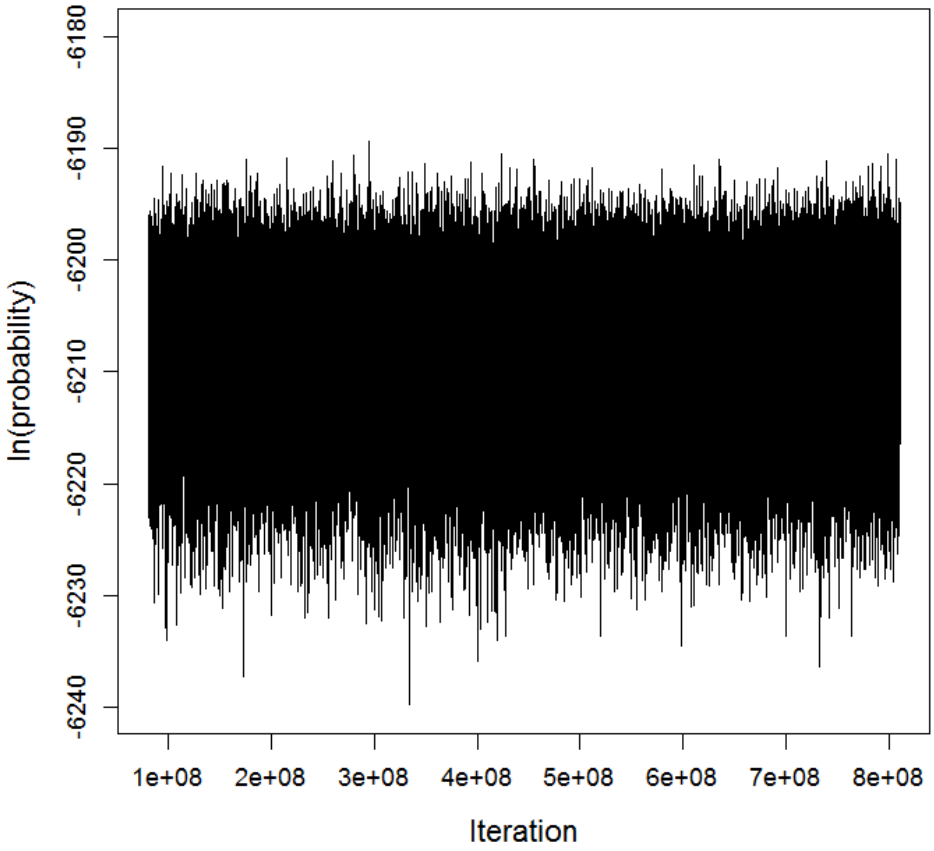


Figure 5.7 Trace plot of the posterior probability (y-axis) for the BEAST (v2.2.1) analysis of B-function MADS-box genes for each iteration of the MCMC run (x-axis).

The chain has converged around a stationary distribution (Figure 5.7) suitable for inferring parameter estimates such as divergence dates. If considering the trace plot of our parameter estimate of interest (the *GLO*-*GLO*^{T} divergence date) (Figure 5.8) then the same is true of that.

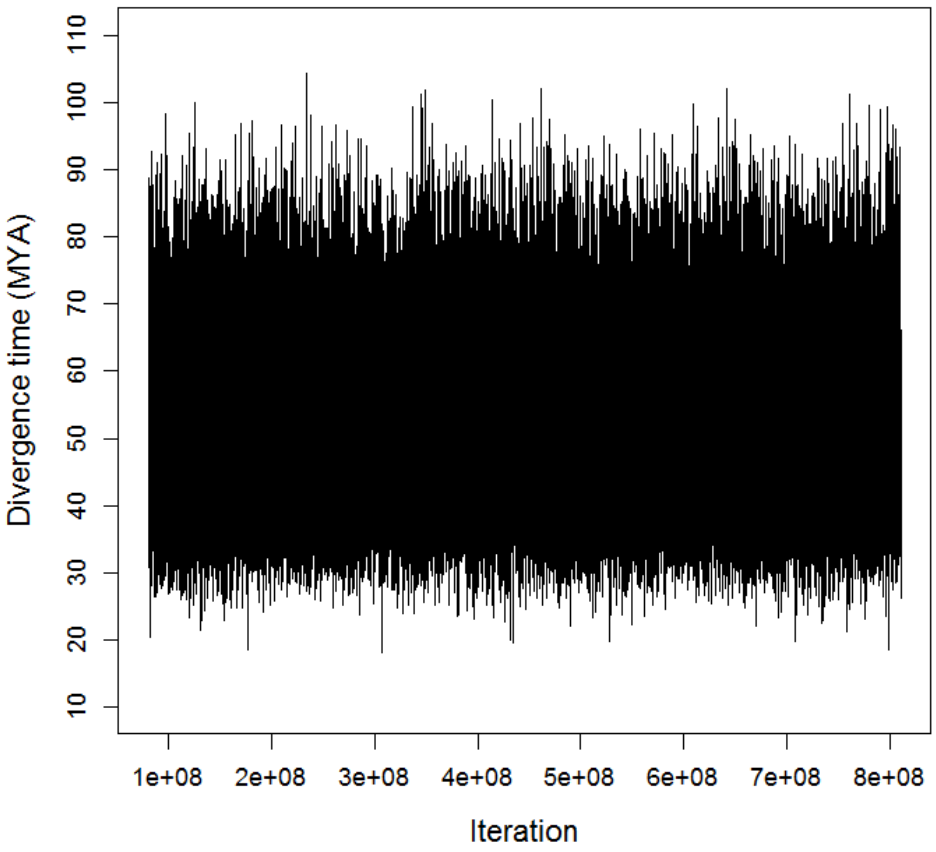


Figure 5.8 Trace plot for the *GLO-GLO*^{T} node age parameter in the BEAST (v2.2.1) (Bouckaert et al., 2014) analysis of B-function MADS-box genes; the mean estimate for the divergence date at this node (y-axis) is shown for each iteration of the MCMC run (x-axis).

The mean (5–95% Highest Posterior Density) coefficient of variation as visualised in Tracer (v1.6) (http://tree.bio.ed.ac.uk/software/tracer/) was 0.35 (0.14-0.59) (Figure 5.9). It has been reported that a coefficient of variation > 0.1 is suggestive of branch rate heterogeneity, thus supporting the selection of a relaxed molecular clock (Drummond and Bouckaert, 2015). The coefficient of variation of 0.35 demonstrates that the substitution rate varies by 35% of the molecular clock rate (Drummond and Bouckaert, 2015). In addition, the uncorrelated lognormal relaxed clock applied in this study has often been employed in angiosperm-based phylogenetic analyses (Bell et al., 2010, Magallón et al., 2010, Smith et al., 2010, Magallón et al., 2013, Magallón et al., 2015). The coefficient of variation in this analysis has no appreciable mass about zero, which would otherwise

suggest that the data cannot be used to reject the use of a strict molecular clock (Drummond and Bouckaert, 2015).

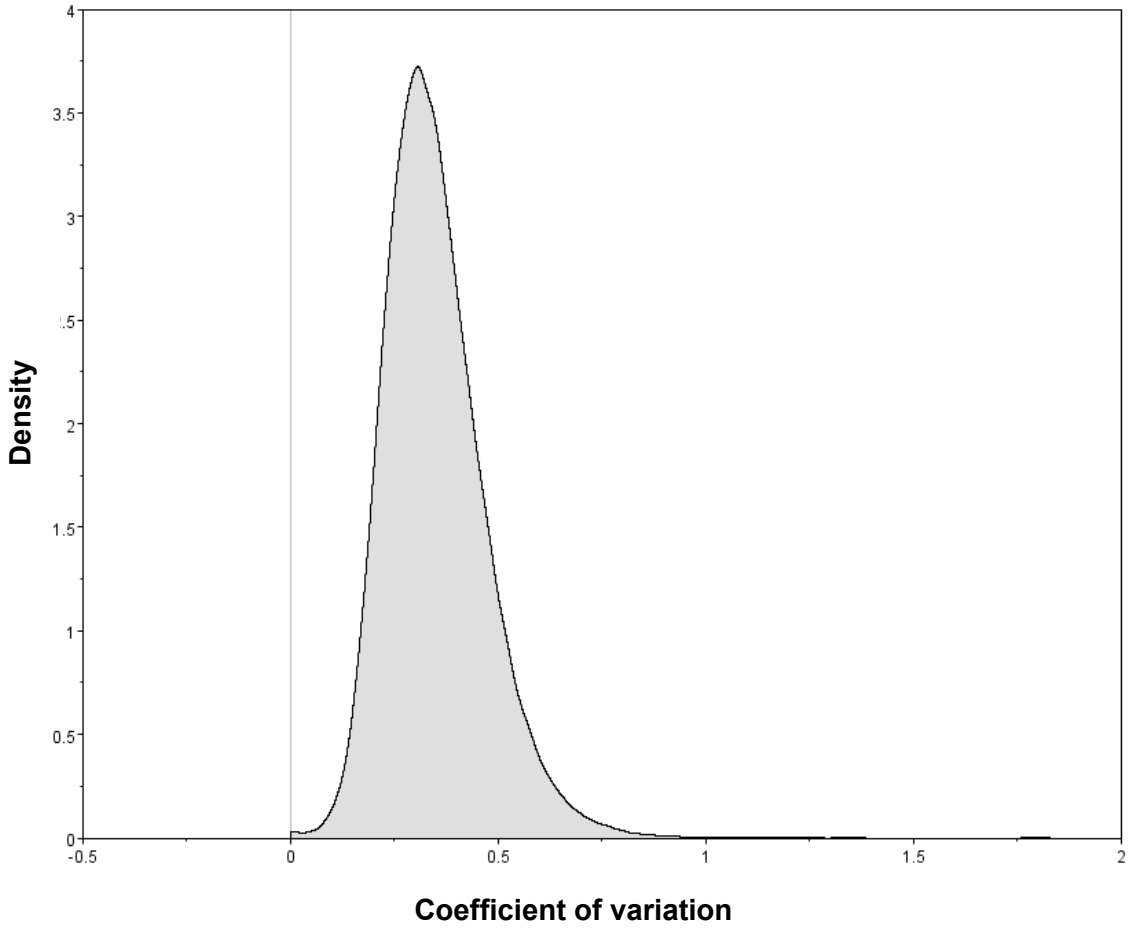


Figure 5.9 Density plot of the coefficient of variation across branches for the Bayesian phylogenetic analysis.

In the final tree, the divergence dates for the nodes used as secondary calibration points are reciprocally consistent with the estimated divergence dates for those nodes in the original studies (Table 5.2). In addition, across the *DEF* and *GLO* clades, the use of the same species and calibrations reveals mean (5–95% Highest Posterior Density) node ages that are consistent. Finally, the marginal posterior distribution of the calibrating nodes (not shown) and that of the *GLO*-*GLO*^{T} node of interest (Figure 5.10) are uni-modal, with that of the calibrating nodes being close to the prior ages; this suggests the priors on the parameters are informing but not adversely influencing the posterior distribution

(Drummond and Bouckaert, 2015). If the traces were bi-modal then this would suggest that the prior is too strong, and that the data does not agree with it, but this is not the case in the current study (Drummond and Bouckaert, 2015).

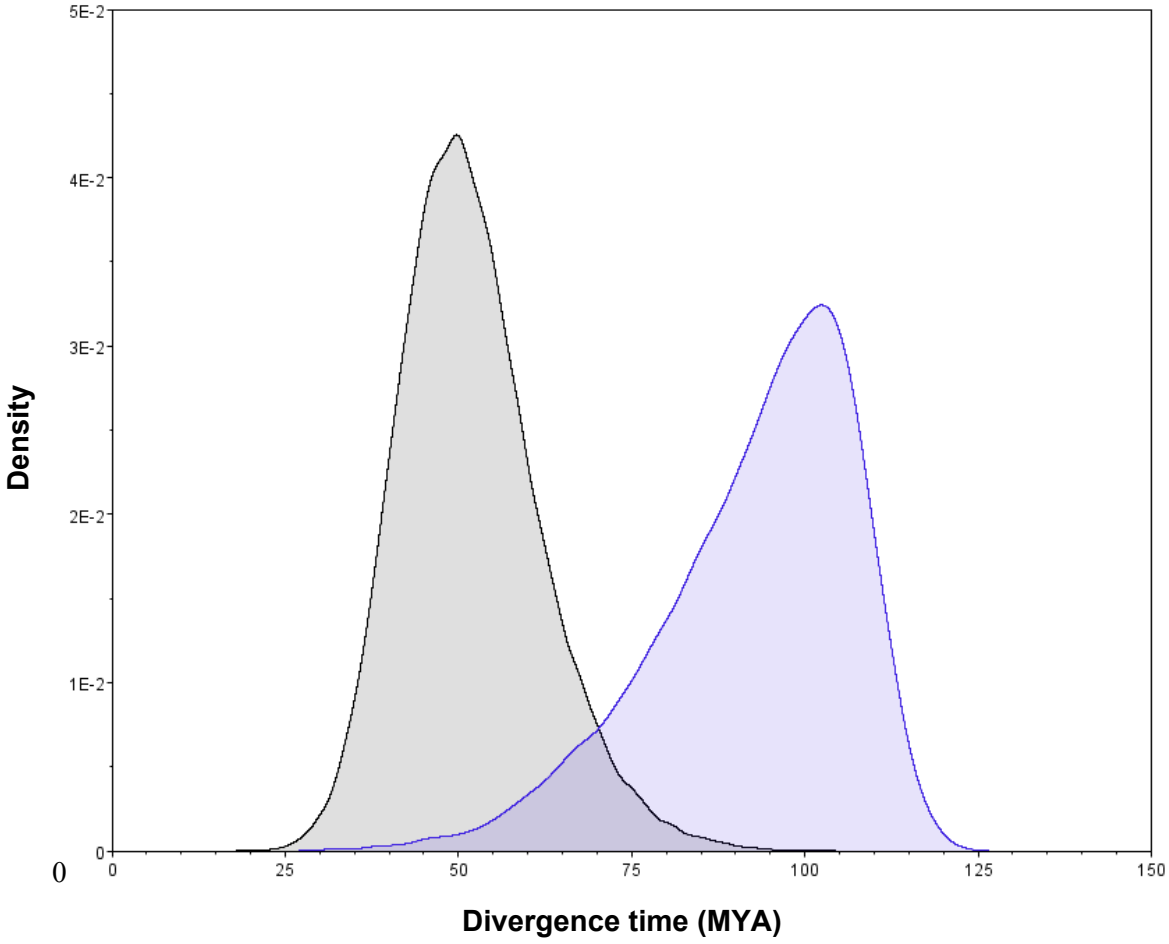


Figure 5.10 Density plot of divergence time (MYA) for the posterior distribution (grey) and prior distribution (purple) for the *GLO*-*GLO*^{T} node.

5.4.7 The comparative analysis of *CFB* flanking genes

Following the estimation and validation of the date for *GLO*-*GLO*^{T} duplication (51.7 MYA), further analysis to deduce the ancestral steps leading to the evolution of reciprocal herkogamy is an intriguing next step. The observation of direct repeats flanking the *S* locus is an interesting finding that has up to now escaped any great attention in this thesis. In

chapter four it was noted that the right and left flanking *CFB* genes are 98% similar; the pin CFB^{P} gene is more similar to CFB^{TL} (99%) than it is to CFB^{TR} (97%).

In a preliminary analysis with the recombination detection program RDP4 (Martin et al., 2015), the pin CFB^{P} gene is predicted to result from potential recombination between CFB^{TR} and CFB^{TL}. The input for this analysis was a multiple sequence alignment of CFB^{TR}, CFB^{TL} and CFB^{P}, generated with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).

5.5 Discussion

The results in this chapter effectively demonstrate that the duplication of *GLO* to GLO^{T} precedes the origin of heterostyly in the Primulaceae, thereby representing a landmark event in the establishment of the *S* locus supergene in this clade. MADS-box family transcription factors and their complex regulatory network interactions have played a central role in the evolution and diversification of angiosperm flower development (Roque et al., 2016). The B-function MADS-box genes *PISTILLA* (*GLOBOSA*) and *APETALA3* (*DEFICIENS*) are involved in the specification of petal and stamen identity (Roque et al., 2016). *PISTILLA* and *APETALA3* lineages result from an ancient duplication event following divergence of extant gymnosperms and angiosperms (Roque et al., 2016); within the *PISTILLATA* (*PI*) clade there is evidence to support further ancient duplication events during asterid evolution based on paralogous clades in phylogenies of *PI* genes from various species (Viaene et al., 2009).

In light of the instability of genetic redundancy, some of these *PI* duplicates have been lost, whilst in other cases they have persisted and diversified in function, or further duplicated (Viaene et al., 2009, Roque et al., 2016). In some organisms it is predicated that up to 30% of genes are duplicates, suggesting a means for them to persist and innovate must exist (Roque et al., 2016). The duplication of *PISTILLATA*-like genes has been noted as a mechanism for floral diversification, such as the evolution of novel floral morphologies (Viaene et al., 2009). In *Petunia x hybrida*, *PI* paralogs *FBP1* and *PMADS2* act redundantly in petal and stamen development, but *FBP1* has further diversified in function to regulate fusion of stamen filaments to the floral tube (Vandenbussche et al., 2004). In *Aquilegia*, the B-function MADS-box genes are reportedly required for modification of the

staminodium, a floral organ comprising sterile stamens, suggesting gene duplication has facilitated dramatic functional elaboration of the B-function MADS-box genes to determine a novel organ identity (Kramer et al., 2007, Sharma et al., 2014). These examples illustrate that the B-function MADS-box lineage is capable of diversification to specify novel floral morphologies in addition to the normal role of these genes in the specification of floral organ identity (Viaene et al., 2009). The duplication of the *PISTILLATA* orthologue *GLOBOSA* to form GLO^T in *Primula* species is an example of one of the more contemporary duplication events within the *PI* lineage (Viaene et al., 2009). It is a duplication that appears to be specific to the Primulaceae, which given the plasticity of B-function MADS-box gene duplicates, could feasibly have neofunctionalized and diversified to control a specific element of heterostylous floral morphology.

The 278 kb thrum-specific *S* locus region identified in *P. vulgaris* is also hemizygous in *P. veris* thrum genomic reads mapped to the region. Furthermore, the five *S* locus genes isolated in *P. vulgaris* have been identified, and shown to have thrum-specific expression in *P. veris*. These data suggest that the *S* locus supergene is hemizygous in both *P. vulgaris* and *P. veris*. In addition, based on the evidence of a shared ancestral *GLO*-GLO^T duplication for the Primulaceae, and the presence of this gene in four additional species, it seems that the hemizygosity of the *S* locus region in thrums is intimately tied to the evolution of heterostyly within this clade. If this is the case, then one possible route to heterostyly in the Primulaceae might comprise the ancestral steps outlined below.

Firstly, besides GLO^T three of the other four genes at the *S* locus, including the candidate gene for the control of style height CYP^T, are much less similar to related sequences in the genome as compared to the similarity of *GLO* and GLO^T (Figure 4.13). Notwithstanding lineage-specific and quite substantial differences in the substitution rate for these three genes (CYP^T, KFB^T and PUM^T), or loss of more similar genes from which they duplicated, perhaps they were already present at the time of the *GLO*-GLO^T duplication (52 MYA). The analysis of additional dated molecular phylogenies for these genes will help to establish whether this is the case.

Style height is under the control of CYP^T, and anther height GLO^T. It may be that the remaining gene (CCM^T) is responsible for the SI reaction, differences in pollen size, or some other ancillary trait that evolved after the *GLO*-GLO^T duplication, such that it retains similarity to the gene from which it diverged following a more contemporary duplication

event; it has been suggested that such features evolved as a result of coadaptation between the interacting pollen and pistil. *CCM*^{T} shows > 91% similarity to a homologous gene elsewhere in the genome.

The above suggests that without *GLO*^{T}, the remaining genes at the *S* locus would result in the development of flowers with a short style, and anthers that are not elevated due to the action of *CYP*^{T}. *KFB*^{T} and *PUM*^{T} may control other morphological features such as pollen size, or could otherwise have contributed to the ancestral phenotype in subtle ways; alternatively, perhaps they came under control of *GLO*^{T} following duplication of *GLO*.

Given improved phylogenies since the work of Lloyd and Webb (1992a), it is possible to observe that the sister lineage of the heterostylous Primulaceae and Androsace families contain plants that have more open flowers, with short anthers and stigma at the same height, and a corolla that is not extended into a tube (Philip Gilmartin, personal communication). This appears to be the case in *Ardisiandra*, *Lysimachia*, *Asterolinon*, *Glaux*, and *Cyclamen*. In contrast, based on thorough examination of the character states associated with heterostylous families, Lloyd and Webb (1992a) reported that the ancestral condition prior to the evolution of reciprocal herkogamy was most likely an approach herkogamous (pin-like) flower with stigma above the anthers and a long floral tube. However, this was based on the heterostylous groups themselves, not the closely related sister groups from which heterostylous species evolved, which by their own admission would have been the best approach if the sister groups of the heterostylous taxa were known at the time (Lloyd and Webb, 1992a). Indeed, Mast et al. (2006) concluded that the most recent common ancestor of *Primula* was distylous, but also suggested that the sampling of character states outside of the *Primula* lineage would be required to understand the morphology of monomorphic ancestors prior to the onset of heterostyly. In summary, some species in heterostylous families may often have the characteristic features described by Lloyd and Webb (1992a), but this does not mean that this was the primitive state.

If the duplication of *GLO*^{T} was accompanied by a mutation that resulted in the open flower becoming a long petal tube then this might result in a transition from a short homostyle-like flower with an open flower to a flower with thrum morphology. Indeed, a mutant *P. vulgaris* plant with a short corolla tube was identified by Margaret Webster in the National Collection of *Primula*, British Floral Variants (Philip Gilmartin, personal communication),

suggesting that such a transition is feasible. If the lack of short homostyle species is due a selective disadvantage of this morphological configuration (Dowrick, 1956) rather than a result of purely genetical considerations (Charlesworth and Charlesworth, 1979a) then this might explain why the ancestral phenotype might comprise a more open flower. The preliminary analysis of species outside of heterostylous families in the *Primula* lineage suggests that the ancestral condition might be a short homostyle-like flower that lacks a long floral tube; this conforms with the expected phenotype in the absence of GLO^T, provided the *GLO*-GLO^T duplication coincides with an adaptation to the long floral tube that is characteristic of most heterostylous species (Lloyd and Webb, 1992a).

Finally, evolution to the null pin allele of the *S* locus could perhaps be achieved in a single evolutionary step. It is notable that the *CFB* genes that flank the *S* locus are apparent direct repeats that are very similar in sequence identity (> 98%). The direct repeats are arranged in the same orientation such that excision by homologous recombination could occur in a way that is analogous to deletion of DNA segments using site-specific recombinase technologies. In the similar CRE-loxP and FLP-FRT site-specific recombinase technologies, FLP recombinase derived from the yeast *Saccharomyces cerevisiae*, or CRE recombinase derived from bacteriophage P1, catalyse recombination at target DNA recognition sites named LOXP or FRT, respectively (Nagy, 2000, Turan et al., 2011). If target sites are oriented as direct repeats, then recombination between them results in deletion of the intervening DNA segment. DNA insertion on the other hand is possible with one recombinase target site at the genomic location of interest and another identical target site in a circular "donor plasmid" containing a gene cassette (Nagy, 2000, Campo et al., 2002, Turan et al., 2011). The role of intra-molecular homologous recombination in genomic rearrangement suggests that two copies of a duplicated sequence can recombine to delete the intervening sequences (Bishop and Schiestl, 2000, Woodhouse et al., 2010). In the absence of heat shock-inducible recombinases, integration of sequences using site-specific recombination technologies would result in immediate re-excision with persistent recombinase activity (Hans et al., 2011, Turan et al., 2011); presumably the possible excision of the *S* locus region would occur in a much less efficient manner than these methods as there is a considerable emphasis on the efficiency of the recombinases selected for such systems (Akbudak and Srivastava, 2011). This would mean that the thrum haplotype of the *S* locus could be maintained, and not 0being persistently excised.

If this excision process was still able to occur, then perhaps it could be detected as follows: excision in thrum plants would be difficult to detect due to the presence of a pin allele, but primers specific to the flanking regions beyond the *CFB* genes could be used in PCR-amplification of pollen from the homozygous long-homostyle to detect a resulting pin allele; if Taq DNA polymerase (1 minute/kb; New England Biolabs (NEB)) was used with an extension time of 3-4 minutes for example, then the resulting ~3 kb pin *CFB* gene could be amplified over so that detection of the expected product size would confirm excision of the intervening 278 kb region; a higher-fidelity polymerase such as Phusion (New England Biolabs (NEB)) could be used to increase the chances of obtaining such a product. The use of pooled pollen for DNA extraction would significantly increase the chances of observing a *CFB* repeat-mediated excision, thus confirming the feasibility of such a mechanism. The use of primers designed for target sequences outside of the *CFB* genes, and thus high-fidelity polymerases capable of amplifying over such a region, is suggested as presumably recombination between the *CFB* genes, and thus excision, could occur at any point. If excision still occurs, then its frequency could be measured.

The frequency of direct repeat-mediated deletions is sensitive to the extent of sequence heterology, and the length of both the repeats and intervening sequence (Phadnis et al., 2005, Oliveira et al., 2008). The *CFB* flanking genes retain 98% homology over 3 kb, and therefore may provide a reasonable substrate for such excision. Furthermore, preliminary analyses reveal CFB^{P} as a predicted recombinant of CFB^{TL} and CFB^{TR}. This supports the suggestion that CFB^{TR} and CFB^{TL} could recombine to produce CFB^{P}, and that CFB^{P} could be more similar to one gene or the other depending on the crossover point; as it happens, CFB^{P} is more similar to the left flanking CFB^{TL} gene in thrum (~99%) than it is to the right (~97%). CFB^{TL} was predicted as the major parent in recombination analyses with RDP4 (Martin et al., 2015). Frequent excision would presumably lead to loss of the thrum allele, but this is not the case as pins and thrums occur with a 1:1 ratio in wild primrose populations (Ornduff, 1979). However, perhaps there is some evidence that excision is still detectable, even if it is infrequent enough such that heterostyly persists: in addition to the anomalous recessive to dominant *Primula S* allele transitions noted in the discussion of the previous chapter, there were also six documented cases of an apparent transition from *GPA* to *gpa* (Ernst, 1957, Lewis and Jones, 1992); deletion of the entire 278 kb thrum haplotype would provide a straightforward way of explaining such an occurrence, and perhaps suggests that excision of this region is not only possible, but that it might still occur.

From an evolutionary perspective, it could be argued that there is no selective advantage for a thrum plant in a population of self-fertile short homostyle-like plants, as whilst an increase in the distance between stigma and anthers may at least partially reduce selfing (Piper and Charlesworth, 1986), there is no reciprocal morph such that cross-pollination is promoted; the short-homostyle would still be likely to self-fertilize. For this reason, Charlesworth (1979) suggested that SI evolved prior to heterostyly, whilst Lloyd and Webb (1992) assumed an approach herkogamous ancestor in their model instead rather than a homostyle ancestor. If *CFB*-repeat mediated excision occurs immediately after the thrum haplotype (including flanking *CFB* genes) has been assembled, however, then the intervening genes could have excised soon after the duplication of *GLO*. This would produce the pin and establish reciprocal herkogamy in one evolutionary step rather than divergence through slow accumulation of mutations, and result in promotion of cross-pollination and perhaps a partial reduction of selfing prior to the onset of SI. Furthermore, linkage constraints that ensure favourable combinations of alleles are maintained in a genomic island of divergence under divergence hitchhiking theory would not necessarily be required, as simulations show that clusters are more likely to form through genomic rearrangements that bring coadapted loci close together under biologically realistic time scales (Yeaman, 2013).

In summary, the above suggests that *GLO* duplicated to form GLO^T with the gene for thrum style length control (CYP^T) already present at the *S* locus alongside PUM^T and KFB^T (Figure 5.11). These latter two genes may have come under the control of GLO^T to control anther positioning or SI, or could otherwise control subtle aspects of the pin and thrum morphologies. In the short homostyle with mutated GLO^T, KFB^T and PUM^T are seemingly downregulated, suggesting the former may be true (Jinhong Li, personal communication). The duplication of *GLO* to GLO^T was perhaps followed by excision of the flanking *CFB* genes by homologous recombination between these direct repeats. These speculative suggestions could be tested by demonstrating excision of the 278 kb thrum-specific region in the pollen of the homozygous long-homostyle, as well as further tests aimed towards determining the duplication dates of other genes at th0e *P. vulgaris S* locus. It could be that CCM^T arrived following a duplication event after or around the same time as the duplication of *GLO* due to the presence of a paralogue for CCM^T with over 91% sequence similarity. CCM^T could therefore be responsible for an ancillary trait, or perhaps an SI response that evolved after reciprocal herkogamy.

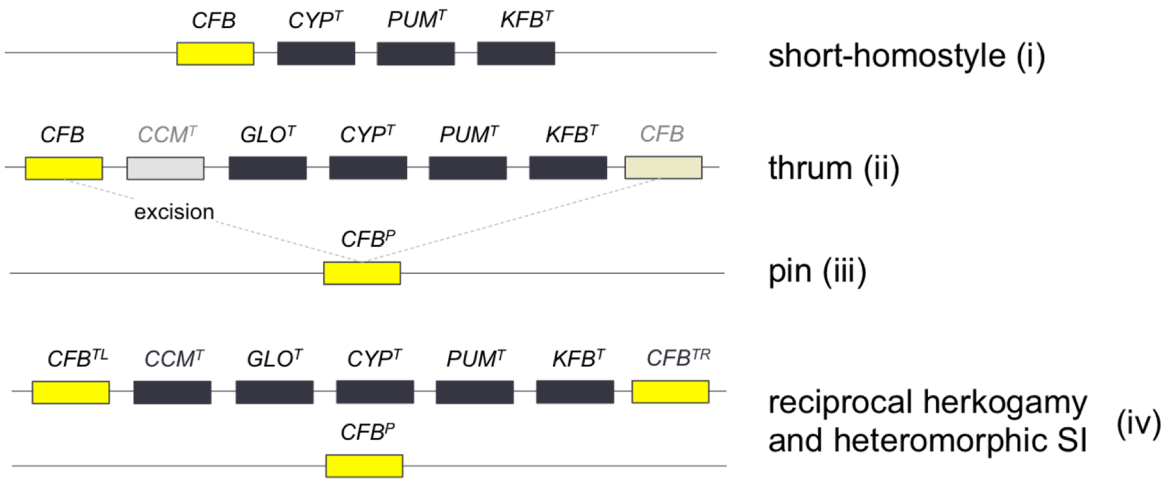


Figure 5.11 Possible ancestral steps in the evolution of the *S* locus: (i) a short homostyle-like progenitor with either CFB^{TL} or CFB^{TR}; (ii) *GLO*-GLO^T duplication and establishment of the thrum allele of the *S* locus, with CCM^T and the second *CFB* gene possibly joining the *S* locus assemblage after or around the same time as GLO^T; (iii) *CFB* direct repeat-mediated excision produces the pin allele of the *S* locus; (iv) reciprocal herkogamy is established and the full thrum haplotype is in place with CCM^T (or other genes under the control of CYP^T or GLO^T) contributing to di-allelic self-incompatibility, facilitated by an increase in cross-pollination brought about by heterostyly. In this scenario *CFB* duplication and excision may precede CCM^T duplication and/or divergence such that SI or another characteristic potentially controlled by CCM^T evolves subsequent to reciprocal herkogamy.

It would arguably be more difficult to date the assembly of the other genes at the *S* locus in comparison to the analysis of GLO^T, as the B-function MADS-box genes are relatively well-studied with regards to dated phylogenies, offering numerous calibration sources and an in-depth understanding of the phylogenetic relationships between true orthologues. However, in the absence of known dates for divergence events within the gene families in question, or the inability to incorporate species divergence dates given the difficulty of identifying true orthologues in other species, it could still be possible to obtain a rough age estimate using homologous genes within the *P. vulgaris* genome and an approximate substitution rate. Perhaps duplication of *GLO*, the CCM^T progenitor, and the *CFB* genes

occurred around the same time as part of a genome-wide increase in duplication events. There is no evidence to suggest whole genome duplication (WGD) event(s) in *Primula* as in *Actinidia chinensis* (Huang et al., 2013), but even in the absence of this it has been demonstrated that segmental duplications have had a major impact on the expansion of angiosperm gene families, with tandem duplication processes leading to large-scale expansion of the Nucleotide Binding Site-Leucine Rich Repeat (NBS-LRR) subset of plant resistance genes for example (Rodgers-Melnick et al., 2012). It seems this might more plausible than WGD, as there is a significantly lower number of annotated genes in *P. vulgaris* and *P. veris* compared to the closely related *Actinidia chinensis* (see Chapter 2). Examples of how these gene duplications can occur includes unequal recombination between nonallelic sequences due to chromosome misalignment, or retro-transposition by flanking transposons that form a composite transposon around a gene (Rodgers-Melnick et al., 2012, Van Zee et al., 2016). The latter often results in pseudogenisation and lack of gene structure due to removal of introns and reintegration of DNA derived from an RNA intermediate; as such, it seems that unequal crossing over might be a more reasonable suggestion for the mechanism leading to the GLO^T duplicate, with *GLO* situated relatively close by in the BAC assembly that comprises *S* locus-linked markers (Chapter 3).

To determine whether duplication events leading to GLO^T, CCM^T and the *CFB* flanking genes potentially form part of a large-scale duplication event, the genome-wide analysis of genetic distances between gene pairs based on all-vs-all alignments will reveal the extent of gene duplication, and could reveal peaks representing bursts of new gene duplicates that have undergone divergence (Huang et al., 2009, Vanneste et al., 2013). If such an occurrence is evident then the coordinated assembly of the complete thrum allele (including CCM^T) and flanking *CFB* genes could be followed by subsequent excision to produce the pin in a short space of time, resulting in establishment and selection of reciprocal herkogamy due to the promotion of cross-pollination, which could be reinforced by an SI system soon after if CCM^T was duplicated around the same time and does indeed have a role in the SI response. It could be that Mather and De Winton (1941) were correct in their proposition that heterostyly and SI arose together. Further to this, it is commonly noted that Darwin (1877) proposed the evolution of SI subsequent to heterostyly as an incidental by-product, but he also suggested that this might have occurred "almost simultaneously". The above suggestion could be important if *CFB*-mediated excision of the thrum haplotype has not remained active, but instead, assuming it could occur to begin

with, was a one-time occurrence. *GLO* and *CFB* duplication could be coordinated to establish the thrum haplotype of the *S* locus flanked by *CFB* direct repeats, followed by a landmark excision event that resulted in pins, and thus the foundation of heterostyly (Figure 5.12).

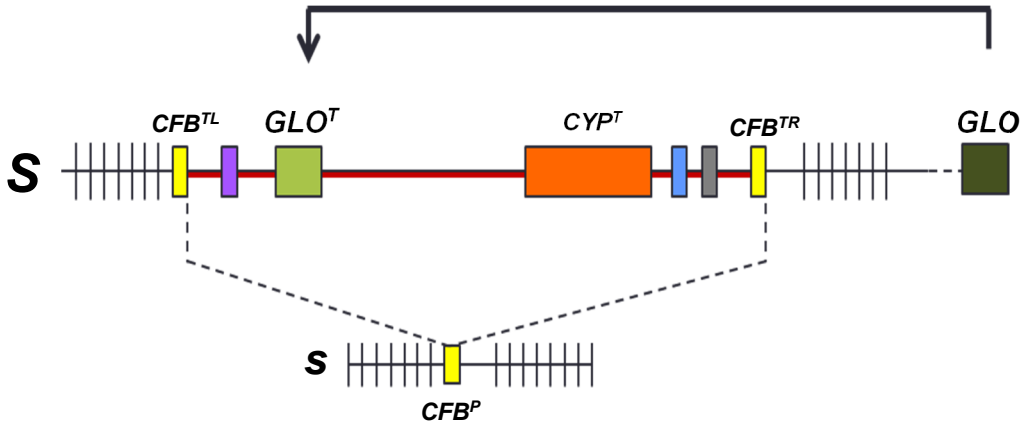


Figure 5.12 Diagram of potential *GLO* duplication and *S* haplotype excision. *GLO* duplication to GLO^T establishes the (pseudo) thrum (*S*) allele of the *S* locus followed by excision of the 278 kb region to form the pin allele (*s*). The yellow boxes represent the *CFB* genes in thrum and pin, and the five coloured boxes in the centre represent the genes at the *S* locus (purple = CCM^T, blue = PUM^T, grey = KFB^T). CCM^T (and perhaps other genes) may have been absent from the *S* locus at the time of *GLO* duplication, however. Black vertical lines represent genes in the regions flanking the *S* locus.

There are numerous alternative possibilities for the evolution of heterostyly given the new molecular data, but the evolution by an excision process discussed above appears to represent a parsimonious solution towards producing the pin allele of the *S* locus in one step, as opposed to the gradual progression towards pin and thrum from a short homostyle ancestor; it would be difficult to imagine how herkogamy would be selected for in the latter situation without SI in place (Lloyd and Webb, 1992a). The evidence of recombination between pin and thrum in the regions immediately flanking the *S* locus (Chapter 4) suggests a lack of absolute repression of recombination. Perhaps considering the distribution of thrum SNPs to determine the level of heterozygosity around the *S* locus might shed light on whether the *S* locus is or was at some point under at least partial

recombination suppression, with increased heterozygosity expected for regions close to the *S* locus under recombination suppression due to divergence between sequences in coupling with the pin and thrum alleles of the *S* locus. The hemizygosity of the *S* locus in its current state guarantees the absence of recombination between pin and thrum, but there was not necessarily any suppression of recombination prior to this.

The idea of thrum haplotype excision is speculative, but if it proves to be correct then the hemizygosity of the *S* locus region in thrums is central to the evolution of heterostyly in the Primulaceae. Despite the diverse origins of this system across the angiosperms, it is tempting to suggest that heterostyly in other families could be determined by hemizygosity. The methods presented in this thesis offer a viable avenue through which to exploit the above, providing a means for the discovery of morph-specific genomic regions in the agronomically important *Fagopyrum esculentum* (buckwheat) for example. In ongoing efforts to reveal the heterostyly genes in *F. esculentum*, *S-ELF3* was identified as closely-linked to the *S* locus and exclusively present in genomes of short-styled (thrum) buckwheat plants (Yasui et al., 2012); this suggests the *S* locus in *F. esculentum* might also be thrum-specific. It seems that heterostyly in buckwheat is not controlled by genes similar to those in *Primula* however: in a preliminary comparative analysis, alignments of the *Primula S* locus genes to the recently sequenced *Fagopyrum esculentum* genome (by JMC) did not reveal close homologues (Yasui et al., 2016).

The presence of heterostyly across diverse angiosperm families is indeed a remarkable example of convergent evolution, and it could be that hemizygosity is just one possible route to this phenomenon. Perhaps chromosomal inversions or splice site variations of a single gene may have been employed in other heterostylous species for the purposes of progression towards dimorphic polymorphism under linkage constraints, as in mimetic butterflies (see Chapter 4). The finding that the *S* locus in *Primula* is hemizygous in thrums is an exciting discovery that may apply to other heterostylous species, but there may be more surprises.

6

General discussion and conclusions

Debate over the evolutionary origin, function and molecular basis of heterostyly has spanned three centuries: beginning with the insights of Darwin (1862), and continuing through to the advancement of biomolecular science and whole genome sequencing (Cohen, 2010, Nowak et al., 2015, Li et al., 2015a, Cocker et al., 2017). Darwin (1877) first put forward a hypothesis suggesting that the purpose of heterostyly was in promoting outcrossing: he proposed that (di-allelic) self-incompatibility (SI) must have followed heterostyly and in fact could see no benefit to SI given that it precluded fertilization with half the population. Bateson and Gregory (1905) defined the genetic model for the inheritance of heterostyly, and others expanded on this (Lewis, 1954, Dowrick, 1956, Lewis and Jones, 1992), predicting the number of genes at the *S* locus and uncovering linked genes (De Winton and Haldane, 1931). In addition, the chromosomal location of the *S* locus, and the idea of tightly-linked genes and putative recombination suppression at the so-called "supergene" was established (Pellow, 1928, Darlington, 1931, Darlington and Mather, 1949). Following this early work, theoretical models for the evolution of heterostyly were defined (Charlesworth and Charlesworth, 1979b, Lloyd and Webb, 1992a, Lloyd and Webb, 1992b), whilst efforts in generating genetic maps and associated BAC assemblies advanced progress towards the identification of the determining genes (Li et al., 2011b, Li et al., 2015a).

The main questions arising from the above studies concern the identification of the *S* locus and its constituent genes, determining the genomic architecture and extent of recombination suppression in the region, and establishing the order and importance of heterostyly, di-allelic SI, and the intermediate ancestral steps. This thesis details an array of genome assemblies and associated genomic studies (Chapter 2) that have underpinned the assembly and analysis of the *Primula S* locus (Chapter 4 and 5). This constitutes a large

group of resources that have instigated the first steps in functional analysis of the heterostyly-determining genes, and redefinition of the evolutionary model.

The *Primula vulgaris S* locus is a 278 kb thrum-specific region that is surrounded by sequences that are homogenised between pin and thrum due to recombination (Chapter 4). The hemizygosity of the *S* locus in thrums precludes recombination with the null pin allele. This means there is no reason for suppression of recombination in the *S* locus or its surrounding sequences; a further mechanism to prevent recombination such as a chromosomal inversion (Mather, 1950) is not required in this species. Furthermore, it is possible through the excision of the entire thrum haplotype (Chapter 5) that the emergence of the null pin allele occurred immediately after *GLO* duplication. If this is the case, a mechanism to suppress recombination in the region, such that pin and thrum alleles could accumulate mutations and diverge over time without homogenisation by recombination, may never have been present.

Darwin realised the advantages of self-sterility, but suggested that SI in *Primula* could offer no benefit due to preventing each plant from fertilizing half the population (Darwin, 1877): "although it may be beneficial to an individual plant to be sterile with its own pollen, cross-fertilisation being thus ensured, how can it be any advantage to a plant to be sterile with half its brethren". Furthermore, he proposed that SI was incidentally, rather than selectively acquired due to the apparent plasticity of the system under changing conditions (Darwin, 1876). Darwin noted that it was "incredible that so peculiar a form of mutual infertility should have been specially acquired unless it were highly beneficial to the species", and concluded that self-sterility could only be beneficial after a plant had become adapted for cross-pollination: "it would manifestly be injurious to a plant that its stigma should fail to receive its own pollen, unless it had already become well adapted for receiving pollen from another individual" (Darwin, 1876, Darwin, 1877). This was supported due to the absence of SI in many species that were adapted for cross-pollination.

Darwin had the foresight to realise the roles of cross- and self-fertilization as mechanisms that could promote or hinder the vigour and fertility of resultant progeny; significantly, this was prior to a wider understanding of the principles of genetics that provide an underlying cause for the negative effects of inbreeding (Darwin, 1876, Charlesworth and Charlesworth, 1987). He concluded, as did Lloyd and Webb (1992b), that SI must follow heterostyly in *Primula* (Darwin, 1877). Lloyd and Webb (1992a) note that almost all

authors in the 20th century reject the idea that SI evolved after heterostyly, including Charlesworth and Charlesworth (1979b).

The findings in this thesis present a plausible scenario in which the pin allele could have emerged soon after the thrum allele, thus providing an adaptive advantage through the promotion of cross-pollination between the two morphs. This most likely preceded the evolution of a di-allelic SI system because the breakdown of heterostyly seen in homostyles must be a result of mutation and not recombination at the *S* locus, suggesting the morphology-associated genes also regulate the SI determinants. These insights support Darwin's original postulate for the evolution of heterostyly prior to SI, and in contrast to the findings of Charlesworth and Charlesworth (1979b) and Lloyd and Webb (1992a), tentatively suggest that the ancestral state might have been a short homostyle-like plant, followed by duplication of *GLO* and, perhaps, excision of the entire *S* haplotype.

From the brief discussion above, it is clear that these studies provide extensive insight into the key questions posed in the literature, revealing a number of unexpected findings that will provoke significant debate regarding the evolution of heterostyly. It is interesting to speculate on the possible ancestral evolutionary events and mechanisms for maintenance of the hemizygous *S* locus given that there is no opportunity for homology-based intra-chromosomal repair. The previous chapters touch on a number of methods for obtaining empirical evidence for this proposal, including the analysis of pollen from homozygous long homostyle plants to detect excision of the *S* haplotype, the dating and analysis of individual and large-scale duplication events, and the analysis of the *S* locus in distinct *Primula* populations using sequence capture approaches. The latter will facilitate detection of polymorphisms between this region and the freely recombining flanking sequences to quantify the extent of degeneration, and assist the formulation of hypotheses for the maintenance and breakdown of the *S* locus region in distinct populations and species.

Besides the functional analysis of the heterostyly-determining genes, including a strong bioinformatic element in the accompanying expression, assembly, and downstream regulatory network studies, the above analyses represent truly interesting directions in which to proceed with research into the remarkable genomic architecture of the *S* locus, and the evolution of heterostyly. These findings form a platform for research into heterostyly and hemizygousity in *Primula*, and perhaps other angiosperm families as well; potential speculative applications for bioinformatic approaches in identifying the molecular

basis of phenomena such as apomixis are revealed, as well as suggestions for the utility of the genes themselves in the floral engineering of crop plants. Focused analyses towards defining the basis of mechanisms to prevent inbreeding and promote outcrossing are of great importance in gaining an understanding of the origins of biodiversity (Potts et al., 2010, Baldock et al., 2015). In the face of pollinator decline, the maintenance and breakdown of pollination syndromes is a significant consideration in food security and ecology due to the fruits and seeds that result from pollination, and the reliance of food and horticultural crop breeding systems on pollinators (Klein et al., 2007, Potts et al., 2010).

In combination with the vast array of molecular and classical genetic studies into heterostyly, the genomic resources presented in this thesis demonstrate the importance of taking a comprehensive, collaborative, and holistic view of such fundamental studies, in order to encompass broad methodologies in seeking to understand the molecular and evolutionary basis of the mechanisms underpinning diverse phenomena. This approach has facilitated the identification of the *S* locus and its constituent genes in *Primula*, and provided the author with a captivating doctoral project which presents a preliminary picture of the genomic and evolutionary basis of heterostyly. The studies point towards countless possibilities for future research into the interesting genomic architecture of the *S* locus, the analysis of downstream pathways involved in specifying the precise architectures of the two floral morphs, and the evaluation of the ancestral steps leading to the emergence of this classic 19th century model for convergent evolution.

Bibliography

ABBASI, N., PARK, Y. I. & CHOI, S. B. 2011. Pumilio Puf domain RNA-binding proteins in *Arabidopsis*. *Plant signaling & behavior,* 6, 364-368.

ABRUSAN, G., GRUNDMANN, N., DEMESTER, L. & MAKALOWSKI, W. 2009. TEclass — a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics,* 25, 1329-1330.

AITKEN, R. J. & MARSHALL GRAVES, J. A. 2002. Human spermatozoa: The future of sex. *Nature,* 415, 963-963.

AKBUDAK, M. A. & SRIVASTAVA, V. 2011. Improved FLP recombinase, FLPe, efficiently removes marker gene from transgene locus developed by Cre–lox mediated site-specific gene integration in rice. *Molecular Biotechnology,* 49, 82-89.

ALLEN, A. M. & HISCOCK, S. J. 2008. Evolution and phylogeny of self-incompatibility systems in angiosperms. *In*: FLANKIN-TONG, V. E. (Ed.) *Self-Incompatibility in Flowering Plants: Evolution, Diversity, and Mechanisms.* Berlin, Heidelberg: Springer Verlag.

ALTENBURG, E. 1916. Linkage in *Primula sinensis*. *Genetics,* 1, 354-366.

ALVAREZ-JUBETE, L., WIJNGAARD, H., ARENDT, E. K. & GALLAGHER, E. 2010. Polyphenol composition and *in vitro* antioxidant activity of amaranth, quinoa buckwheat and wheat as affected by sprouting and baking. *Food Chemistry,* 119, 770-778.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biology,* 11, R106.

ANDERS, S., PYL, P. T. & HUBER, W. 2014. HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv,* 31, 166-169.

ANGENENT, G. C., BUSSCHER, M., FRANKEN, J., DONS, H. J. & VAN TUNEN, A. J. 1995. Functional interaction between the homeotic genes *fbp1* and *pMADS1* during petunia floral organogenesis. *The Plant Cell,* 7, 507-516.

AOKI, S., UEHARA, K., IMAFUKU, M., HASEBE, M. & ITO, M. 2004. Phylogeny and divergence of basal angiosperms inferred from *APETALA3*- and *PISTILLATA*-like MADS-box genes. *Journal of Plant Research,* 117, 229-244.

ARNAUDEAU, C., LUNDIN, C. & HELLEDAY, T. 2001. DNA double-strand breaks associated with replication forks are predominantly repaired by homologous recombination involving an exchange mechanism in mammalian cells. *Journal of Molecular Biology,* 307, 1235-1245.

ÁVILA-ARCOS, M. C., SANDOVAL-VELASCO, M., SCHROEDER, H., CARPENTER, M. L., MALASPINAS, A.S., WALES, N., PEÑALOZA, F., BUSTAMANTE, C. D. & GILBERT, M. T. P. 2015. Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries. *Methods in Ecology and Evolution,* 6, 725-734.

AVNER, P. & HEARD, E. 2001. X-chromosome inactivation: counting, choice and initiation. *Nature Reviews Genetics,* 2, 59-67.

BACHTROG, D. 2013. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nature reviews Genetics,* 14, 113-124.

BAKER, H. 1966. The evolution, functioning and breakdown of heteromorphic incompatibility systems. I. The Plumbaginaceae. *Evolution,* 20, 349-368.

BALAJI, S., BABU, M. M., IYER, L. M. & ARAVIND, L. 2005. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Research,* 33, 3994-4006.

BALDOCK, K. C. R., GODDARD, M. A., HICKS, D. M., KUNIN, W. E., MITSCHUNAS, N., OSGATHORPE, L. M., POTTS, S. G., ROBERTSON, K. M., SCOTT, A. V., STONE, G. N., VAUGHAN, I. P. & MEMMOTT, J. 2015. Where is the UK's pollinator biodiversity? The importance of urban areas for flower-visiting insects. *Proceedings of the Royal Society B: Biological Sciences,* 282, 20142849.

BANKAR, K. G., TODUR, V. N., SHUKLA, R. N. & VASUDEVAN, M. 2015. Ameliorated *de novo* transcriptome assembly using Illumina paired end sequence data with Trinity Assembler. *Genomics Data,* 5, 352-359.

BARRETT, S. C. & GLOVER, D. E. 1985. On the Darwinian hypothesis of the adaptive significance of tristyly. *Evolution,* 39, 766-774.

BARRETT, S. C. & SHORE, J. S. 1987. Variation and evolution of breeding systems in the *Turnera ulmifolia* L. complex (Turneraceae). *Evolution,* 41, 340-354.

BARRETT, S. C. & WOLFE, L. M. 1986. Pollen heteromorphism as a tool in studies of the pollination process in *Pontederia cordata* L. *In:* MULCAHY, D. L., MULCAHY, G. B. & OTTAVIANO, E. (eds.) *Biotechnology and Ecology of Pollen.* New York: Springer.

BARRETT, S. C. H. 1992. Heterostylous genetic polymorphisms: model systems for evolutionary analysis. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* Berlin: Springer Verlag.

BARRETT, S. C. H. 2010. Darwin's legacy: the forms, function and sexual divesity of flowers. *Philisophical Transactions of The Royal Society B,* 365, 351-368.

BARRETT, S. C. H. & CRUZAN, M. B. 1994. Incompatibility in heterostylous plants. *In:* WILLIAMS, E. G., CLARKE, A. E. & KNOX, R. B. (eds.) *Genetic control of self-incompatibility and reproductive development in flowering plants.* Dordrecht: Springer Netherlands.

BARRETT, S. C. H. & HOUGH, J. 2012. Sexual dimorphism in flowering plants. *Journal of Experimental Botany*, 63(2), 695–709.

BARRETT, S. C. H. & SHORE, J. S. 2008. New insights on heterostyly: Comparative biology, ecology and genetics. *In*: FLANKIN-TONG, V. D. (ed.) *Self-Incompatibility in Flowering Plants: Evolution, Diversity and Mechanisms.* Berlin: Springer Verlag

BATESON, W. 1902. *Mendel's Principles of Heredity,* Cambridge, Cambridge University press.

BATESON, W. & GREGORY, R. P. 1905. On the inheritance of heterostylism in *Primula*. *Proceedings of the Royal Society of London B Series,* 76, 581-586.

BEHNKE, H. D., LÜTTGE, U., ESSER, K., KADEREIT, J. W. & RUNGE, M. 2012. *Progress in Botany / Fortschritte der Botanik: Structural Botany Physiology Genetics Taxonomy Geobotany / Struktur Physiologie Genetik Systematik Geobotanik*. Berlin, Heidelberg: Springer Verlag

BEILSTEIN, M. A., NAGALINGUM, N. S., CLEMENTS, M. D., MANCHESTER, S. R. & MATHEWS, S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences,* 107, 18724-18728.

BELL, C. D., SOLTIS, D. E. & SOLTIS, P. S. 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany,* 97, 1296-1303.

BELLES-BOIX, E., HAMANT, O., WITIAK, S. M., MORIN, H., TRAAS, J. & PAUTOT, V. 2006. *KNAT6*: an Arabidopsis homeobox gene involved in meristem activity and organ separation. *Plant Cell,* 18, 1900-1907.

BHARATHAN, G., JANSSEN, B. J., KELLOGG, E. A. & SINHA, N. 1999. Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. *Molecular Biology and Evolution,* 16, 553-563.

BHARATHAN, G., GOLIBER, T. E., MOORE, C., KESSLER, S., PHAM, T. & SINHA, N. R. 2002. Homologies in leaf form inferred from *KNOXI* gene expression during development. *Science,* 296, 1858-1860.

BI, R., LIU, P. 2016. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics*, 17, 146.

BISHOP, A. J. R. & SCHIESTL, R. H. 2000. Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. *Human Molecular Genetics,* 9, 2427-2334.

BJÖRKMAN, T., SAMIMY, C. & PEARSON, K. J. 1995. Variation in pollen performance among plants of *Fagopyrum esculentum*. *Euphytica,* 82, 235-240.

BLAINEY, P., KRYZWIKSI, M., ALTMAN, N. 2014. Points of significance: Replication. *Nature Methods*, 11, 879-880.

BODMER, W. F. 1960. The genetics of homostyly in populations of *Primula vulgaris*. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences,* 242, 517-549.

BOUCKAERT, R., HELED, J., KÜHNERT, D., VAUGHAN, T., WU, C. H., XIE, D., SUCHARD, M. A., RAMBAUT, A. & DRUMMOND, A. J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol,* 10, e1003537.

BRADNAM, K. R., FASS, J. N., ALEXANDROV, A., BARANAY, P., BECHNER, M., BIROL, I., BOISVERT, S., CHAPMAN, J. A., CHAPUIS, G. & CHIKHI, R. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience,* 2, 1.

BRANZEI, D. & FOIANI, M. 2008. Regulation of DNA repair throughout the cell cycle. *Nature Reviews Molecular Cell Biology,* 9, 297-308.

BRENNEMAN, M. A., WAGENER, B. M., MILLER, C. A., ALLEN, C. & NICKOLOFF, J. A. 2002. XRCC3 Controls the Fidelity of Homologous Recombination: Roles for XRCC3 in Late Stages of Recombination. *Molecular Cell,* 10, 387-395.

BRIDGES, C. B. 1914. The chromosome hypothesis of linkage applied to cases in sweetpeas and *Primula*. *American Naturalist,* 48, 524-534.

BROMHAM, L. & PENNY, D. 2003. The modern molecular clock. *Nature Reviews Genetics,* 4, 216-224.

BUTLER, J., MACCALLUM, I., KLEBER, M., SHLYAKHTER, I. A., BELMONTE, M. K., LANDER, E. S., NUSBAUM, C. & JAFFE, D. B. 2008. ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Research,* 18, 810-820.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10, 1-9.

CAMPO, N., DAVERAN-MINGOT, M. L., LEENHOUTS, K., RITZENTHALER, P. & LE BOURGEOIS, P. 2002. Cre-loxP recombination system for large genome rearrangements in *Lactococcus lactis*. *Applied and Environmental Microbiology,* 68, 2359-2367.

CHAISSON, M. J. & PEVZNER, P. A. 2008. Short read fragment assembly of bacterial genomes. *Genome Research,* 18, 324-330.

CHAN, S. S. K. & KYBA, M. 2013. What is a Master Regulator? *Journal of stem cell research & therapy,* 3, 114.

CHANG, Z., LI, G., LIU, J., ZHANG, Y., ASHBY, C., LIU, D., CRAMER, C. L. & HUANG, X. 2015. Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. *Genome Biology,* 16, 1-10.

CHARLESWORTH, B. & CHARLESWORTH, D. 1979a. Maintenance and breakdown of distyly. *American Naturalist,* 114, 499-513.

CHARLESWORTH, D. 2010. Self-incompatibility. *F1000 Biology Reports,* 2, 68.

CHARLESWORTH, D. & CHARLESWORTH, B. 1979b. Model for the evolution of distyly. *American Naturalist,* 114, 467-498.

CHARLESWORTH, D. & CHARLESWORTH, B. 1987. Inbreeding Depression and its Evolutionary Consequences. *Annual Review of Ecology and Systematics,* 18, 237-268.

CHEN, K. Y., CONG, B., WING, R., VREBALOV, J. & TANKSLEY, S. D. 2007. Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes. *Science,* 318, 643-645.

CHOO, K. H. A. 1998. Why Is the Centromere So Cold? *Genome Research,* 8, 81-82.

CHUCK, G., LINCOLN, C. & HAKE, S. 1996. *KNAT1* induces lobed leaves with ectopic meristems when overexpressed in *Arabidopsis*. *Plant Cell,* 8, 1277-1289.

CLARK, M. B., AMARAL, P. P., SCHLESINGER, F. J., DINGER, M. E., TAFT, R. J., RINN, J. L., PONTING, C. P., STADLER, P. F., MORRIS, K. V. & MORILLON, A. 2011. The reality of pervasive transcription. *PLoS Biol,* 9, e1000625.

CLARKE, J., WU, H. C., JAYASINGHE, L., PATEL, A., REID, S. & BAYLEY, H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology,* 4, 265-270.

CLUSIUS, C. 1583. *Rariorum aliquot stirpium, per Pannoniam, Austriam, & vicinas quasdam provincias observatarum historia, quator libris expressa,* Antwerp, Officina Christophori Plantini.

COCK, P. J. A., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research,* 38, 1767-1771.

COCKER, J., WEBSTER, M. A., LI, J., WRIGHT, J., KAITHAKOTTIL, G. G., SWARBRECK, D. & GILMARTIN, P. M. 2015. *Oakleaf:* an *S* locus-linked mutation of *Primula vulgaris* that affects leaf and flower development. *New Phytologist*, 208, 149-161.

COCKER, J. M., WRIGHT, J., LI, J., SWARBRECK, D., DYER, S., CACCAMO, M. & GILMARTIN, P. M. 2017. The *Primula vulgaris* genome (in preparation).

COHEN, J. I. 2010. "A case to which no parallel exists": The influence of Darwin's Different Forms of Flowers. *American Journal of Botany,* 97, 701-716.

COLLINS, F. S., MORGAN, M. & PATRINOS, A. 2003. The Human Genome Project: lessons from large-scale biology. *Science,* 300, 286-290.

COLLINS, F. S. & WEISSMAN, S. M. 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proceedings of the National Academy of Sciences,* 81, 6812-6816.

CONESA, A., GOTZ, S., GARCIA-GOMEZ, J. M., TEROL, J., TALON, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21, 3674-3676.

CONROY, C. J. & VAN TUINEN, M. 2003. Extracting time from phylogenies: positive interplay between fossil and genetic data. *Journal of Mammalogy,* 84, 444-455.

CROSBY, J. L. 1940. High proportions of homostyle plants in populations of *Primula vulgaris*. *Nature,* 145, 672-673.

CROSBY, J. L. 1949. Selection of an unfavourable gene complex. *Evolutionary Ecology Research,* 3, 212-230.

CURTIS, W. 1777-1798. *Flora Londinensis,* London, William Curtis.

DARLINGTON, C. D. 1931. Meiosis in diploid and tetraploid *Primula sinensis*. *Journal of genetics,* 24, 65-95.

DARLINGTON, C. D. & MATHER, K. 1949. *The Elements of Genetics*, Schocken Books.

DARRIBA, D., TABOADA, G. L., DOALLO, R. & POSADA, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods,* 9, 772-772.

DARWIN, C. 1876. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. London: John Murray.

DARWIN, C. R. 1860. Darwin Correspondence Database Letter 2785. Cambidge Digital Library: http://cudl.lib.cam.ac.uk/collections/darwin_mss.

DARWIN, C. R. 1862. On the two forms or dimorphic condition in the species of *Primula*, and on their remarkable sexual relations. *Journal of the Proceedings of the Linnean Society, Botany,* 6, 77-96.

DARWIN, C. R. 1863. On the existence of two forms, and on their reciprocal sexual relation, in several species of the genus *Linum*. *Journal of the Proceedings of the Linnean Society, Botany,* 7, 69-83.

DARWIN, C. R. 1877. *The different forms of flowers on plants of the same species.* London: John Murray.

DE NETTANCOURT, D. 1997. Incompatibility in angiosperms. *Sexual Plant Reproduction,* 10, 185-199.

DE VEGA, J. J., AYLING, S., HEGARTY, M., KUDRNA, D., GOICOECHEA, J. L., ERGON, Å., ROGNLI, O. A., JONES, C., SWAIN, M., GEURTS, R., LANG, C., MAYER, K. F. X., RÖSSNER, S., YATES, S., WEBB, K. J., DONNISON, I. S., OLDROYD, G. E. D., WING, R. A., CACCAMO, M., POWELL, W., ABBERTON, M. T. & SKØT, L. 2015. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Scientific Reports,* 5, 17394.

DE VOS, J. M., HUGHES, C. E., SCHNEEWEISS, G. M., MOORE, B. R. & CONTI, E. 2014. Heterostyly accelerates diversification via reduced extinction in primroses. *Proceedings of the Royal Society B: Biological Sciences,* 281: 20140075.

DE WINTON, D. 1928. Inheritance in *Primula sinensis*. *The Fourth Primula Conference 1928,* 1, 84-90.

DE WINTON, D. & HALDANE, J. B. S. 1931. Linkage in the tetraploid *Primula sinensis*. *Journal of Genetics,* 24, 121-144.

DE WINTON, D. & HALDANE, J. B. S. 1933. The Genetics of *Primula Sinensis*. II. Segregation and interaction of factors in the Diploid. *Journal of Genetics,* 27, 1-44.

DE WINTON, D. & HALDANE, J. B. S. 1935. The genetics of *Primula sinensis*. III. Linkage in the diploid. *Journal of Genetics,* 31, 67-100.

DEL BIANCO, M., GIUSTINI, L. & SABATINI, S. 2013. Spatiotemporal changes in the role of cytokinin during root development. *New Phytologist,* 199, 324-338.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. 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. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* 43, 491-498.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29, 15-21.

DODD, M. E., SILVERTOWN, J. & CHASE, M. W. 1999. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. *Evolution,* 53, 732-744.

DONG, Y., XIE, M., JIANG, Y., XIAO, N., DU, X., ZHANG, W., TOSSER-KLOPP, G., WANG, J., YANG, S., LIANG, J., CHEN, W., CHEN, J., ZENG, P., HOU, Y., BIAN, C., PAN, S., LI, Y., LIU, X., WANG, W., SERVIN, B., SAYRE, B., ZHU, B., SWEENEY, D., MOORE, R., NIE, W.,

SHEN, Y., ZHAO, R., ZHANG, G., LI, J., FARAUT, T., WOMACK, J., ZHANG, Y., KIJAS, J., COCKETT, N., XU, X., ZHAO, S., WANG, J. & WANG, W. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nature Biotechnology,* 31, 135-141.

DOWRICK, V. P. J. 1956. Heterostyly and homostyly in *Primula obconica*. *Heredity,* 10, 219-236.

DRUMMOND, A. J. & BOUCKAERT, R. R. 2015. *Bayesian Evolutionary Analysis with BEAST*, Cambridge University Press.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. & RAMBAUT, A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution,* 29, 1969-1973.

DUCHÊNE, S., LANFEAR, R. & HO, S. Y. W. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular Phylogenetics and Evolution,* 78, 277-289.

DULBERGER, R. 1964. Flower dimorphism and self-incompatibility in *Narcissus tazetta* L. *Evolution,* 18, 361-363.

DULBERGER, R. 1975. *S*-gene action and the significance of characters in the heterostylous syndrome. *Heredity,* 35, 407-415.

EARL, D., BRADNAM, K., ST. JOHN, J., DARLING, A., LIN, D., FASS, J., YU, H. O. K., BUFFALO, V., ZERBINO, D. R., DIEKHANS, M., NGUYEN, N., ARIYARATNE, P. N., SUNG, W. K., NING, Z., HAIMEL, M., SIMPSON, J. T., FONSECA, N. A., BIROL, İ., DOCKING, T. R., HO, I. Y., ROKHSAR, D. S., CHIKHI, R., LAVENIER, D., CHAPUIS, G., NAQUIN, D., MAILLET, N., SCHATZ, M. C., KELLEY, D. R., PHILLIPPY, A. M., KOREN, S., YANG, S. P., WU, W., CHOU, W. C., SRIVASTAVA, A., SHAW, T. I., RUBY, J. G., SKEWES-COX, P., BETEGON, M., DIMON, M. T., SOLOVYEV, V., SELEDTSOV, I., KOSAREV, P., VOROBYEV, D., RAMIREZ-GONZALEZ, R., LEGGETT, R., MACLEAN, D., XIA, F., LUO, R., LI, Z., XIE, Y., LIU, B., GNERRE, S., MACCALLUM, I., PRZYBYLSKI, D., RIBEIRO, F. J., YIN, S., SHARPE, T., HALL, G., KERSEY, P. J., DURBIN, R., JACKMAN, S. D., CHAPMAN, J. A., HUANG, X., DERISI, J. L., CACCAMO, M., LI, Y., JAFFE, D. B., GREEN, R. E., HAUSSLER, D., KORF, I. & PATEN, B. 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research,* 21, 2224-2241.

EDDY, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Informatics,* 2009. 205-211.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32, 1792-1797.

EDGAR, R. C. & BATZOGLOU, S. 2006. Multiple sequence alignment. *Current Opinion in Structural Biology,* 16, 368-373.

ERNST, A. 1928a. Genetische Studien über Calycanthemie bei *Primula*. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zurich,* 15, 665-704.

ERNST, A. 1928b. Zur Vererbung der morphologischen Heterostylie merkmale. *Berichte der Deutschen Botanischen Gesellschaft,* 46, 573-588.

ERNST, A. 1936a. Erblichkeitsforschung an calycanthemen Primeln. *Der Zuchter,* 8, 281-294.

ERNST, A. 1936b. Heterostylie-Forschung Versuche zur genetischen analyse eines organisations und 'Anpassungs' merkmales. *Zeitschrift für Induktive Abstammungs und Vererbungslehre*, 71, 156-230.

ERNST, A. 1936c. Weitere untersuchungen zur Phänanalyse zum Fertilitätsproblem und zur Genetik heterostyler Primeln. II. *Primula hortensis*. *Archive der Julius Klaus Stiftung für Vererbungsforschung Sozialanthropologie und Rassenhygiene,* 11, 1-280.

ERNST, A. 1942. Vererbung durch labile gene. *Archive der Julius Klaus Stiftung für Vererbungsforschung Sozialanthropologie und Rassenhygiene,* 17, 1-567.

ERNST, A. 1955. Self-fertility in monomorphic Primulas. *Genetica,* 27, 391-448.

ERNST, A. 1957. Aberranten in Erbgangen der Heterostylie Merkmale bei Primeln und ihre Bedutung fur Vererbungs und Evolutionsprobleme. *Archive der Julius Klaus Stiftung für Vererbungsforschung Sozialanthropologie und Rassenhygiene,* 32, 16-217.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research,* 8, 175-185.

FARLOW, A., MEDURI, E. & SCHLÖTTERER, C. 2011. DNA double-strand break repair and the evolution of intron density. *Trends in Genetics,* 27, 1-6.

FESENKO, N., FESENKO, N. & ROMANOVA, O. 2006. Genofond i selektsiya krupyanykh kul'tur. Grechikha (Gene Pool and Breeding of Groat Crops. Buckwheat). *St. Petersburg: VIR*.

FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., HEGER, A., HETHERINGTON, K., HOLM, L., MISTRY, J., SONNHAMMER, E. L. L., TATE, J. & PUNTA, M. 2014. Pfam: the protein families database. *Nucleic Acids Research,* 42, D222-D230.

FISHER, R. A. 1949. A Theoretical system of selection for homostyle *Primula*. *Sankhya,* 9, 325-342.

FLINT-GARCIA, S. A., THORNSBERRY, J. M. & BUCKLER, E. S. 2003. Structure of linkage disequilibrium in plants. *Annual Review Plant Biology,* 54, 357-374.

FU, L., NIU, B., ZHU, Z., WU, S. & LI, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28, 3150-3152.

GANDERS, F. R. 1974. Disassortative pollination in the distylous plant *Jepsonia heterandra*. *Canadian Journal of Botany,* 52, 2401-2406.

GANDERS, F. R. 1976. Pollen flow in distylous populations of *Amsinckia* (Boraginaceae). *Canadian Journal of Botany,* 54, 2530-2535.

GANDERS, F. R. 1979. The biology of heterostyly. *New Zealand Journal of Botany,* 17, 607-635.

GIBBS, P. E. 1986. Do homomorphic and heteromorphic self-incompatibility systems have the same sporophytic mechanism? *Plant Systematics and Evolution,* 154, 285-323.

GILMARTIN, P. M. 2015. On the origins of observations of heterostyly in *Primula*. *New Phytologist,* 208, 39-51.

GILMARTIN, P. M. & LI, J. 2010. Homing in on heterostyly. *Heredity,* 105, 161-162.

GOEL, S., CHEN, Z., CONNER, J. A., AKIYAMA, Y., HANNA, W. W. & OZIAS-AKINS, P. 2003. Delineation by fluorescence *in situ* hybridization of a single hemizygous chromosomal region associated with aposporous embryo sac formation in *Pennisetum squamulatum* and *Cenchrus ciliaris*. *Genetics,* 163, 1069-1082.

GOLYNSKAYA, E. L., BASHRIKOVA, N. V. & TOMCHUK, N. N. 1976. Phytomaemagglutinins of the pistil in *Primula* as possible proteins of generative incompatibility. *Soviet Journal of Plant Physiology,* 23, 69-76.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17, 333-351.

GORDON, G. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 10, 645-656.

GOTO, K. & MEYEROWITZ, E. M. 1994. Function And Regulation Of The Arabidopsis Floral Homeotic Gene Pistillata. *Genes & Development,* 8, 1548-1560.

GRABHERR, M. G., HAAS, B. J., YASSOUR, M., LEVIN, J. Z., THOMPSON, D. A., AMIT, I., ADICONIS, X., FAN, L., RAYCHOWDHURY, R., ZENG, Q., CHEN, Z., MAUCELI, E., HACOHEN, N., GNIRKE, A., RHIND, N., DI PALMA, F., BIRREN, B. W., NUSBAUM, C., LINDBLAD-TOH, K., FRIEDMAN, N. & REGEV, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology,* 29, 644-U130.

GRAVES, J. A. M. 2006. Sex chromosome specialization and degeneration in mammals. *Cell,* 124, 901-914.

GRAVES, J. A. M. & DISTECHE, C. M. 2007. Does gene dosage really matter? *Journal of Biology,* 6, 1.

GREGORY, R. P. 1911. Experiments with *Primula sinensis*. *Journal of Genetics,* 1, 73-132.

GREGORY, R. P., DE WINTON, D. & BATESON, M. A. 1923. Genetics of *Primula sinensis*. *Journal of Genetics,* 13, 219-253.

GRIFFITHS-JONES, S., GROCOCK, R. J., VAN DONGEN, S., BATEMAN, A. & ENRIGHT, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research,* 34, D140-D144.

GROVE, M. D., SPENCER, G. F., ROHWEDDER, W. K., MANDAVA, N., WORLEY, J. F., WARTHEN, J. D., STEFFENS, G. L., FLIPPEN-ANDERSON, J. L. & COOK, J. C. 1979. Brassinolide, a plant growth-promoting steroid isolated from *Brassica napus* pollen. *Nature,* 281, 216-217.

GUGGISBERG, A., MANSION, G. & CONTI, E. 2009. Disentangling reticulate evolution in an arctic–alpine polyploid complex. *Systematic Biology,* 58, 55-73.

GURNEY, M., PRESTON, C., BARRETT, J. & BRIGGS, D. 2007. Hybridisation between Oxlip *Primula elatior* (L.) Hill and Primrose *P. vulgaris* Hudson, and the identification of their variable hybrid *P.* ×*digenea* A. Kerner. *Watsonia,* 26, 239-252.

GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S. & REGEV,

A. 2010. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology,* 28, 503-510.

HALDANE, J. B. S. 1933. Two new allelomorphs for heterostylism in *Primula*. *American Naturalist,* 67, 559-560.

HALDER, A., JAIN, M., CHAUDHARY, I. & KABRA, M. 2010. Prevalence of 22q11.2 microdeletion in 146 patients with cardiac malformation in a referral hospital of North India. *BMC Medical Genetics,* 11, 101-101.

HALLAB, A. 2015. Protein function prediction using phylogenomics, domain architecture analysis, data integration, and lexical scoring. *PhD thesis, Universitäts-und Landesbibliothek Bonn, Germany.*

HANS, S., FREUDENREICH, D., GEFFARTH, M., KASLIN, J., MACHATE, A. & BRAND, M. 2011. Generation of a non-leaky heat shock-inducible Cre line for conditional Cre/lox strategies in zebrafish. *Developmental Dynamics,* 240, 108-115.

HARDER, L. D. & WILSON, W. G. 1998. Theoretical consequences of heterogeneous transport conditions for pollen dispersal by animals. *Ecology,* 79, 2789-2807.

HAREVEN, D., GUTFINGER, T., PARNIS, A., ESHED, Y. & LIFSCHITZ, E. 1996. The making of a compound leaf: Genetic manipulation of leaf architecture in tomato. *Cell*, 84, 735-744.

HARLEY, V. R., CLARKSON, M. J. & ARGENTARO, A. 2003. The molecular action and regulation of the testis-determining factors, SRY (Sex-determining Region on the Y chromosome) and SOX9 [SRY-related high-mobility group (hmg) box 9]. *Endocrine Reviews,* 24, 466-487.

HAY, A., KAUR, H., PHILLIPS, A., HEDDEN, P., HAKE, S. & TSIANTIS, M. 2002. The gibberellin pathway mediates KNOTTED1-type homeobox function in plants with different body plans. *Current Biology,* 12, 1557-1565.

HAY, A. & TSIANTIS, M. 2006. The genetic basis for differences in leaf form between *Arabidopsis thaliana* and its wild relative *Cardamine hirsuta*. *Nature Genetics,* 38, 942-947.

HAY, A. & TSIANTIS, M. 2010. *KNOX* genes: versatile regulators of plant development and diversity. *Development,* 137, 3153-3165.

HAYTA, S., SMEDLEY, M. A., LI, J., HARWOOD, W. A. & GILMARTIN, P. M. 2016. Plant regeneration from leaf-derived callus cultures of primrose (*Primula vulgaris*). *Horticularal Science,* 51, 558-562.

HERNÁNDEZ-HERNÁNDEZ, T., MARTÍNEZ-CASTILLA, L. P. & ALVAREZ-BUYLLA, E. R. 2007. Functional diversification of B MADS-box homeotic regulators of flower development: adaptive evolution in protein–protein interaction domains after major gene duplication events. *Molecular Biology and Evolution,* 24, 465-481.

HIRT, R. P., LOGSDON, J. M., HEALY, B., DOREY, M. W., DOOLITTLE, W. F. & EMBLEY, T. M. 1999. Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences,* 96, 580-585.

HISCOCK, S. J. 2002. Pollen recognition during the self-incompatibility response in plants. *Genome Biology*, 3, reviews1004.1–1004.6.

HO, S. Y. M. 2007. Calibrating molecular estimates of substitution rates and divergence times in birds. *Journal of Avian Biology,* 38, 409-414.

HO, S. Y. W. & DUCHÊNE, S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology,* 23, 5947-5965.

HOFF, K. J., LANGE, S., LOMSADZE, A., BORODOVSKY, M. & STANKE, M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics,* 32, 767-769.

HOLT, C. & YANDELL, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics,* 12, 1-14.

HUANG, S., DING, J., DENG, D., TANG, W., SUN, H., LIU, D., ZHANG, L., NIU, X., ZHANG, X., MENG, M., YU, J., LIU, J., HAN, Y., SHI, W., ZHANG, D., CAO, S., WEI, Z., CUI, Y., XIA, Y., ZENG, H., BAO, K., LIN, L., MIN, Y., ZHANG, H., MIAO, M., TANG, X., ZHU, Y., SUI, Y., LI, G., SUN, H., YUE, J., SUN, J., LIU, F., ZHOU, L., LEI, L., ZHENG, X., LIU, M., HUANG, L., SONG, J., XU, C., LI, J., YE, K., ZHONG, S., LU, B. R., HE, G., XIAO, F., WANG, H. L., ZHENG, H., FEI, Z. & LIU, Y. 2013. Draft genome of the kiwifruit *Actinidia chinensis*. *Nature Communications,* 4, 2640.

HUANG, S., LI, R., ZHANG, Z., LI, L., GU, X., FAN, W., LUCAS, W. J., WANG, X., XIE, B., NI, P., REN, Y., ZHU, H., LI, J., LIN, K., JIN, W., FEI, Z., LI, G., STAUB, J., KILIAN, A., VAN DER VOSSEN, E. A. G., WU, Y., GUO, J., HE, J., JIA, Z., REN, Y., TIAN, G., LU, Y., RUAN, J., QIAN, W., WANG, M., HUANG, Q., LI, B., XUAN, Z., CAO, J., ASAN, WU, Z., ZHANG, J., CAI, Q., BAI, Y., ZHAO, B., HAN, Y., LI, Y., LI, X., WANG, S., SHI, Q., LIU, S., CHO, W. K., KIM, J. Y., XU, Y., HELLER-USZYNSKA, K., MIAO, H., CHENG, Z., ZHANG, S., WU, J., YANG, Y., KANG, H., LI, M., LIANG, H., REN, X., SHI, Z., WEN, M., JIAN, M., YANG, H., ZHANG, G., YANG, Z., CHEN, R., LIU, S., LI, J., MA, L., LIU, H., ZHOU, Y., ZHAO, J., FANG, X., LI, G., FANG, L., LI, Y., LIU, D., ZHENG, H., ZHANG, Y., QIN, N., LI, Z., YANG, G., YANG, S., BOLUND, L., KRISTIANSEN, K., ZHENG, H., LI, S., ZHANG, X., YANG, H., WANG, J., SUN, R., ZHANG, B., JIANG, S., WANG, J., DU, Y. & LI, S. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet,* 41, 1275-1281.

HUNTER, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in science and engineering,* 9, 90-95.

JACK, T., FOX, G. L. & MEYEROWITZ, E. M. 1994. Arabidopsis homeotic gene *APETALA3 ectopic expression* - transcriptional and post-transcriptional regulation determine floral organ identity. *Cell,* 76, 703-716.

JACKMAN, S. D., VANDERVALK, B. P., MOHAMADI, H., CHU, J., YEO, S., HAMMOND, S. A., JAHESH, G., KHAN, H., COOMBE, L., WARREN, R. L. & BIROL, I. 2016. ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. *bioRxiv,* 10.1101/068338

JHALA, A. J., BHATT, H., TOPINKA, K. & HALL, L. M. 2011. Pollen-mediated gene flow in flax (*Linum usitatissimum* L.): can genetically engineered and organic flax coexist? *Heredity,* 106, 557-566.

JOHNSON, N. A. & LACHANCE, J. 2012. The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. *Annals of the New York Academy of Sciences,* 1256, E1-22.

JONES, P., BINNS, D., CHANG, H. Y., FRASER, M., LI, W., MCANULLA, C., MCWILLIAM, H., MASLEN, J., MITCHELL, A., NUKA, G., PESSEAT, S., QUINN, A. F., SANGRADOR-

VEGAS, A., SCHEREMETJEW, M., YONG, S. Y., LOPEZ, R. & HUNTER, S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics,* 30, 1236-1240.

JORON, M., FREZAL, L., JONES, R. T., CHAMBERLAIN, N. L., LEE, S. F., HAAG, C. R., WHIBLEY, A., BECUWE, M., BAXTER, S. W., FERGUSON, L., WILKINSON, P. A., SALAZAR, C., DAVIDSON, C., CLARK, R., QUAIL, M. A., BEASLEY, H., GLITHERO, R., LLOYD, C., SIMS, S., JONES, M. C., ROGERS, J., JIGGINS, C. D. & FFRENCH-CONSTANT, R. H. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature,* 477, 203-206.

KARLSSON, M. 2001. *Primula* culture and production. *Horticularal Technology,* 11, 627-635.

KEES, U. R., TERRY, P. A., FORD, J., EVERETT, J., MURCH, A., DE KLERK, N. & BAKER, D. L. 2005. Detection of hemizygous deletions in genomic DNA from leukaemia specimens for the diagnosis of patients. *Leukaemia Research,* 29, 165-171.

KELLER, B., THOMSON, J. D. & CONTI, E. 2014. Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. *Functional Ecology,* 28, 1413-1425.

KENT, W. J. 2002. BLAT - The BLAST-like alignment tool. *Genome Research,* 12, 656-664.

KERSTETTER, R., VOLLBRECHT, E., LOWE, B., VEIT, B., YAMAGUCHI, J. & HAKE, S. 1994. Sequence-analysis and expression patterns divide the maize *KNOTTED1*-like homeobox genes into 2 classes. *Plant Cell,* 6, 1877-1887.

KIM, D., LANGMEAD, B. & SALZBERG, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods,* 12, 357-360.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013a. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology,* 14, 1-13.

KIM, H. J., CHIANG, Y. H., KIEBER, J. J. & SCHALLER, G. E. 2013b. SCF^{KMD} controls cytokinin signaling by regulating the degradation of type-B response regulators. *Proceedings of the National Academy of Sciences,* 110, 10028-10033.

KIM, S., YOO, M. J., ALBERT, V. A., FARRIS, J. S., SOLTIS, P. S. & SOLTIS, D. E. 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *American Journal of Botany,* 91, 2102-2118.

KIMURA, M. 1984. *The neutral theory of molecular evolution*, Cambridge University Press.

KLEIN, A. M., VAISSIÈRE, B. E., CANE, J. H., STEFFAN-DEWENTER, I., CUNNINGHAM, S. A., KREMEN, C. & TSCHARNTKE, T. 2007. Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society B: Biological Sciences,* 274, 303-313.

KOHN, D., MURRELL, G., PARKER, J. & WHITEHORN, M. 2005. What Henslow taught Darwin. *Nature,* 436, 643-645.

KOLE, C., MUTHAMILARASAN, M., HENRY, R., EDWARDS, D., SHARMA, R., ABBERTON, M., BATLEY, J., BENTLEY, A., BLAKENEY, M., BRYANT, J., CAI, H., CAKIR, M., CSEKE, L., COCKRAM, J., OLIVEIRA, A., PACE, C., DEMPEWOLF, H., ELLISON, S., GEPTS, P., GREENLAND, A., HALL, A., HORI, K., HUGHES, S.,

HUMPHREYS, M., IORIZZO, M., ISMAIL, A., MARSHALL, A., MAYES, S., NGUYEN, H., OGBONNAYA, F., ORTIZ, R., PATERSON, A., SIMON, P., TOHME, J., TUBEROSA, R., VALLIYODAN, B., VARSHNEY, R., WULLSCHLEGER, S., YANO, M. & PRASAD, M. 2015. Application of genomics-assisted breeding for generation of climate resilient crops: progress and prospects. *Frontiers in Plant Science,* 6.

KONISHI, T., IWATA, H., YASHIRO, K., TSUMURA, Y., OHSAWA, R., YASUI, Y. & OHNISHI, O. 2006. Development and characterization of microsatellite markers for common buckwheat. *Breeding Science,* 56, 277-285.

KOOPMAN, P., GUBBAY, J., VIVIAN, N., GOODFELLOW, P. & LOVELL-BADGE, R. 1991. Male development of chromosomally female mice transgenic for *Sry*. *Nature,* 351, 117-121.

KOPYLOVA, E., NOE, L. & TOUZET, H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics,* 28, 3211-3217.

KOREN, S., SCHATZ, M. C., WALENZ, B. P., MARTIN, J., HOWARD, J. T., GANAPATHY, G., WANG, Z., RASKO, D. A., MCCOMBIE, W. R., JARVIS, E. D. & PHILLIPPY, A. M. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology,* 30.

KRAMER, E. M., HOLAPPA, L., GOULD, B., JARAMILLO, M. A., SETNIKOV, D. & SANTIAGO, P. M. 2007. Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot *Aquilegia*. *The Plant Cell,* 19, 750-766.

KREFT, I. & LUTHAR, Z. 1990. Buckwheat — a low input plant. *In:* EL BASSAM, N., DAMBROTH, M. & LOUGHMAN, B. C. (eds.) *Genetic Aspects of Plant Mineral Nutrition.* Dordrecht: Springer Netherlands.

KRIZEK, B. A. & MEYEROWITZ, E. M. 1996. The Arabidopsis homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the B class organ identity function. *Development,* 122, 11-22.

KUBAT, Z., ZLUVOVA, J., VOGEL, I., KOVACOVA, V., CERMAK, T., CEGAN, R., HOBZA, R., VYSKOT, B. & KEJNOVSKY, E. 2014. Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. *New Phytologist,* 202, 662-678.

KUNTE, K., ZHANG, W., TENGER-TROLANDER, A., PALMER, D. H., MARTIN, A., REED, R. D., MULLEN, S. P. & KRONFORST, M. R. 2014. *doublesex* is a mimicry supergene. *Nature,* 507, 229-232.

KURIAN, V. 1996. Investigations into the breeding system supergene in *Primula*. *PhD Thesis, University of Newcastle upon Tyne, UK*.

KURIAN, V. & RICHARDS, A. J. 1997. A new recombinant in the heteromorphy '*S*' supergene in *Primula*. *Heredity,* 78, 383-390.

KURSEL, L. E. & MALIK, H. S. 2016. Centromeres. *Current Biology,* 26, R487-R490.

LABOMBARDA, P., BUSTI, A., CACERES, M. E., PUPILLI, F. & ARCIONI, S. 2002. An AFLP marker tightly linked to apomixis reveals hemizygosity in a portion of the apomixis-controlling locus in *Paspalum simplex*. *Genome,* 45, 513-519.

LABONNE, J. D. J., GOULTIAEVA, A. & SHORE, J. S. 2009. High-resolution mapping of the *S* locus in *Turnera* leads to the discovery of three genes tightly associated with the *S* alleles. *Molecular Genetics and Genomics,* 281, 673-685.

LABONNE, J. D. J. & SHORE, J. S. 2011. Positional cloning of the *s* haplotype determining the floral and incompatibility phenotype of the long-styled morph of distylous *Turnera subulata*. *Molecular Genetics and Genomics,* 285, 101-111.

LABONNE, J. D. J., TAMARI, F. & SHORE, J. S. 2010. Characterization of X-ray-generated floral mutants carrying deletions at the *S* locus of distylous *Turnera subulata*. *Heredity,* 105, 235-243.

LABONNE, J. D. J., VAISMAN, A. & SHORE, J. S. 2008. Construction of a first genetic map of distylous *Turnera* and a fine-scale map of the *S* locus region. *Genome,* 51, 471-478.

LANFEAR, R., HUA, X. & WARREN, D. L. 2016. Estimating the effective sample size of tree topologies from bayesian phylogenetic analyses. *Genome Biology and Evolution,* 8, 2319-2332.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23.

LEWIS, D. 1949. Incompatibility in flowering plants. *Biological Reviews,* 24, 471-496.

LEWIS, D. 1954. Comparative incompatibility in angiosperms and fungi. *Advances in Genetics Incorporating Molecular Genetic Medicine,* 6, 235-285.

LEWIS, D. 1982. Incompatibility, stamen movement and pollen economy in a heterostyled tropical forest tree, *Cratoxylum formosum* (Guttiferae). *Proceedings of the Royal Society of London B: Biological Sciences,* 214, 273-283.

LEWIS, D. & JONES, D. A. 1992. The genetics of heterostyly. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* Berlin: Springer Verlag.

LI, E., BHARGAVA, A., QIANG, W., FRIEDMANN, M. C., FORNERIS, N., SAVIDGE, R. A., JOHNSON, L. A., MANSFIELD, S. D., ELLIS, B. E. & DOUGLAS, C. J. 2012. The Class II *KNOX* gene *KNAT7* negatively regulates secondary wall formation in *Arabidopsis* and is functionally conserved in *Populus*. *New Phytologist,* 194, 102-115.

LI, E., WANG, S., LIU, Y., CHEN, J. G. & DOUGLAS, C. J. 2011a. OVATE FAMILY PROTEIN4 (OFP4) interaction with KNAT7 regulates secondary cell wall formation in *Arabidopsis thaliana*. *Plant Journal,* 67, 328-341.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv,* 1303.3997.

LI, H. & 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. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, 2078-2079.

LI, J. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics,* 13, 523-538.

LI, J., COCKER, J. M., WRIGHT, J., WEBSTER, M. A., MCMULLAN, M., AYLING, S., SWARBRECK, D., CACCAMO, M., OOSTERHOUT, C. & GILMARTIN, P. M. 2016. Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nature Plants,* 2, 16188.

LI, J., DUDAS, B., WEBSTER, M. A., COOK, H. E., DAVIES, B. H. & GILMARTIN, P. M. 2010. *Hose in Hose,* an *S* locus-linked mutant of *Primula vulgaris* is caused by an unstable mutation at the *Globosa* locus. *Proceedings of the National Academy of Sciences,* 107, 5664-5668.

LI, J., WEBSTER, M., DUDAS, B., COOK, H., MANFIELD, I., DAVIES, B. & GILMARTIN, P. M. 2008. The *S* locus-linked *Primula* homeotic mutant *sepaloid* shows characteristics of a B-function mutant but does not result from mutation in a B-function gene. *Plant Journal,* 56, 1-12.

LI, J., WEBSTER, M. A., FURUYA, M. & GILMARTIN, P. M. 2007. Identification and characterization of pin and thrum alleles of two genes that co-segregate with the *Primula S* locus. *Plant Journal,* 51, 18-31.

LI, J., WEBSTER, M. A., SMITH, M. C. & GILMARTIN, P. M. 2011b. Floral heteromorphy in *Primula vulgaris*: progress towards isolation and characterization of the *S* locus. *Annals of Botany,* 108, 715-726.

LI, J., WEBSTER, M. A., WRIGHT, J., COCKER, J. M., M., S., BADAKSHI, F., HESLOP-HARRISON, P. & GILMARTIN, P. M. 2015a. Integration of genetic and physical maps of the *Primula vulgaris S* locus and localization by chromosome *in situ* hybridisation *New Phytologist,* 208, 137-148.

LI, R., YU, C., LI, Y., LAM, T. W., YIU, S. M., KRISTIANSEN, K. & WANG, J. 2009b. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25, 1966-1967.

LI, W., TUCKER, A. E., SUNG, W., THOMAS, W. K. & LYNCH, M. 2009c. Extensive, recent intron gains in *Daphnia* populations. *Science,* 326, 1260-1262.

LI, Z. F., ZHANG, Y. C. & CHEN, Y. Q. 2015b. miRNAs and lncRNAs in reproductive development. *Plant Science,* 238, 46-52.

LIEBER, M. R. 2010. The mechanism of double-strand dna break repair by the nonhomologous DNA end joining pathway. *Annual review of biochemistry,* 79, 181-211.

LINCOLN, C., LONG, J., YAMAGUCHI, J., SERIKAWA, K. & HAKE, S. 1994. A *KNOTTED1*-like homeobox gene in Arabidopsis is expressed in the vegetative meristem and dramatically alters leaf morphology when over-expressed in transgenic plants. *Plant Cell,* 6, 1859-1876.

LINNAEUS, C. 1792. *Botanicorum Principis, Philisophia Botanica.* London, Andrew Murray (pp. 240-241)

LIU, B., SHI, Y., YUAN, J., HU, X., ZHANG, H., LI, N., LI, Z., CHEN, Y., MU, D. & FAN, W. 2013. Estimation of genomic characteristics by analyzing *k*-mer frequency in *de novo* genome projects. *arXiv,* 1308.2012

LIU, C. M., WONG, T., WU, E., LUO, R., YIU, S. M., LI, Y., WANG, B., YU, C., CHU, X., ZHAO, K., LI, R. & LAM, T. W. 2012. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics,* 28, 878-879.

LLOYD, D. G. & WEBB, C. J. 1992a. The Evolution of Heterostyly. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* Berlin: Springer Verlag.

LLOYD, D. G. & WEBB, C. J. 1992b. The Selection of Heterostyly. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* Berlin: Springer Verlag.

LONG, J. A., MOAN, E. I., MEDFORD, J. I. & BARTON, M. K. 1996. A member of the KNOTTED class of homeodomain proteins encoded by the *STM* gene of *Arabidopsis*. *Nature,* 379, 66-69.

LÓPEZ-PÉREZ, M., MARTIN-CUADRADO, A. B. & RODRIGUEZ-VALERA, F. 2014. Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes. *Frontiers in Genetics,* 5, 147.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y., TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, D. W., YIU, S. M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T. W. & WANG, J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience,* 1, 18-18.

MAGALLÓN, S., GÓMEZ-ACEVEDO, S., SÁNCHEZ-REYES, L. L. & HERNÁNDEZ-HERNÁNDEZ, T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist,* 207, 437-453.

MANFIELD, I. W., PAVLOV, V. K., LI, J. H., COOK, H. E., HUMMEL, F. & GILMARTIN, P. M. 2005. Molecular characterization of DNA sequences from the *Primula vulgaris S* locus. *Journal of Experimental Botany,* 56, 1177-1188.

MAPLESON, D., GARCIA ACCINELLI, G., KETTLEBOROUGH, G., WRIGHT, J. & CLAVIJO, B. 2016. KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *bioRxiv,* 10.1101/064733

MARÇAIS, G., & KINGSFORD, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics,* 27, 764-770.

MARTIN, D. P., MURRELL, B., GOLDEN, M., KHOOSAL, A. & MUHIRE, B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution,* 1, vev003.

MAST, A. R., KELSO, S. & CONTI, E. 2006. Are any primroses (*Primula*) primitively monomorphic? *New Phytologist,* 171, 605-616.

MAST, A. R., KELSO, S., RICHARDS, A. J., LANG, D. J., FELLER, D. M. S. & CONTI, E. 2001. Phylogenetic relationships in *Primula* L. and related genera (Primulaceae) based on noncoding chloroplast DNA. *International Journal of Plant Sciences,* 162, 1381-1400.

MATHER, K. 1939. Crossing over and heterochromatin in the X chromosome of *Drosophila melanogaster*. *Genetics,* 24, 413-435.

MATHER, K. 1950. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution,* 4, 340-352.

MATHER, K. & DE WINTON, D. 1941. Adaptation and counter-adaptation of the breeding system in *Primula*. *Annals of Botany,* 5, 197-311.

MAYER, K. F. X., ROGERS, J., DOLEŽEL, J., POZNIAK, C., EVERSOLE, K., FEUILLET, C., GILL, B., FRIEBE, B., LUKASZEWSKI, A. J., SOURDILLE, P., ENDO, T. R., KUBALÁKOVÁ, M., ČÍHALÍKOVÁ, J., DUBSKÁ, Z., VRÁNA, J., ŠPERKOVÁ, R., ŠIMKOVÁ, H., FEBRER, M., CLISSOLD, L., MCLAY, K., SINGH, K., CHHUNEJA, P.,

SINGH, N. K., KHURANA, J., AKHUNOV, E., CHOULET, F., ALBERTI, A., BARBE, V., WINCKER, P., KANAMORI, H., KOBAYASHI, F., ITOH, T., MATSUMOTO, T., SAKAI, H., TANAKA, T., WU, J., OGIHARA, Y., HANDA, H., MACLACHLAN, P. R., SHARPE, A., KLASSEN, D., EDWARDS, D., BATLEY, J., OLSEN, O. A., SANDVE, S. R., LIEN, S., STEUERNAGEL, B., WULFF, B., CACCAMO, M., AYLING, S., RAMIREZ-GONZALEZ, R. H., CLAVIJO, B. J., WRIGHT, J., PFEIFER, M., SPANNAGL, M., MARTIS, M. M., MASCHER, M., CHAPMAN, J., POLAND, J. A., SCHOLZ, U., BARRY, K., WAUGH, R., ROKHSAR, D. S., MUEHLBAUER, G. J., STEIN, N., GUNDLACH, H., ZYTNICKI, M., JAMILLOUX, V., QUESNEVILLE, H., WICKER, T., FACCIOLI, P., COLAIACOVO, M., STANCA, A. M., BUDAK, H., CATTIVELLI, L., GLOVER, N., PINGAULT, L., PAUX, E., SHARMA, S., APPELS, R., BELLGARD, M., CHAPMAN, B., NUSSBAUMER, T., BADER, K. C., RIMBERT, H., WANG, S., KNOX, R., KILIAN, A., ALAUX, M., ALFAMA, F., COUDERC, L., GUILHOT, N., VISEUX, C., LOAEC, M., KELLER, B. & PRAUD, S. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science,* 345, 1251788.

MCCUBBIN, A. 2008. Heteromorphic self-Incompatibility in *Primula*: Twenty-first Century tools promise to unravel a classic nineteenth Century model system. *In:* FRANKLIN-TONG, V. E. (ed.) *Self-Incompatibility in Flowering Plants – Evolution, Diversity and mechanisms.* Berlin: Springer Verlag.

MCCUBBIN, A. G., LEE, C. & HETRICK, A. 2006. Identification of genes showing differential expression between morphs in developing flowers of *Primula vulgaris*. *Sexual Plant Reproduction,* 19, 63-72.

MEANEY, S. 2005. Is C-26 hydroxylation an evolutionarily conserved steroid inactivation mechanism? *FASEB Journal,* 19, 1220-1224.

MILJUS-DUKIC, J., NINKOVIC, S., RADOVIC, S., MAKSIMOVIC, V., BRKLJACIC, J. & NESKOVIC, M. 2004. Detection of proteins possibly involved in self-incompatibility response in distylous buckwheat. *Biologia Plantarum,* 48, 293-296.

MIZOI, J., SHINOZAKI, K. & YAMAGUCHI-SHINOZAKI, K. 2012. AP2/ERF family transcription factors in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms,* 1819, 86-96.

MOSTOVOY, Y., LEVY-SAKIN, M., LAM, J., LAM, E. T., HASTIE, A. R., MARKS, P., LEE, J., CHU, C., LIN, C., DZAKULA, Z., CAO, H., SCHLEBUSCH, S. A., GIORDA, K., SCHNALL-LEVIN, M., WALL, J. D. & KWOK, P. Y. 2016. A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat Meth,* 13, 587-590.

NAGARAJAN, N. & POP, M. 2013. Sequence assembly demystified. *Nature Reviews Genetics,* 14, 157-167.

NAGPAL, P., ELLIS, C. M., WEBER, H., PLOENSE, S. E., BARKAWI, L. S., GUILFOYLE, T. J., HAGEN, G., ALONSO, J. M., COHEN, J. D., FARMER, E. E., ECKER, J. R. & REED, J. W. 2005. Auxin response factors ARF6 and ARF8 promote jasmonic acid production and flower maturation. *Development,* 132, 4107-4118.

NAGY, A. 2000. Cre recombinase: the universal reagent for genome tailoring. *Genesis,* 26, 99-109.

NAIKI, A. 2012. Heterostyly and the possibility of its breakdown by polyploidization. *Plant Species Biology,* 27, 3-29.

NAM, J., DEPAMPHILIS, C. W., MA, H. & NEI, M. 2003. Antiquity and Evolution of the MADS-Box Gene Family Controlling Flower Development in Plants. *Molecular Biology and Evolution,* 20, 1435-1447.

NG, P. C. & HENIKOFF, S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research,* 31, 2812-3814.

NONG, G. & ZHANG, S. 2007. Efficient algorithms for the inverse sort transform. *IEEE Transactions on Computers,* 56, 1564-1574.

NOWAK, M. D., RUSSO, G., SCHLAPBACH, R., HUU, C. N., LENHARD, M. & CONTI, E. 2015. The draft genome of *Primula veris* yields insight into the molecular basis of heterostyly. *Genome Biology,* 16, 16.

NURMBERG, P. L., KNOX, K. A., YUN, B. W., MORRIS, P. C., SHAFIEI, R., HUDSON, A. & LOAKE, G. J. 2007. The developmental selector AS1 is an evolutionarily conserved regulator of the plant immune response. *Proceedings Of The National Academy Of Sciences,* 104, 18795-18800.

NYLANDER, J. A. A., WILGENBUSCH, J. C., WARREN, D. L. & SWOFFORD, D. L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics,* 24, 581-583.

OHTA, M., OHME-TAKAGI, M. & SHINSHI, H. 2000. Three ethylene-responsive transcription factors in tobacco with distinct transactivation functions. *The Plant Journal,* 22, 29-38.

OLESEN, J. M. 1979. Floral morphology and pollen flow in the heterostylous species *Pulmonaria obscura* Dumort (Boraginaceae). *New Phytologist,* 82, 757-767.

OLIVEIRA, P. H., LEMOS, F., MONTEIRO, G. A. & PRAZERES, D. M. F. 2008. Recombination frequency in plasmid DNA containing direct repeats — predictive correlation with repeat and intervening sequence length. *Plasmid,* 60, 159-165.

ORNDUFF, R. 1979. Pollen flow in a population of *Primula vulgaris* Huds. *Botanical Journal of the Linnean Society,* 78, 1-10.

ORNDUFF, R. 1992. Historical Perspectives on Heterostyly. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* London: Springer Verlag.

OTA, T., MORI, M., MORI, M., MATSUMOTO, D., OHNISHI, O., CAMPBELL, C. & YASUI, Y. 2006. Positional cloning of the genes determining heterostylous self-incompatibility in buckwheat. *Genes & Genetic Systems,* 81, 444.

OZIAS-AKINS, P., ROCHE, D. & HANNA, W. W. 1998. Tight clustering and hemizygosity of apomixis-linked molecular markers in *Pennisetum squamulatum* implies genetic control of apospory by a divergent locus that may have no allelic form in sexual genotypes. *Proceedings of the National Academy of Sciences,* 95, 5127-5132.

PARRA, G., BRADNAM, K. & KORF, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,* 23, 1061-1067.

PARRA, G., BRADNAM, K., NING, Z., KEANE, T. & KORF, I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res,* 37, 289-297

PASAM, R. K., SHARMA, R., MALOSETTI, M., VAN EEUWIJK, F. A., HASENEYER, G., KILIAN, B. & GRANER, A. 2012. Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biology,* 12, 1-22.

PATON, A. J., BRUMMITT, N., GOVAERTS, R., HARMAN, K., HINCHCLIFFE, S., ALLKIN, B. & LUGHADHA, E. N. 2008. Towards target 1 of the global strategy for plant conservation: a working list of all known plant species - progress and prospects. *Taxon,* 57, 602-611.

PAZ, M. M. & VEILLEUX, R. E. 1999. Influence of culture medium and in vitro conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. *Plant Breeding,* 118, 53-57.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. 2011. Scikit-learn: machine learning in python. *The Journal of Machine Learning Research,* 12, 2825-2830.

PELLOW, C. 1928. *Report for the Year 1928.* Merton, UK, The John Innes Horticultural Institution.

PHADNIS, N., SIA, R. A. & SIA, E. A. 2005. Analysis of repeat-mediated deletions in the mitochondrial genome of *Saccharomyces cerevisiae*. *Genetics,* 171, 1549-1559.

PHILIPPE, H., BRINKMANN, H., LAVROV, D. V., LITTLEWOOD, D. T. J., MANUEL, M., WÖRHEIDE, G. & BAURAIN, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology,* 9, e1000602.

PIPER, J. G., CHARLESWORTH, B. & CHARLESWORTH, D. 1984. A high-rate of self-fertilization and increased seed fertility of homostyle primroses. *Nature,* 310, 50-51.

PIPER, J. G. & CHARLESWORTH, D. 1986. Breeding system evolution in *Primula vulgaris* and the role of reproductive assurance. *Heredity,* 56, 207-217.

POTTS, S. G., BIESMEIJER, J. C., KREMEN, C., NEUMANN, P., SCHWEIGER, O. & KUNIN, W. E. 2010. Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution,* 25, 345-353.

PRYSZCZ, L. P. & GABALDÓN, T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, 44, e113.

PRYSZCZ, L. P., NÉMETH, T., GÁCSER, A. & GABALDÓN, T. 2014. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biology and Evolution,* 6, 1069-1078.

PUTTICK, M. N., CLARK, J. & DONOGHUE, P. C. J. 2015. Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms. *Proceedings of the Royal Society B: Biological Sciences,* 282, 20152289.

QU, W., HASHIMOTO, S. I. & MORISHITA, S. 2009. Efficient frequency-based *de novo* short-read clustering for error trimming in next-generation sequencing. *Genome Research,* 19, 1309-1315.

RAINERI, E., FERRETTI, L., ESTEVE-CODINA, A., NEVADO, B., HEATH, S. & PÉREZ-ENCISO, M. 2012. SNP calling by sequencing pooled samples. *BMC Bioinformatics,* 13, 239-239.

RAMBAUT, A., LAM, T. T., CARVALHO, L. M. & PYBUS, O. G. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution,* 2, vew007.

RAY, A. & LANGER, M. 2002. Homologous recombination: ends as the means. *Trends in Plant Science,* 7, 435-440.

RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13, 278-289.

RICHARDS, A. J. 1997. Plant Breeding Systems 2nd edition. *Chapman and Hall, London.*

RICHARDS, A. J. 2001. Does low biodiversity resulting from modern agricultural practice affect crop pollination and yield? *Annals of Botany,* 88, 165-172.

RICHARDS, A. J. & EDWARDS, B. 2003. *Primula*. Portland (OR): Timber Press.

RICHARDS, J. 2014. *Primula*. London: Pavilion Books.

RICHARDS, J. H. & BARRETT, S. C. H. 1992. The Development of Heterostyly. *In:* BARRETT, S. C. H. (ed.) *Evolution and Function of Heterostyly.* Berlin: Springer Verlag.

RICHARDS, S. A., WILLIAMS, N. M. & HARDER, L. D. 2009. Variation in pollination: causes and consequences for plant reproduction. *The American Naturalist,* 174, 382-398.

ROBERTSON, G., SCHEIN, J., CHIU, R., CORBETT, R., FIELD, M., JACKMAN, S. D., MUNGALL, K., LEE, S., OKADA, H. M., QIAN, J. Q., GRIFFITH, M., RAYMOND, A., THIESSEN, N., CEZARD, T., BUTTERFIELD, Y. S., NEWSOME, R., CHAN, S. K., SHE, R., VARHOL, R., KAMOH, B., PRABHU, A. L., TAM, A., ZHAO, Y., MOORE, R. A., HIRST, M., MARRA, M. A., JONES, S. J. M., HOODLESS, P. A. & BIROL, I. 2010. *De novo* assembly and analysis of RNA-seq data. *Nature Methods,* 7, 909-912.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26: 139-40

RODGERS-MELNICK, E., MANE, S. P., DHARMAWARDHANA, P., SLAVOV, G. T., CRASTA, O. R., STRAUSS, S. H., BRUNNER, A. M. & DIFAZIO, S. P. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research,* 22, 95-105.

ROQUE, E., FARES, M. A., YENUSH, L., ROCHINA, M. C., WEN, J., MYSORE, K. S., GÓMEZ-MENA, C., BELTRÁN, J. P. & CAÑAS, L. A. 2016. Evolution by gene duplication of *Medicago truncatula PISTILLATA*-like transcription factors. *Journal of Experimental Botany,* 67, 1805-1817.

ROSOV, S. & SCREBTSOVA, N. Honey bees and selective fertilization of plants. *XVII International Beekeeping Congress*, 1958. 494-501.

ROZEN, S., SKALETSKY, H., MARSZALEK, J. D., MINX, P. J., CORDUM, H. S., WATERSTON, R. H., WILSON, R. K. & PAGE, D. C. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature,* 423, 873-876.

RUFFALO, M., LAFRAMBOISE, T. & KOYUTÜRK, M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics,* 27, 2790-2796.

RUTSCHMANN, F., ERIKSSON, T., SALIM, K. A. & CONTI, E. 2007. Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points. *Systematic Biology,* 56, 591-608.

SALEH, A. & PAGÉS, M. 2003. Plant AP2/ERF transcription factors. *Genetika,* 35, 37-50.

SATO, S., TABATA, S., HIRAKAWA, H., ASAMIZU, E., SHIRASAWA, K. & ISOBE, S. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature,* 485, 635-541.

SAYADI, A., IMMONEN, E., BAYRAM, H. & ARNQVIST, G. 2016. The *de novo* transcriptome and its functional annotation in the eed beetle *Callosobruchus maculatus*. *PLoS ONE,* 11, e0158565.

SCHADT, E. E., TURNER, S. & KASARSKIS, A. 2010. A window into third-generation sequencing. *Human Molecular Genetics,* 19, R227-R240.

SCHATZ, M. C., DELCHER, A. L. & SALZBERG, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research,* 20, 1165-1173.

SCHENK, J. J. 2016. Consequences of Secondary Calibrations on Divergence Time Estimates. *PLoS ONE,* 11, e0148228.

SCHLOTTERER, C., TOBLER, R., KOFLER, R. & NOLTE, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics,* 15, 749-763.

SCHULZ, M. H., ZERBINO, D. R., VINGRON, M. & BIRNEY, E. 2012. Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics,* 28, 1086-1092.

SCHURCH, N.J., SCHOFIELD, P., GIERLIŃSKI, M., COLE, C., SHERSTNEV, A., SINGH, V., WROBEL, N., GHARBI, K., SIMPSON, G.G., OWEN-HUGHES, T., BLAXTER, M., BARTON, G. J. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA,* 22(6): 839-851.

SCHWANDER, T., LIBBRECHT, R. & KELLER, L. 2014. Supergenes and complex phenotypes. *Current Biology,* 24, R288-R294.

SCHWARTZ, D., LI, X., HERNANDEZ, L., RAMNARAIN, S., HUFF, E. & WANG, Y. 1993. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science,* 262, 110-114.

SERIKAWA, K. A., MARTINEZLABORDA, A., KIM, H. S. & ZAMBRYSKI, P. C. 1997. Localization of expression of *KNAT3*, a class 2 *Knotted1*-like gene. *Plant Journal,* 11, 853-861.

SERIKAWA, K. A., MARTINEZLABORDA, A. & ZAMBRYSKI, P. 1996. Three *knotted1*-like homeobox genes in *Arabidopsis*. *Plant Molecular Biology,* 32, 673-683.

SHANI, E., BURKO, Y., BEN-YAAKOV, L., BERGER, Y., AMSELLEM, Z., GOLDSHMIDT, A., SHARON, E. & ORI, N. 2009. Stage-specific regulation of *Solanum lycopersicum* leaf maturation by Class 1 KNOTTED1-like homeobox Proteins. *Plant Cell*, 21, 3078-3092.

SHARMA, B., YANT, L., HODGES, S. A. & KRAMER, E. M. 2014. Understanding the development and evolution of novel floral form in *Aquilegia*. *Current Opinion in Plant Biology,* 17, 22-27.

SHIVANNA, K. R., HESLOP-HARRISON, J. & HESLOP-HARRISON, Y. 1981. Heterostyly in *Primula*. 2. Sites of pollen inhibition, and effects of pistil constituents on compatible and incompatible pollen tube growth. *Protoplasma,* 107, 319-337.

SILJAK-YAKOVLEV, S., PUSTAHIJA, F., ŠOLIĆ, E., BOGUNIĆ, F., MURATOVIĆ, E., BAŠIĆ, N., CATRICE, O. & BROWN, S. 2010. Towards a genome size and chromosome number database of Balkan flora: C-values in 343 taxa with novel values for 242. *Advanced Science Letters,* 3, 190-213.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31, 3210-3212.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J. & BIROL, I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Reearch,* 19, 1117-1123.

SINCLAIR, A. H., BERTA, P., PALMER, M. S., HAWKINS, J. R., GRIFFITHS, B. L., SMITH, M. J., FOSTER, J. W., FRISCHAUF, A. M., LOVELL-BADGE, R. & GOODFELLOW, P. N. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature,* 346, 240-244.

SKALETSKY, H., KURODA-KAWAGUCHI, T., MINX, P. J., CORDUM, H. S., HILLIER, L., BROWN, L. G., REPPING, S., PYNTIKOVA, T., ALI, J., BIERI, T., CHINWALLA, A., DELEHAUNTY, A., DELEHAUNTY, K., DU, H., FEWELL, G., FULTON, L., FULTON, R., GRAVES, T., HOU, S. F., LATRIELLE, P., LEONARD, S., MARDIS, E., MAUPIN, R., MCPHERSON, J., MINER, T., NASH, W., NGUYEN, C., OZERSKY, P., PEPIN, K., ROCK, S., ROHLFING, T., SCOTT, K., SCHULTZ, B., STRONG, C., TIN-WOLLAM, A., YANG, S. P., WATERSTON, R. H., WILSON, R. K., ROZEN, S. & PAGE, D. C. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* 423, 825-837.

SLATER, G. & BIRNEY, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics,* 6, 31.

SMALL, K. S., BRUDNO, M., HILL, M. M. & SIDOW, A. 2007. A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome biology,* 8, R41.

SOMMER, H., BELTRAN, J. P., HUIJSER, P., PAPE, H., LÖNNIG, W., SAEDLER, H. & SCHWARZ-SOMMER, Z. 1990. *Deficiens*, a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus*: the protein shows homology to transcription factors. *The EMBO Journal,* 9, 605.

SONESON, C. & DELORENZI, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics,* 14, 1-18.

SONG, F., LI, H., JIANG, P., ZHOU, X., LIU, J., SUN, C., VOGLER, A. P. & CAI, W. 2016. Capturing the phylogeny of holometabola with mitochondrial genome data and bayesian site-heterogeneous mixture models. *Genome Biology and Evolution,* 8, 1411-1426.

SONODA, E., HOCHEGGER, H., SABERI, A., TANIGUCHI, Y. & TAKEDA, S. 2006. Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair,* 5, 1021-10029.

STANKE, M. & MORGENSTERN, B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research,* 33, W465-W467.

STEVENS, V. A. M. & MURRAY, B. G. 1982. Studies on heteromorphic self-incompatibility systems: Physiological aspects of the incompatibility system of *Primula obconica. Theoret. Appl. Genetics,* 61, 245.

STEWART, J. B., FREYER, C., ELSON, J. L., WREDENBERG, A., CANSU, Z., TRIFUNOVIC, A. & LARSSON, N.G. 2008. Strong purifying selection in transmission of mammalian mitochondrial dna. *PLoS Biology,* 6, e10.

STONE, J. L. 1995. Pollen donation patterns in a tropical distylous shrub (*Psychotria suerrensis*; Rubiaceae). *American Journal of Botany*, 1390-1398.

SU, Y. H., LIU, Y. B., ZHOU, C., LI, X. M. & ZHANG, X. S. 2016. The microRNA167 controls somatic embryogenesis in *Arabidopsis* through regulating its target genes *ARF6* and *ARF8*. *Plant Cell, Tissue and Organ Culture (PCTOC),* 124, 405-417.

SUPPLE, M. A., HINES, H. M., DASMAHAPATRA, K. K., LEWIS, J. J., NIELSEN, D. M., LAVOIE, C., RAY, D. A., SALAZAR, C., MCMILLAN, W. O. & COUNTERMAN, B. A. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Research*, 23, 1248-1257.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution,* 30, 2725-2729.

TANG, H., LYONS, E. & TOWN, C. D. 2015. Optical mapping in plant comparative genomics. *GigaScience,* 4, 1-6.

TEMSCH, E. M., TEMSCH, W., EHRENDORFER-SCHRATT, L. & GREILHUBER, J. 2010. Heavy metal pollution, selection, and genome size: the species of the Žerjav study revisited with flow cytometry. *Journal of Botany,* 2010, 596542.

TEN BOSCH, J. R. & GRODY, W. W. 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics,* 10, 484-492.

TEUNE, J. H. & STEGER, G. 2010. NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome. *Journal of Nucleic Acids,* 2010, 495904.

THE ANGIOSPERM PHYLOGENY. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181: 1-20.

THOMAS, J. W., CACERES, M., LOWMAN, J. J., MOREHOUSE, C. B., SHORT, M. E., BALDWIN, E. L., MANEY, D. L. & MARTIN, C. L. 2008. The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. *Genetics,* 179, 1455-1468.

THOMPSON, M. J. & JIGGINS, C. D. 2014. Supergenes and their role in evolution. *Heredity,* 113, 1-8.

TRAPNELL, C., HENDRICKSON, D. G., SAUVAGEAU, M., GOFF, L., J.L., R. & PACHER, L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31, 46-53.

TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25, 1105-1111.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols,* 7, 562-578.

TRIANT, D. A. & PEARSON, W. R. 2015. Most partial domains in proteins are alignment and annotation artifacts. *Genome Biology,* 16, 99.

TROBNER, W., RAMIREZ, L., MOTTE, P., HUE, I., HUIJSER, P., LONNIG, W. E., SAEDLER, H., SOMMER, H. & SCHWARZ-SOMMER, Z. 1992. *GLOBOSA* - A homeotic gene which interacts with DEFICIENS in the control of *Antirrhinum* floral organogenesis. *EMBO Journal,* 11, 4693-4704.

TRUERNIT, E., SIEMERING, K. R., HODGE, S., GRBIC, V. & HASELOFF, J. 2006. A map of *KNAT* gene expression in the Arabidopsis root. *Plant Molecular Biology,* 60, 1-20.

TURAN, S., GALLA, M., ERNST, E., QIAO, J., VOELKEL, C., SCHIEDLMEIER, B., ZEHE, C. & BODE, J. 2011. Recombinase-mediated cassette exchange (RMCE): traditional concepts and current challenges. *Journal of Molecular Biology,* 407, 193-221.

TURK, E. M., FUJIOKA, S., SETO, H., SHIMADA, Y., TAKATSUTO, S., YOSHIDA, S., DENZEL, M. A., TORRES, Q. I. & NEFF, M. M. 2003. CYP72B1 inactivates brassinosteroid hormones: An intersection between photomorphogenesis and plant steroid signal transduction. *Plant Physiology,* 133, 1643-1653.

UNAMBA, C. I. N., NAG, A. & SHARMA, R. K. 2015. Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. *Frontiers in Plant Science,* 6, 1074.

USHIJIMA, K., NAKANO, R., BANDO, M., SHIGEZANE, Y., IKEDA, K., NAMBA, Y., KUME, S., KITABATA, T., MORI, H. & KUBO, Y. 2012. Isolation of the floral morph-related genes in heterostylous flax (*Linum grandiflorum*): the genetic polymorphism and the transcriptional and post-transcriptional regulations of the *S* locus. *Plant Journal,* 69, 317-331.

VAN DE PASSE, C. 1614. *Hortus Floridu*. Utrecht: Hans Woutneel.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics,* 30, 418-426.

VAN DIJK, W. 1943. La decóuverte de l'hétérostylie chez *Primula* par Ch. de l'Écluse et P. Reneaulme. *Nedelandsch Kruidkundig Archief,* 53, 81-85.

VAN ZEE, J. P., SCHLUETER, J. A., SCHLUETER, S., DIXON, P., SIERRA, C. A. B. & HILL, C. A. 2016. Paralog analyses reveal gene duplication events and genes under positive selection in *Ixodes scapularis* and other ixodid ticks. *BMC Genomics,* 17, 241.

VANDENBUSSCHE, M., ZETHOF, J., ROYAERT, S., WETERINGS, K. & GERATS, T. 2004. The duplicated B-class heterodimer model: Whorl-specific effects and complex genetic interactions in *Petunia hybrida* flower development. *Plant Cell,* 16, 741-754.

VANNESTE, K., VAN DE PEER, Y. & MAERE, S. 2013. Inference of Genome Duplications from Age Distributions Revisited. *Molecular Biology and Evolution,* 30, 177-190.

VARSHNEY, R. K., GRANER, A. & SORRELLS, M. E. 2005. Genomics-assisted breeding for crop improvement. *Trends in Plant Science,* 10, 621-630.

VARSHNEY, R. K., MOHAN, S. M., GAUR, P. M., GANGARAO, N. V., PANDEY, M. K., BOHRA, A., SAWARGAONKAR, S. L., CHITIKINENI, A., KIMURTO, P. K., JANILA, P., SAXENA, K. B., FIKRE, A., SHARMA, M., RATHORE, A., PRATAP, A., TRIPATHI, S., DATTA, S., CHATURVEDI, S. K., MALLIKARJUNA, N., ANURADHA, G., BABBAR, A., CHOUDHARY, A. K., MHASE, M. B., BHARADWAJ, C., MANNUR, D. M., HARER, P. N., GUO, B., LIANG, X., NADARAJAN, N. & GOWDA, C. L. 2013. Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances,* 31, 1120-1134.

VERHOEF, N., YOKOTA, T., SHIBATA, K., DE BOER, G. J., GERATS, T., VANDENBUSSCHE, M., KOES, R. & SOUER, E. 2013. Brassinosteroid biosynthesis and signalling in *Petunia hybrida*. *Journal of Experimental Botany,* 64, 2435-2448.

VIAENE, T., VEKEMANS, D., IRISH, V. F., GEERAERTS, A., HUYSMANS, S., JANSSENS, S., SMETS, E. & GEUTEN, K. 2009. Pistillata-duplications as a mode for floral diversification in (basal) asterids. *Molecular Biology and Evolution,* 26, 2627-2645.

VOLLBRECHT, E., VEIT, B., SINHA, N. & HAKE, S. 1991. The developmental gene *KNOTTED-1* is a member of a maize homeobox gene family. *Nature*, 350, 241-243.

WALLBANK, R. W. R., BAXTER, S. W., PARDO-DIAZ, C., HANLY, J. J., MARTIN, S. H., MALLET, J., DASMAHAPATRA, K. K., SALAZAR, C., JORON, M., NADEAU, N., MCMILLAN, W. O. & JIGGINS, C. D. 2016. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biology,* 14, e1002353.

WANG, J., WURM, Y., NIPITWATTANAPHON, M., RIBA-GROGNUZ, O., HUANG, Y. C., SHOEMAKER, D. & KELLER, L. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature,* 493, 664-668.

WANG, S., LORENZEN, M. D., BEEMAN, R. W. & BROWN, S. J. 2008. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biology,* 9, 1-14.

WANG, X. J., REYES, J. L., CHUA, N. H. & GAASTERLAND, T. 2004. Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biology,* 5, R65.

WANG, Z., HOBSON, N., GALINDO, L., ZHU, S., SHI, D., MCDILL, J., YANG, L., HAWKINS, S., NEUTELINGS, G., DATLA, R., LAMBERT, G., GALBRAITH, D. W., GRASSA, C. J., GERALDES, A., CRONK, Q. C., CULLIS, C., DASH, P. K., KUMAR, P. A., CLOUTIER, S., SHARPE, A. G., WONG, G. K. S., WANG, J. & DEYHOLOS, M. K. 2012. The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *The Plant Journal,* 72, 461-473.

WANI, S. H., KUMAR, V., SHRIRAM, V. & SAH, S. K. 2016. Phytohormones and their metabolic engineering for abiotic stress tolerance in crop plants. *The Crop Journal,* 4, 162-176.

WARR, A., ROBERT, C., HUME, D., ARCHIBALD, A., DEEB, N. & WATSON, M. 2015. Exome sequencing: current and future perspectives. *G3: Genes|Genomes|Genetics,* 5, 1543-1550.

WATERS, P. D., WALLIS, M. C. & GRAVES, J. A. M. 2007. Mammalian sex — Origin and evolution of the Y chromosome and SRY. *Seminars in Cell & Developmental Biology,* 18, 389-400.

WEBB, C. & LLOYD, D. G. 1986. The avoidance of interference between the presentation of pollen and stigmas in angiosperms II. Herkogamy. *New Zealand journal of botany,* 24, 163-178.

WEBSTER, M., A. 2005. Floral morphogenesis in *Primula*: Inheritance of mutant phenotypes, heteromorphy, and linkage analysis. *PhD thesis, University of Leeds*.

WEBSTER, M. A. & GILMARTIN, P. M. 2003. A comparison of early floral ontogeny in wild-type and floral homeotic mutant phenotypes of *Primula*. *Planta,* 216, 903-917.

WEBSTER, M. A. & GILMARTIN, P. M. 2006. Analysis of late stage flower development in *Primula vulgaris* reveals novel differences in cell morphology and temporal aspects of floral heteromorphy. *New Phytologist,* 171, 591-603.

WEBSTER, M. A. & GRANT, C. J. 1990. The inheritance of calyx morph variants in *Primula vulgaris* (Huds). *Heredity,* 64, 121-124.

WEDDERBURN, F. & RICHARDS, A. J. 1990. Variation in within-morph incompatibility inhibition sites in heteromorphic *Primula* L. *New Phytologist,* 116, 149-162.

WEDDERBURN, F. M. & RICHARDS, A. J. 1992. Secondary homostyly in *Primula* L.; evidence for the model of the *S*-supergene. *New Phytologist,* 121, 649-655.

WESTERKAMP, C. & GOTTSBERGER, G. 2000. Diversity pays in crop pollination. *Crop Science,* 40, 1209-1222.

WOODHOUSE, M. R., PEDERSEN, B. & FREELING, M. 2010. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet,* 6, e1000949.

WU, H. J., MA, Y. K., CHEN, T., WANG, M. & WANG, X. J. 2012. PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Research,* 40, W22-W28.

WU, J., XIAO, J., WANG, L., ZHONG, J., YIN, H., WU, S., ZHANG, Z. & YU, J. 2013. Systematic analysis of intron size and abundance parameters in diverse lineages. *Science China Life Sciences,* 56, 968-974.

WU, M. F., TIAN, Q. & REED, J. W. 2006. Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. *Development,* 133, 4211-4218.

WU, T. D. & NACU, S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics,* 26, 873-881.

XIA, X. 1998. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Molecular Biology and Evolution,* 15, 336-344.

XIA, X. 2013. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Molecular Biology and Evolution*, 30, 1720-1728.

XIA, X., XIE, Z., SALEMI, M., CHEN, L. & WANG, Y. 2003. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution,* 26, 1-7.

XIE, Y., WU, G., TANG, J., LUO, R., PATTERSON, J., LIU, S., HUANG, W., HE, G., GU, S., LI, S., ZHOU, X., LAM, T. W., LI, Y., XU, X., WONG, G. K. S. & WANG, J. 2014. SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics,* 30, 1660-1666.

XU, S., OMILIAN, A. R. & CRISTESCU, M. E. 2011. High rate of large-scale hemizygous deletions in asexually propagating *Daphnia*: implications for the evolution of sex. *Molecular Biology and Evolution,* 28, 335-342.

XU, W., LI, F., LING, L. & LIU, A. 2013. Genome-wide survey and expression profiles of the AP2/ERF family in castor bean (*Ricinus communis* L.). *BMC Genomics,* 14, 1-15.

XU, X., PAN, S., CHENG, S., ZHANG, B., MU, D., NI, P., ZHANG, G. 2011. Genome sequence and analysis of the tuber crop potato. *Nature,* 475, 189-195.

XUE, W., LI, J. T., ZHU, Y. P., HOU, G. Y., KONG, X. F. & KUANG, Y. Y. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics,* 14, 604.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution,* 24, 1586-1591.

YANG, Z. & NIELSEN, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution,* 46, 409-418.

YASUI, Y., HIRAKAWA, H., UENO, M., MATSUI, K., KATSUBE-TANAKA, T., YANG, S. J., AII, J., SATO, S. & MORI, M. 2016. Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. *DNA Research*, 25, 215-224.

YASUI, Y., MORI, M., AII, J., ABE, T., MATSUMOTO, D., SATO, S., HAYASHI, Y., OHNISHI, O. & OTA, T. 2012. *S-LOCUS EARLY FLOWERING 3* is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility. *Plos One,* 7, e31264.

YASUI, Y., MORI, M., MATSUMOTO, D., OHNISHI, O., CAMPBELL, C. G. & OTA, T. 2008. Construction of a BAC library for buckwheat genome research - An application to positional cloning of agriculturally valuable traits. *Genes & Genetic Systems,* 83, 393-401.

YASUI, Y., WANG, Y. J., OHNISHI, O. & CAMPBELL, C. G. 2004. Amplified fragment length polymorphism linkage analysis of common buckwheat (*Fagopyrum esculentum*) and its wild self-pollinated relative *Fagopyrum homotropicum*. *Genome,* 47, 345-351.

YEAMAN, S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences,* 110, E1743-E1751.

YEO, P. F. 1975. Some aspects of heterostyly. *New Phytologist,* 75, 147-153.

YOSHIDA, Y., UENO, S., HONJO, M., KITAMOTO, N., NAGAI, M., WASHITANI, I., TSUMURA, Y., YASUI, Y. & OHSAWA, R. 2011. QTL analysis of heterostyly in *Primula sieboldii* and its application for morph identification in wild populations. *Annals of Botany,* 108, 133-142.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research,* 18, 821-829.

ZHANG, S. V., ZHUO, L. & HAHN, M. W. 2016. AGOUTI: improving genome assembly and annotation using transcriptome data. *GigaScience,* 5, 1-12.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology,* 7, 203-214.

ZHAO, X., MO, D., LI, A., GONG, W., XIAO, S., ZHANG, Y., QIN, L., NIU, Y., GUO, Y., LIU, X., CONG, P., HE, Z., WANG, C., LI, J. & CHEN, Y. 2011. Comparative analyses by sequencing of transcriptomes during skeletal muscle development between pig breeds differing in muscle growth rate and fatness. *PLoS One,* 6, e19774.

ZHOU, W., BARRETT, S. C. H., WANG, H. & LI, D. Z. 2015. Reciprocal herkogamy promotes disassortative mating in a distylous species with intramorph compatibility. *New Phytologist,* 206, 1503-1512.

ZHU, J. Y., SAE-SEAW, J. & WANG, Z. Y. 2013. Brassinosteroid signalling. *Development,* 140, 1615-1620.