Structural and molecular analyses of heterostyly in Linum tenue (Linaceae)

FOROOZANI, ALIREZA

Structural and molecular analyses of heterostyly in *Linum tenue* (Linaceae)

***Linum tenue*.** Image by Stuart Brooker and Alireza Foroozani

By Alireza Foroozani

Department of Biosciences

Durham University

December 2018

Submitted for the degree of Doctor of Philosophy

Declaration

The material contained within this thesis has not previously been submitted for a degree at Durham University or any other university. The research reported within this thesis has been conducted by the author unless indicated otherwise.

Alireza Foroozani
December 2018

Acknowledgements

I would first like to deeply thank my primary supervisor, Dr. Adrian Brennan, for his continued guidance, patience and support throughout the entire project. Adrian, there has never been a time where I felt unable to come to you for help. You have given me the opportunity and the freedom for me to take my work in whatever direction I saw fit. I hope that we can remain friends and collaborators into the future.

I would like to give special thanks to my secondary supervisor Prof. Marc Knight and all the members of Lab 19 throughout the years for their friendships and the memories we've shared.

I would like to pay my gratitude to our collaborators: Dr. Juan Orroyo at Universidad de Sevilla, for his guidance and recommendations on the fieldwork; and to Dr. Rocio Perrez-Barrales at the University of Portsmouth, for sharing her data and knowledge of the complexity of *Linum* ecology.

I would also like to thank the members of the Durham Ecology and Durham Molecular Ecology groups, whose friendships have been among the most important of my time in Durham.

Finally, I owe a great personal thanks to Albert Lahat, whose mastery of Python was an invaluable help, and without his teaching and advice I would not have developed a lot of the skills necessary to conduct the work in these pages. David Dolan, for his bioinformatics guidance and for teaching me some of the most important lessons in computing! And last but not least Tom Batstone, for his many late-night discussions on bioinformatics and words of encouragement throughout the years.

Dedication

I dedicate this thesis to my mother, for her love and sacrifices.

General Abstract

Flowering plant mating systems are as varied as they are complex, and are generally considered the leading force behind angiosperm diversity and evolution. Establishing the genetic and molecular mechanisms behind phenotypic traits is important to understand how they have developed, evolved and spread across populations and taxa. Until recently, our understanding of the genetic basis of heterostyly, whereby reciprocal polymorphisms in the relative positioning of stigmas and anthers in certain angiosperm species, has been largely unknown; though has since been shown to be determined by the multiallelic *S*-locus. Using the study species of *Linum tenue* (Linaceae), this thesis investigates the ecology, trait variance and developmental progression of heterostyly in the flowers of *L. tenue*, making inferences on the specific control of heterostyly in this species and general speculative theories for its evolution in this taxa. Using a transcriptomic dataset derived from short-read Illumina sequence data, this thesis presents an automated method for reassembling consensus unigene sequences for the creation of a high-quality refernce transcriptome from mRNA data, providing a useful tool for a challenging aspect of gene expression studies in non-model organisms. This method is exemplified through the creation of a transcriptomic reference sequence of *L. tenue* vegetative and floral tissues, and through the *de novo* assembly of the *Arabidopsis thaliana* transcriptome. Next, global and differential expression analyses are used to investigate patterns in expression between polymorphic *L. tenue* flowers to discover candidate and proto-candidate loci determining the heterostyly syndrome and study differences in expression behaviour between the two morphotypes. This thesis provides evidence that heterostyly in *L. tenue* is pleiotropically controlled, and that the non-recombining nature of the *S*-locus can result in interesting patterns of morph-specific expression.

Table of contents

List of abbreviations and acronyms

ABC	type-A, type-B, and type-C function genes (ABC model of flower development)
ACB	Adrian Christopher Brennan
AF	Alireza Foroozani
APG	angiosperm phylogeny group
BALLISTA	abstraction of allelic and isoform-level variation for transcriptome analyses
BED	browser extensible data (genome annotation file format)
BLAST	basic local alignment search tool
BLASTN	BLAST search using nucleotide query against a nucleotide database
BLASTP	BLAST search using an amino acid (protein) query against an amino acid database
BLASTX	BLAST search using a nucleotide query against an amino acid database
BUSCO	benchmarking universal single-copy orthologs
CAP3	contig assembly program v3
CDD	(NCBI) conserved domain database
CDS	coding sequence
EMBL	European molecular biology laboratory
EST	expressed sequence tag
FDR	false discovery rate
GFF3	gene feature format (version 3)
GLO1,2	*GLOBOSA1, GLOBOSA2* (floral identity genes in *Antirrhinum* and *Primula*)
GO	gene ontology
GTF	gene transfer format
LgAP1	*LINUM GRANDIFLORUM APETALA1*
LgGLX1	*LINUM GRANDIFLORUM GLOXAL OXIDASE 1*
LgMYB21	*LINUM GRANDIFLORUM MYB21*
LS	long-styled (heterostylous morph)
LSU	large sub unit (ribosomal)
M(1-4)	conserved amino acid motif (1-4) on VUP and homologs
NCBI	national centre for biotechnology information
ORF	open reading frame
ORNA	optimized read normalization algorithm [sic]
PCA	principal components analyses
PE	paired-end
RR	reduced redundancy
RPB	Rocío Pérez Barrales
RPS	reversed position specific (BLAST search)

SCF	Skp, Cullin, F-box containing complex
SE	single-end
SH-aLRT	Shimodaira-Hasegawa approximate likelihood-ratio test
SRK	stigma-specific S locus receptor kinase
SS	short-styled (heterostylous morph)
SSG	*SHORT-STYLE-SPECIFIC GENE*
SSU	small sub unit (ribosomal)
STM	scaffolding by translation mapping
TBLASTN	BLAST search using a translated nucleotide query against a translated nucleotide database
TPP1	*THRUM POLLEN PREDOMINANT GENE*
t-SNE	t-distributed stochastic neighbour embedding
TSS1	*THRUM STYLE-SPECIFIC GENE*
UPGMA	unweighted pair group method with arithmetic mean
VUP(1-4)	*VACULAR-RELATED PROTEIN (1-4)*

- Chapter 1 -
General Introduction

Mediterranean Sun – Apiaceae *cf.* Image by Alireza Foroozani

1.0 General Introduction

"Equipped with flowers, the new seed plants spread from the forests to the polar tundra and invaded freshwaters; they returned even to the sea, which no moss, liverwort, fern, fern ally, cycad, or conifer had been able to do. Emancipated, they transformed the vast gloomy flowerless world with gaiety, feast, and song into a new factory of life."

E. J. H. Corner (1964) *The Life of Plants*

1.1 Introduction to angiosperm reproductive biology

1.1.1 Angiosperm diversity

Understanding the origins of phylogenetic diversity and ecological success are long-standing challenges in biology, yet precisely defining either is difficult as both are fairly ambiguous and multi-factorial. However, by any standard, both may be used to aptly describe the flowering plants (angiosperms). Composed of an estimated 352,000 species, angiosperms comprise around 90% of all extant species of land plants (embryophytes) (Niklas, 1997); they represent every known plant body plan and growth form, from small herbaceous dandelions to towering oaks; and they have radiated into every terrestrial biome, where they play key roles in ecological food webs and species-species interactions. The dominance and diversity of angiosperms is nowhere more lucidly described than in Edred J. H. Corner's '*The life of plants*' (1964), which, forgiving the misconception at its time of writing that embryophytes descended from marine ancestors, provides the aspiring botanist with a comprehensive overview of angiosperm biology entwined with deep passion and profound understanding. Since their emergence in the fossil record during the early Cretaceous only c.120 mya (Hughes, 1994; Sun *et al.*, 2002), angiosperms appear to have undergone rapid early diversification (Friis, Pedersen and Crane, 2005); with the origin of key clades, such as the eudicots (130 mya) rosids (108-121 mya) and asterids (101-119 mya) dating to roughly the same period (Bell, Soltis and Soltis, 2010).

The evolutionary forces that have begot and shaped the staggering angiosperm diversity we see were branded by Darwin as an *"abominable mystery"*, and still remain elusive today.

Up until twenty years ago there was no solid consensus in the scientific community regarding the relationships between angiosperm groups (Heywood, 1978), and the application of predominantly morphology-based classification did little to bridge deep-rooted schisms among botanists. Due to the nature of the characters used for classification, the inability to distinguish between homology and convergence (and, indeed, the weighting of their importance) prevented cladistic analyses from creating accurate phylogenies. Subsequent advances in palynology (Donoghue and Doyle, 1989), molecular systematics (Crane, 1993; Qiu *et al.*, 1999; Soltis, Soltis and Chase, 1999) and embryology (Williams and Friedman, 2002; Friedman, 2006) have provided major insights into the evolutionary relationships among angiosperm taxa. Allowing clear and objective conclusions to be drawn on the phylogenetic distribution of defined angiosperm characters paved the way to a complete revaluation of the angiosperm phylogeny. The work presented by these, and many other, scientists has been summarised by the Angiosperm Phylogeny Group (APG), an (informal) international collaboration of plant systematists. The latest findings of the APG IV (Chase *et al.*, 2016) are regarded as the most reliable reference for and holistic depiction of angiosperm phylogenetic relationships to date.

1.1.2 Angiosperm mating systems

The vast diversity seen in angiosperms is mirrored in the complexity of their reproductive biology. Determining the reproductive fate of the plant, there is often very strong selection acting on flowers. By shifting to a new pollinator, or becoming intimately specialised with

one, plant populations can achieve rapid reproductive isolation. It is not surprising, then, to see that animal-pollinated lineages show the highest levels of species richness and can frequently undergo rapid adaptive radiations (Johnson and Steiner, 2000, 2003; Richardson *et al.*, 2001). I believe this can be seen as akin to sexual selection: through reward, scent, shape and colour the flower has evolved increasing ornamentation to attract the pollinator, which carries with it the unmindful pollen grain. Viewed in this way, the flower serves a function similar to that of the peacock's tail. It explains how the evolution of floral traits can be exacerbated, and how rapid speciation can take place among animal-pollinated taxa.

1.1.3 The dimensions of plant mating strategies

It is first important to understand how trajectories of male and female function can either align or diverge based on the associated cost of the mating system. Parker, Baker and Smith (1972) showed how an evolutionary stable strategy favours gamete dimorphism (anisogamy) under conditions where parental resources are limited, as is always the case in nature; resulting in small, abundant, motile male gametes, and large, few, sessile female gametes. The natural corollary of this is that male fitness is determined by the *number* of gametes produced, and female fitness by the *quality* of gametes produced. In many mammals, for example, there is heavy investment in sexual form and function, spurred often by sexual selection: males compete to defend territories or harems, while females invest heavily through maternal care and weaning. In most plants, however, sexual costs account for relatively low amounts of the total energy budget (Richards, 1997), and thus male and female forms can successfully coexist in a hermaphroditic individual. It is not surprising, then, to see the evolution of dioecy in situations of differential sexual investment, i.e. expensive fruit production in tropical trees (Barrett and Hough, 2013).

Roughly 90% of angiosperms are cosexual and hermaphroditic, giving weight to the theory that the first flowers were cosexual (Frohlich and Parker, 2000). Assuming an absence of other constraints, cosexual flowers can self-pollinate, which guarantees reproduction. Selfing is a sexual process, involving meiosis to form both gametes, and so does not result in the same lack of genetic variability that is found in asexual reproduction. This is quite possibly what allowed angiosperms, which first evolved on the peripheries of plant communities (Sun *et al.*, 2002), to invade and conquer the gymnosperm forests.

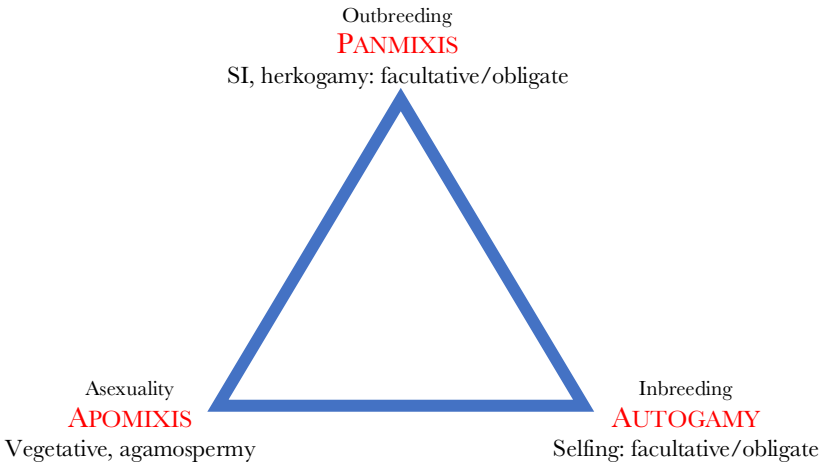


Figure 1.1 The 'eternal triangle' of angiosperm breeding system interfaces (Richards, 1997). There are three distinct breeding systems, but it is rare to find such extremes in nature. Most species will fall further along one of the edges; even in populations considered obligate inbreeders, such as *Arabidopsis thaliana*, there is still a limited degree of outcrossing that will occur. In this way, many species adopt a 'mixed' mating strategy.

As shown in Figure 1.1, angiosperm mating strategies tend to lean towards one of three broad types. Apomixis will not be discussed in detail here, but it is interesting to note how this mating strategy has led to the evolution of complex communities and endemic taxa, notably in *Sorbus* populations of the Avon Gorge, Bristol (Robertson *et al.*, 2010), and the East Lyn Valley, Devon (Hamston *et al.*, 2017). The third strategy is outcrossing between different individuals in a large enough population (panmixis).

Each strategy has consequences for gene flow dynamics among populations. Outcrossers will have higher levels of genetic diversity, with large numbers of heterozygotes for any

polymorphic locus. This genetic variability may offer the potential for continual evolution and allow diversification to fill many different niches. Selfing populations will have lower levels of genetic diversity and fewer heterozygotes, although occasional outcrossing may slightly ameliorate this. As a result, most selfing species may lack adaptive or evolutionary potential (Goldberg *et al.*, 2010). Asexual populations will also possess a further lack of genetic variation, unless, as has been seen in *Sorbus*, a number of different asexual lines coexist. Depending on the progenitor genotype, heterozygosity may be low or high, but, given their vegetative nature, they (theoretically) have almost no evolutionary potential.

The balance between the relative importance of genetic variation and reproductive assurance differs between species, primarily due to habitat, life history and niches occupied (Charlesworth, 2006). It is the conflict between long-term evolutionary persistence and short-sighted evolutionary change that characterises the distribution of mating strategies.

1.1.4 Promoting outbreeding

Usually bearing hermaphroditic flowers, most plants face the possibility of self-pollination. This can involve either pollen from an anther landing on the stigma of the *same* flower (autonogamy) or pollen from an anther landing on the stigma of a *different* flower of the same individual (geitonogamy). It has long been known that inbred offspring are less fit than outbred offspring, as a result of *i)* accumulation of deleterious recessive alleles and *ii)* a lack of genetic variation to enhance adaptive potential; this is known as inbreeding depression (Charlesworth and Charlesworth, 1987). To overcome this, the evolution of sophisticated mating systems conducive to outcrossing have evolved multiple times across

the angiosperm phylogeny; some of which seem to have evolved in parallel, others convergently.

Two of the most common forms are self-incompatibility (SI) (the *recognition and rejection* of self-pollen) and herkogamy (the *separation* of male and female sexual organs in space). SI involves a series of molecular mechanisms, whereby pollen from the same, or a closely related, individual is recognised on the stigma and either fails to germinate or the pollen tube is actively destroyed. In contrast, herkogamy is purely structural and often takes the form of heterostyly, which involves polymorphisms in relative anther and stigma heights, where morphotypes generally can only mate reciprocally. SI and heterostyly are under tight genetic control, and different species may employ either or both of these strategies to facilitate outcrossing. Other systems also exist, such as dichogamy (the separation of male and female organs in time) and dicliny (the development of unisexual flowers, on either monoecious or dioecious individuals), but will not be discussed further here.

By Fisher's fundamental theorem, where the rate of evolution increases with additive genetic variance, high levels of genetic diversity make it easier for a population to traverse the adaptive landscape. The proclivity to evolve such diverse ranges in reproductive strategy, and to enforce outcrossing in particular, is a key attribute of angiosperms, and has undoubtedly contributed to their diversity, plasticity and success.

1.1.4.1 Self-incompatibility

It has been reported that up to 50% of angiosperm species bear some form of SI (McClure and Franklin-Tong, 2006). In the 90 angiosperm families where SI has been studied, the control of fertilisation can often be determined by a single genetic locus, the S locus. The S locus is a closely linked chromosomal region containing multiple genes, which together

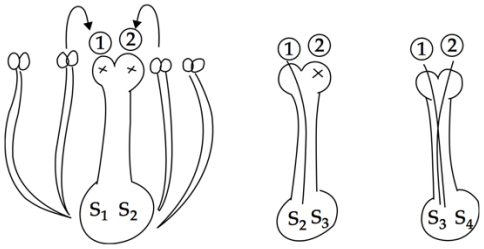


Figure 1.2 Gametophytic self-incompatibility. Mating type of the pollen is determined by the haplotype of the pollen gamete (Glover, 2007).

control SI, and exists in multiple haplotypes (although sometimes referred to as 'alleles' in the literature). A pollen-pistil union where one, or both, of the parent haplotypes are the same results in an incompatible pollination. SI can be split into two main categories: *i)* gametophytic SI (GSI), where SI is determined by the haplotype of the pollen (see Figure 1.2), and *ii)* sporophytic SI (SSI) where SI is determined by the diploid genomes of the parents (Figure 1.3).

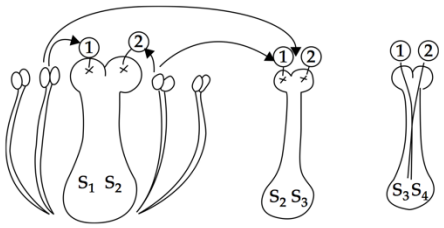


Figure 1.3 Sporophytic self-incompatibility. Here, the mating type of the pollen is determined by the diploid genotype of the parents (Glover, 2007).

Best characterised in the Solanaceae, but also found in the Rosaceae and Papaveraceae (among others), GSI inhibits pollen tube growth in the *tissue of the style*. The female determinant was first discovered in *Nicotiana alata* (Anderson *et al.*, 1986), and later confirmed to be acting as an RNase (named S-RNase) (Gray *et al.*, 1991). Characterising the male determinant of GSI has been more difficult. Early work found *S locus F box* (*SLF*) genes to affect SI (Lai *et al.*, 2002; Qiao *et al.*, 2004; Sijacic *et al.*, 2004); F box proteins target specific proteins for ubiquitination by the SCF E3 ubiquitin ligase complex; the proposition being that non-self S-RNases are targeted for degradation. A more in-depth study has since shown that multiple *SLF* alleles are linked to the S locus and expressed in

the pollen, allowing a variety of S-RNases to be targeted (Kubo *et al.*, 2010). Given the relatively recent discovery of the male determinant, much work remains to be done to fully characterise the system and that of completely different mechanisms described in other families (Foote *et al.*, 1994; Wheeler *et al.*, 2009).

SSI has been heavily studied in the Brassicaceae, but is also found in other major families, including the Asteraceae and Caryophyllaceae, although the molecular mechanisms by which it governs in them are yet unclear. In contrast to GSI, SSI inhibits pollen tube germination on the *stigmatic surface* before it enters the style. Dominance, co-dominance and recessiveness may be at play among S locus haplotypes, making understanding the relationships between them a rather complicated affair (Hiscock and McInnis, 2003). Two highly polymorphic S locus genes have been identified in *Brassica*: *S locus receptor kinase* (*SLK*) (Stein *et al.*, 1991) and *S locus glycoprotein* (*SLG*) (Nasrallah *et al.*, 1985; Kandasamy *et al.*, 1989). SRK proteins have been found to be localised to the plasma membrane of stigmatic tissues (Stein *et al.*, 1996) and are the female specificity determinants, with SLG enhancing the recognition process (Takasaki *et al.*, 2000). The male determinant has been found to be *S locus cystine-rich* (*SCR*) (Schopfer, Nasrallah and Nasrallah, 1999), also known as *S-locus pollen protein 11* (*SP11*) (Takayama *et al.*, 2000), the proteins of which were found to be expressed within the tapetum and rubbed into the exine walls of developing pollen grains. When pollen lands on the stigma, SCR and SRK will bind together if they are encoded by genes from the same haplotype (Takayama *et al.*, 2001) to initiate an intracellular signal transduction cascade, which results in the degradation of proteins essential for pollen tube growth by E3 ubiquitin ligase (Stone *et al.*, 2003).

Studies in *Senecio* (Asteraceae) have focused on isolating glycoproteins with stigma-specific expression and S locus segregation through transcriptome-based approaches (Allen *et al.*, 2011). Interestingly, potential homologs of *SLK* and *SLG* are ubiquitously expressed and

do not segregate with the S locus (Hiscock *et al.*, 2003), suggesting SSI has evolved multiple times across the angiosperm phylogeny with different molecular mechanisms.

1.1.4.2 Heteromorphic self-incompatibility

Many angiosperm families have been reported to have a heteromorphic SI system, where SI type is associated with morphological polymorphisms in the relative positioning of the stamen and the stigma. This is often also referred to as heterostyly and will be discussed in greater detail in section 1.2.

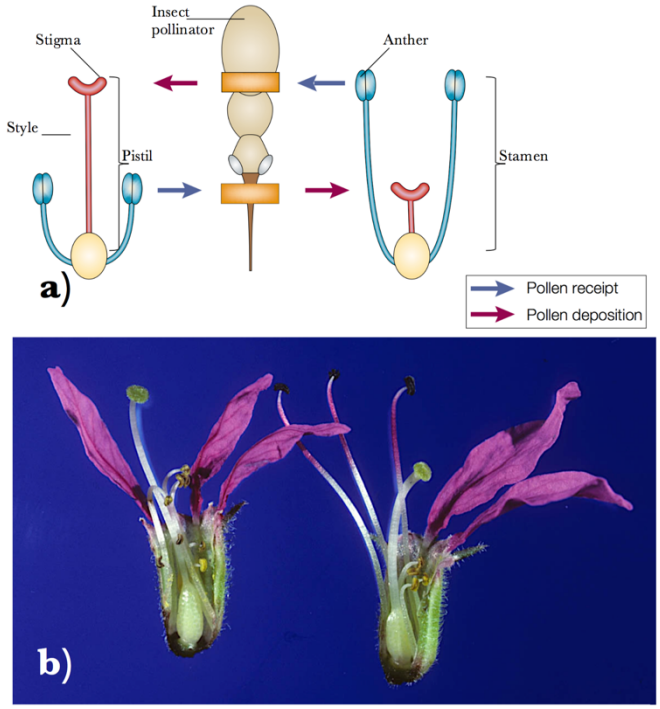


Figure 1.4 Heterostyly, a form of reciprocal herkogamy, in angiosperms. **a)** A schematic representation of distyly. Pin (left) and thrum (right) polymorphisms provide efficient pollen transfer with a reduced chance of self-pollination (Barrett, 2002). **b)** Photograph showing heterostyly in long (left) and medium (right) morphs of tristylous *Lythrum slaicaria* (Lythraceae) (Nickrent *et al.*, 2006 onwards)

Usually heterostyly takes the form of distyly, where two morphs are present in a species, although tristyly (three morphs: *e.g. Oxalis eckloniana*, *Lythrum slaicaria*) is also known to occur. As shown in Figure 1.4a, distyly involves 'pin' (long-styled) and 'thrum' (short-styled) individuals in a population, and only *reciprocal pollination* results in a legitimate union. The classic example is that of the common primrose, *Primula vulgaris* (Primulaceae),

which Charles Darwin studied in great detail. He postulated that the function of heterostyly was to cover the anterior and posterior ends of the insect with pin and thrum pollen respectively, so entry into a flower would be likely to deposit pollen onto the reciprocal stigma. *P. vulgaris* has a SSI system directed by only two haplotypes, which also control the relative style and anther lengths. These haplotypes display Mendelian dominance-recessive behaviour, with SS and Ss individuals producing thrum flowers and ss individuals producing pin flowers. This works in a very similar way to the XY sex determination system in mammals, and similarly produces populations with 50-50 proportions of each morph.

There is some evidence that molecular SI mechanisms may differ between floral morphs, as incompatible pollen has shown to be inhibited in different tissues dependant on whether the maternal plant is a pin (stigmatic tissues) or thum (style tissues) in *Pentanisia prunelloides* (Rubiaceae) (Massinga, Johnson and Harder, 2005). Until recent years, most research has focused on the population dynamics and evolution of heterostyly, as opposed to its molecular evolution and genetic basis.

1.1.5 Evolutionary transitions to self-compatibility

A further striking feature of plant mating systems is their evolutionary lability. Outcrossing is dependent on the pollen of a *compatible* mate to *reach* the stigma, which will fail in the absence of a large enough population and/or an effective pollinator. It is not surprising, then, to see that about 20% of angiosperm species are primarily selfers, with transitions from SI to self-compatibility (SC) and from heterostyly to homostyly being commonplace (Barrett, 2002).

One fascinating example of this is in the bee orchid *Ophrys apifera*. *Ophrys spp.* are well known for their visual and olfactory mimicry of female insects – usually bees – which in turn attracts males of the same species to attempt copulation with the flower, transferring pollen in the process. Given the complexity required for such sexual deception, the plant-pollinator interaction is highly species-specific and has developed through tight coevolution. However, due to the extinction of the pollinator, the stalks bearing the pollen sacs have adapted their development to grow longer and thinner, allowing the pollen sacs to draw themselves down under their own weight and contact the stigma (Stebbins, 1957). The stark irony here demonstrates both the blind directionality of natural selection and the plasticity of angiosperm reproductive systems: millions of years of evolution geared towards heavy investment in traits that promote outcrossing have culminated in a plant that selfs.

The genetic basis for the break-down of SI has been characterised in *Arabidopsis thaliana*, and attributed to an inversion within the *SCR* locus (Tsuchimatsu *et al.*, 2010). The shift to selfing is often accompanied by the loss of traits that attract pollinators, as the plant no longer has to invest in these (sometimes) expensive characters. This can include a reduction in flower size, a reduction in scent or nectar production, and a loss of any stylar polymorphisms. Traits that enhance self-pollination may also become more pronounced: flowers may become increasingly cleistogamous (indehiscent) or lose dichogamous maturation (Kalisz *et al.*, 2012). Any alleles that promote selfing may spread quickly through a population due to natural selection favouring the reproductive assurance which autogamy provides. In some cases the free-living gametophyte phase of the plant life-cycle is can purge deleterious alleles from a population (Szövényi *et al.*, 2014), although, as explained in section 1.1.3, inbreeding can perpetuate genetic invariability and thus a lack of evolutionary vigour.

The evolution of SC from SI is largely unidirectional, and reversions from selfing back to outcrossing are rare; Dollo's law postulates this is due to the unlikelihood of such complex traits being regained (Barrett, 2013). A study by Goldberg *et al.* (2010) investigated this directionality in the evolution of selfing, and found the rates of transitions from SI to SC were so much greater than SC to SI that it was in fact surprising that SI persists at all. Upon taking a macroevolutionary perspective, it can be seen that selection acts antagonistically on different levels of the taxonomic hierarchy. Species may also be units of selection: they produce 'offspring' (incipient species); there will be variation in the offspring, but traits will also be inherited from the progenitor species; species may have differential success (give rise to more/less incipients); and they can also 'die' (become extinct). Goldberg *et al.* found that SI species have a tendency to beget more incipient species and were much less likely to become extinct than SC species. The short-sightedness of natural selection driving the evolution of SC is, thereby, completely displaced by species selection favouring SI in the long-term.

1.2 Introduction to heterostyly

"I do not think anything in my scientific life has given me so much satisfaction as making out the meaning and structure of these [heterostylous] plants."

Charles Darwin (1876)

1.2.1 Reciprocal herkogamy: the Darwinian hypothesis

From at least the 16th century people had already made observations of the arrangements of the sexual organs of primrose (*Primula*) species, but botanical interest in heterostyly was piqued when Charles Darwin published a paper entitled *'On the two forms, or dimorphic condition, in the species of Primula, and on their remarkable sexual relations'* in 1862 – three years after the publication of *'On the origin of species'*. Darwin claimed to have first noticed this unusual trait over twenty years prior in *Linum flavum*, where the species is composed of pin

(long-styled, LS) and thrum (short-styled, SS) individuals, but originally believed it to be a function of the natural variation present in a population. However, upon an examination of *Primula spp.*, he discovered that the two forms were *"much too regular and constant"* to be present as a result of random variability.

His original assumption was that heterostyly represented a transitional stage between cosexual flowers and dioecy; that the shorter organs on each morph were decreasing in functionality. At his home at Down House, Orpington UK, Darwin began to test this hypothesis, growing up hundreds of *Primula* flowers in flower beds at the bottom of his garden. Through a series of controlled crosses, whereby there are four possible pollen-pistil unions (LS pollen *x* LS stigma; LS pollen *x* SS stigma; SS pollen *x* LS stigma; SS pollen *x* SS stigma), it became clear that there was no fitness cost associated with a shorter organ or an intermorph mating; which would be a prerequisite for diversifying selection to drive segregation of the sex organs. In fact, the pollination studies revealed the opposite: as demonstrated in Figure 1.5, intermorph unions where organs were of reciprocal heights (i.e. LS pollen *x* SS stigma or SS pollen *x* LS stigma) produced dramatically larger and higher quality seed sets (had greater reproductive success) than those from intramorph matings or self-pollinations.

Darwin's work clearly indicated the function of heterostyly was to encourage outcrossing, and in 1877, after further extensive examination of heterostyly in many other families such as the Linaceae, Rubiaceae, and Polygonaceae (buckwheat) (Darwin, 1877), concluded that the mechanism behind this is to encourage deposition of pollen from different style morphs on different ends of the insect pollinator. A year prior, Darwin had published a book entitled *'The effects of cross and self-fertilisation in the vegetable kingdom'* (1876) and was well aware of the effects of inbreeding. For this reason, crosses were always between unrelated individuals to prevent confounding influences on seed set measures. In

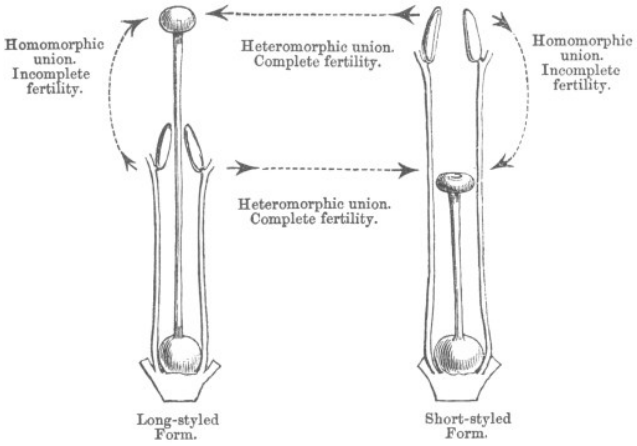


Figure 1.5 An illustration of Darwin's model of heterostyly from his 1862 paper *'On the two forms of Primula'*. The dotted arrows represent various possibilities of pollen transfer, where intermorph (heteromorphic) pollinations result in a much higher seed set than intramorph (homomorphic) pollinations. This suggests selection is favouring intermorph (disassortative) mating.

addition to the lengths of the sexual organs, a number of other ancillary morphological variations were found specific to each morphotype. Specifically the large size difference between pin and thrum pollen in heterostylous *Primula spp.*, along with the different [inwards/outwards-facing] orientations of the stigmas, convinced Darwin that dimorphic pollen is distributed along the insect in a segregated fashion: LS pollen on the anterior and SS pollen on the posterior (see Figure 1.4). This then serves, Darwin postulated, to ensure that when the pollinator visits a pin flower, *long-anther pollen* on its rear will be in contact with the *long-styled stigmas*, and *short-anther pollen* on its front will be in contact with *short-styled stigmas* in a thrum.

Darwin thus described an elegant structural mechanism employed by flowering plants to enforce outcrossing and disassortative mating in order to prevent inbreeding depression.

1.2.2 The ecological function: the modern perspective

The ecological function of heterostyly can be difficult to examine, based on the limited number of methods available for precise pollen tracking. Ganders (1974) argued avoidance of self-pollination cannot be the only explanation for the persistence of

heterostyly, as monomorphic populations can equally prevent self-pollination just as effectively. This work also showed, through emasculation of anthers, that a significant percentage of incoming pollen on the distylous stigmas of *Jepsonia heterandra* (Saxifragaceae) was from the opposite morph. Evidence seems to suggest heterostyly functions to promote disassortative mating and not just the avoidance of selfing.

Lloyd and Yates (1982) and Harder and Barrett (2006) studied the effects of self-interference, which occurs when there is competition between the male and female sexual functions within a hermaphroditic individual. In the case of the SI flower, if autonogamy (self-pollen landing on stigma of the *same flower*) occurs, male function is adversely affected as the pollen is 'wasted', and there may also be a cost to female function: there will be less space on the stigma for viable pollen grains. Conversely, male fitness (in terms of the export of pollen) can be compromised if the positioning of the stigma reduces contact between the anther and the visiting pollinator. In this way, the floral architecture that maximises male fitness can be different to the architecture that maximises female fitness (Johnston *et al.*, 2009). Other than opting for mono- or dioecy, herkogamy provides the plant with a resolution for this sexual conflict (Barrett, 2002). Heterostyly therefore functions to *(i)* reduce sexual interference through the spatial separation of sexual organs within the flower, *(ii)* promote disassortative mating through selective pollen transfer using the reciprocal positioning of the sexual organs, and *(iii)* prevent selfing through structural and physiological intramorph incompatibilities, aided by biochemical SI mechanisms.

1.2.3 The genetic basis (in *Primula* – the 'model' heterostylous system)

In the early 20th century, William Bateson 'rediscovered' the work of Gregor Mendel and fully recognised its importance for the understanding of inheritance. Being a zealous

exponent of Mendel's work, Bateson quickly became known as 'Mendel's bulldog', and coined the term 'genetics' in 1905 (from the Greek 'genno': *'to give birth'*); and founded what has today become the scientific discipline of genetics. In the same year, Bateson published *'On the inheritance of heterostylism in Primula'* (Bateson and Gregory, 1905), which demonstrated that the inheritance of distyly in *P. sinensis* nicely follows the principles of Mendelian particulate inheritance. Through the cross- and self-fertilisation of pin and thrum individuals, Bateson determined that (long-styled, LS) pins were homozygous *ss* and (short-styled, SS) thrums were heterozygous *Ss*, where *'S'* represents the allelic locus controlling heterostyly and *S* is the dominant allele and *s* is the recessive. The results showed that self-fertilised pins solely produced LS offspring (homozygosity), whereas self-fertilised thrums produced SS to LS offspring in the expected ratio of 3:1 (heterozygosity); and a pin *x* thrum cross yielded SS to LS offspring in a 1:1 ratio (*S* is dominant over *s*). Bateson's work was replicated and confirmed in various different systems towards the latter half of the 20th century. However, higher levels of intramorph incompatibility sometimes prevented successful crossing experiments from being carried out, such as in *P. vulgaris* (Ornduff, 1992).

The first large breakthrough in heterostyly genetics came from the work of Alfred Ernst, whose exhaustive crossing experiments (Ernst, 1928, 1936, 1955), involving thousands of *Primula* intra- and inter-species pollinations, revealed that three determinant loci were together responsible for the two style morphotypes: *G*: controlling style length, stigmatic papillae type and female compatibility type (with the dominant allele causing short styles); *P*: the pollen and male incompatibility type (with the dominant allele causing larger pollen; and *A*: the position of the anthers (with the dominant allele causing long anthers). This complex of loci has come to be collectively known as the *S* locus. Under Ernst's model, thrums were heterozygous *GPAgpa* and pins were homozygous *gpagpa*, as suggested by Bateson and Gregory (1905). This discovery was primarily due to the categorisation of

novel morphotypes in rare homostylous populations (high anthers in LS homostyles and low anthers in SS homostyles), in addition to similar segregation of pollen size, suggesting separate, yet tightly linked, alleles were responsible for different aspects of the heterostyly syndrome. Specifically, the pollen incompatibility phenotype consistently segregated with pollen size, and the stigma incompatibility type with style length suggesting the GPA order of genes within the *S* locus.

Ernst's work suggested that inheritance of heterostyly was clearly controlled by a number of tightly-linked alleles, most likely a chromosomal region, and that novel morphotypes were arising as a result of mutation in one or more of the constituent alleles, creating new *S* locus genotypes. Further work focussing on the classical genetics of *Primula* in the latter half of the 20th century confirmed and expanded on Ernst's work, and showed that while mutation may be the cause of novel morphotypes in some *Primula* species, recombination events at the *S* locus was more likely to be producing the unusual morphs (Pamela and Dowrick, 1956; Barrett, 1992). Pamela & Dowrick (1956) also worked with the tetraploid *Primula obconica* and noted that the frequency of double reduction, a phenomenon where the complicated pairing of homologous chromosomes during meiosis in polyploid organisms can result in a chromatid and its sister copy to end up in the same gamete (potentially allowing heterozygous individuals to generate homozygous offspring), was extremely low. Given that the centromere of the chromosome can shield against crossing over, the coefficient of double reduction can therefore be a function of distance from the centromere. Pamela & Dowrick (1956) thus predicted that the *S* locus was likely to be situated close to a centromere.

The consensus, then, for the model of heterostyly genetics in *Primula* was that the *S* locus was a diallelic superlocus, with all dominant alleles (responsible for SS thrums), *GPA*, present on one haplotype and all recessive alleles (responsible for LS pins), *gpa*, present on

the other. Recombination at the locus is heavily suppressed, but occasional recombination events can create unusual morphs. It has been suggested (Kurian and Richards, 1997) that additional genes are responsible for sex-linked incompatibility other than *G* and *P* as an unusual morphotype was discovered that produced a mixture of pin and thrum pollen, demonstrating that male compatibility type can be segregated from pollen size.

Population genetic studies predict that negative frequency dependant selection should maintain a balanced level of polymorphism within a population for longer than alleles at other loci not under selection (Charlesworth, 2006). However, the effects of genetic drift should also be stronger on the *S* locus than selection compared to other regions of the genome due to the reduced effective population size of the S allele, and purifying selection to remove deleterious mutations will be weaker at the S allele, as it is only present as a heterozygous genotype in a genomic region that is already under tight suppression of recombination (Uyenoyama, 2004). The corollary of this is that *S* locus genes should present higher levels of sequence divergence relative to other parts of the genome, and will likely have accumulated associated genetic load and regions of repetitive elements and transposons (Kappel, Huu and Lenhard, 2017).

With improvements in molecular techniques and technologies, attention in *Primula* has turned towards sequence and functional genetics-based approaches. Through subtractive transcriptomics (McCubbin, Lee and Hetrick, 2006) and differential display PCR (Li *et al.*, 2007) and RFLP (Manfield *et al.*, 2005) approaches, numerous genes were initially found to be present with morph-specific expression patterns. Some of these were not linked to the *S* locus (McCubbin, Lee and Hetrick, 2006) and thus thought to function downstream. However, *SLL1* and *SLL2* were found to be tightly linked to the *S* locus (Li *et al.*, 2007), and *PvSLP1* was found to be very closely linked repetitive region (Manfield *et al.*, 2005). Overexpression of *GLO1*, the *Primula* ortholog of the class B floral homeotic

gene *GLOBOSA* in *A. thaliana*, was found to be responsible for homeotic mutant phenotypes (Li *et al.*, 2008, 2010). Li *et al.* (2015) then used these loci as part of a high-resolution map of the *S* locus in *P. vulgaris*. It was also confirmed that the *S* locus is located on the largest *Primula* chromosome next to a centromere, as predicted by Pamela & Dowrick (1956). Genome, RAD and transcriptome sequencing of *P. veris* (Nowak *et al.*, 2015) suggest heavy suppression of recombination in genomic regions surrounding the *S* locus and higher levels of sequence divergence were found in thrum morph haplotypes, as predicted by population genetics (Charlesworth, 2006). Of key significance in this study was the absence of these *S* locus sequences in the pin genome, suggestive of a hemizygous determination system. Recently, a build of the *P. vulgaris* genome (Cocker *et al.*, 2018) has revealed the architecture of the *S* locus, and has shown its conservation throughout the genus.

1.2.4 Genetic studies in other heterostylous systems

In most other studied heterostylous systems, SS morphs similarly appear to be determined by a dominant (or possibly hemizygous) haplotype. Classical (Shore and Barrett, 1985) and molecular (Labonne *et al.*, 2010) genetics strongly suggest the presence of a superlocus in *Turnera* (Passifloraceae) with thrum morphs being determined by the dominant haplotype. Thrum style-specific expression of alpha-dioxygenase (unlinked to the *S* locus) and polygalacturonase (linked to but not a part of the *S* locus) (Athanasiou and Shore, 1997; Athanasiou *et al.*, 2003) suggest secondary control by distyly-determinant factors. Labonne et al. (2009) conducted high-resolution mapping of the *S* locus and found genes (including retrotransposons, as predicted by population genetics) to be tightly linked with the *S* locus.

Inheritance patterns in *Fagopyrum* (Polygonaceae) similarly indicate governance by a superlocus (Garber & Quisenberry 1927; Matsui et al. 2003). Many markers have been found to be linked to the *S* locus (Aii *et al.*, 1998; Yasui *et al.*, 2004) along with thrum-specific expression (Miljuš-Đukić *et al.*, 2004). RNAseq was used to identify four *SHORT-STYLE-SPECIFIC GENE* (*SSG*) genes (*SSG1-4*) with thrum-specific expression (Yasui *et al.*, 2012). *SSG3* seems to have arisen through the duplication of a homolog of the closely related *EFL3* in *A. thaliana*. This pattern is also observed in other *Fagopyrum* species, and thus has implications for the selection of heterostyly in the Polygonaceae. Furthermore, mutations in *SSG3* appear to be involved in the breakdown of heterostyly and SI in two independent *Fagopyrum esculentum* homostyles, strongly suggesting it is a functional member of the *S* locus. Genome sequencing of *F. esculentum* (Yasui *et al.*, 2016) further identified large regions of over 5.4 Mb that were SS-specific, suggestive of a large non-recombining hemizygous region, with 75% of the sequence derived from transposable elements, similarly consistent with population genetics predictions.

Within the Boraginaceae heterostyly appears to have evolved independently on numerous occasions, at both the family and genus level (Cohen, 2011). RNAseq found differential expression of numerous genes at different floral developmental stages between morphs, with fewer genes being differentially expressed early in development. Throughout development, there appears to be a shift in expression from genes involved in growth and floral development to genes involved in physiological functions.

Tristyly in the monocot *Eichhornia* (Ponterderiaceae) has been shown to follow a similar dialleleic control system (Arunkumar *et al.*, 2017), whereby tristyly seems to be controlled by two loci: the *S* locus and the *M* (modifier) locus. Here, QTL mapping of floral traits was conducted on the *M* locus, finding a large region (10 Mb) that cosegregates with the

M locus. Whether the *M* locus represents a superlocus or contains a small number of pleiotropic loci is still an area of enquiry.

1.2.5 The origins of heterostyly

Heterostyly is, relatively speaking, a rare phenomenon, but is nonetheless a fascinating study system due to the convergent evolution of different heterostylous systems across the angiosperm phylogeny. Heterostyly seems to have evolved independently at least twenty times (Lloyd and Webb 1992), even numerous times independently within a family such as the Boraginaceae (Cohen 2011). In the Primulaceae, the 'model' heterostyly system, however, a single independent origin is most likely (de Vos et al. 2014). Darwin knew the occurrence of heterostyly was spread across the angiosperm phylogeny, and he proposed that the most parsimonious explanation was multiple independent evolution events.

Darwin's hypothesis was that the first stage in the evolution of heterostyly was that in ancestral progenitor species of extant heterostylous taxa there would have been high amounts of variation in the pistil and stamen lengths, or, as in the case of *Linum grandiflorum*, pistil length alone. He then implied that some degree of SI would have favoured reciprocal herkogamy that improved regular cross-pollination. In the 140 years since, Darwin's hypothesis still forms the basis for various competing hypotheses: is the acquisition of *reciprocal herkogamy* or *SI* the first stage in the evolution of heterostyly?

There are two main competing models for the evolution of heterostyly. Charlesworth & Charlesworth (1979) proposed a mutation for a novel incompatible pollen type arising in a SC homostylous ancestor could spread and establish a polymorphism under conditions where the product of selfing rate and inbreeding depression are high. In contrast, Lloyd

& Webb (1992a; 1992b) propose a model whereby ancestors with approach herkogamous flowers (stigma positioned above anthers) experienced an evolutionary event, possibly the result of a mutation, to shorten the style and subsequent mutations to raise the heights of the anthers. Subsequent work has tended to support models similar to the latter 'approach-herkogamy-first' hypothesis, as it suggests at least partial outcrossing and thus the selection for pollination efficiency (Harder and Barrett, 2006).

1.3 Introduction to *Linum*

1.3.1 The Linaceae

The genus Linum (Linaceae) is a study system which provides plenty of opportunities to explore heterostyly. For over 8 millennia (Hillman 1975) humanity has fostered and exploited the natural qualities of flax, which has been grown for its fibres (Zohary and Hopf 2000) and its seeds (Vaisey-Genser & Morris 2003), and this has led to the breeding of cultivated flax (*Linum usitatissimum*). In recent times, the use of flax for linen fibres has heavily declined due to the boom in cotton and synthetic fabric industries, but *L. usitatissimum* is still grown for its seeds (linseed) and remains an important economic crop for many countries in Europe, Asia, Canada and the USA. Scientific interest in flax has also been renewed due to the discovery of high lignan and α- linolenic acid content of its seeds, which have been shown to be effective in protecting against cardivascular dieases and cancer (Cunnane, 2003; Muir and Westcott, 2003).

Having evolved c.44 mya (McDill et al. 2009) and being comprised of over 180 species inhabiting tropical, sub-tropical and temperate regions around the globe, *Linum* displays classic variation (see Figure 1.1) in its species' reproductive biology. *Linum bienne* (the progenitor of *L. usitatissimum*), *Linum strictum*, *Linum tenue* and *Linum trigynum* are all found

throughout southern Europe and are quite common across the southern Iberian peninsula in particular. *L. tenue* and *L. trigynum* are two heteromorphic SI sister species, whereas *L. strictum* and *L. bienne* are both homomorphic and SC. Being part of two separate *Linum* clades, *L. strictum* and *L. bienne* represent at least two independent evolutionary transitions from SI to SC, thus providing the ideal opportunity to understand the selective forces that have resulted in convergence and have shaped patterns of mating system evolution in this genus. Furthermore, the ranges of these four species span the entirety of the Andalusia region, which encompasses populations living at sea level, in e.g. Marbella, to >3000m in the Sierra Nevada mountain range. This system thus allows observations of how gene-flow dynamics across a species' range can be affected by its mating strategy.

With the exception of some work into the heterostyly syndrome (Armbruster et al., 2006), and recent advances by Ushijima *et al.* (2012, 2015) the genetic basis of SI has been largely overlooked in *Linum* by the broader research community. Through a molecular dissection of the genes involved in the *S* locus, and a study of how these can be exposed to selection within populations, we can begin to paint a more colourful and holistic picture of plant mating system evolution than that which is conventionally offered through model systems.

1.3.2 The genetics of heterostyly in *Linum*

Heterostylous *Linum* species demonstrate all the classic traits associated with the heterostyly syndrome, such as strong SI and different pollen morphologies (Rogers, 1979). The breakdown of heterostyly in *L. tenuifolioum* demonstrates that the long homostyles are more frequent (Nicholls, 1985), consistent with the theory that there is a shift to selection for maintaining female function during heterostyly breakdown (Charlesworth and Charlesworth, 1979). In comparison to other heterostylous systems, the difference in the

anther positioning of some *Linum* species is not as extreme, though strong SI and many of the associated morph-specific polymorphisms are commonly found (Dulberger 1992).

As in many other heterstylous taxa, the working assumption is that SS individuals are hemizygous, with superlocus control of heterostyly being suggested, as is found in *Primula*, due to the identification of thrum-specific genes *L. grandiflorum* (Ushijima *et al.*, 2012, 2015). Combining proteomic and transcriptomic approaches, Ushijima et al. (2012) isolated 12 floral morph-related genes with strong indications for the functional control of heterostyly in *L. grandiflorum*, similarly indicating the presence of a superlocus. Of specific interest is the *TSS1* gene, which is exclusively expressed in thrum styles. Ushijima *et al.* (2015) later demonstrated that *TSS1* is absent in pin genomic DNA. This has implications for a hemizygous system similar to *Primula*, although was discovered before the publication of the *Primula* genome. Given the restricted expression of *TSS1* in thrums, it is likely a part of the dominant haplotype. The function of this gene, however, has remained unknown.

1.4 *Linum tenue*: the study species

To provide greater context for the thesis, and the decisions taken for experimental design of the following data chapters, I will use this section to introduce *Linum tenue* and provide information on the background work I conducted prior to data collection and analyses.

1.4.1 *Linum tenue*: the biology, bauplan and phylogeny

The distylous *L. tenue* is a locally frequent annual forb (herbaceous angiosperm that is not grass or grass-like in morphology) of grassland, frequently found in meadows, olive groves and orchards. Its native range spans southwest Iberia and northwest Africa, though recent

phylogenetic studies suggest the African and European lineages are separate species (Ruiz-Martín *et al.*, 2018). Individual plants can grow 30-150 cm tall and show a range of variation in branching patterns and upright or prostrate growth. In the wild, flowering time generally begins from mid-late spring and lasts until mid-summer, though the flowering phenology of populations at higher elevations will shift forwards by a month or so. In the glasshouse, under controlled temperature and watering conditions, flowering time can extend up to 10 months in our experience. During the flowering period there can be a range of developmental stages on each shoot, from young developing buds to open flowers, with ~2-4 flowers open on a shoot at any given time. The flowers display a degree of nyctinasty, and close slightly from the late afternoon to the next morning, and will also close in the presence of rainfall.

The yellow flowers, up to ~2 cm in diameter, of *L. tenue* are actinomorphic with five-fold symmetry: five sepals, petals, stamens and pistils, that are fused at the ovaries. The sepals and petals form a polypetalous (petals are free from one another) corolla with basal nectaries (Valdés-Castrillón, Talavera-Lozano and Fernández-Galiano, 1987). The heterostyly observed in *L. tenue* individuals follows the typical mode of distyly, pin and thrum morphs bearing tall and short organs of roughly equivalent heights (Figure 2.1). Literature on the mating system of *L. tenue* is scarce, and to the author's knowledge no studies have directly investigated the nature of incompatibility in the species. However, from statements made in the literature (Murray, 1986) and anecdotal glasshouse observations made by researchers, the research community as a whole generally accepts the assumption that *L. tenue* is (at least mostly) self-incompatible.

Phylogenetically within the Linaceae, the family is mostly split into two broad clades: the largest being the core *Linum*, comprised of the genera *Linum*, *Cliococca*, *Hesperolinon*, *Screrolinom* and *Radiola*; and a much smaller clade containing *Humiria*, *Viola* and *Hypericum*

(Ruiz-Martín *et al.*, 2018). Within the subclades of the core *Linum* group, *L. tenue* is part of subclade B, with its close relatives being *L. trigynum*, *L. catharticum* and *L. suffruticosum*. Conversely, other more heavily studied species of *Linum*, such as *L. usitatissimum*, *L. bienne* and *L. grandiflorum*, are within the subclade A (Ruiz-Martín *et al.*, 2018).

1.4.2 Initial crossing experiments

During the early stages of this project a number of different experimental approaches were considered to investigate the mating system of *L. tenue*: a morphological approach (Chapter 2.0), a molecular genetics approach (Chapters 3.0 and 4.0) and a classical genetics approach. The classical genetics approach intended to conduct series of controlled intra- and intermorph crosses between *L. tenue* pin and thrum individuals to determine the nature of the self-incompatibility system.

During the first year of the project, seeds collected from four wild populations in 2013 by ACB were grown in the glasshouse. Multiple crosses were subsequently attempted but unfortunately failed to set seed, most likely due to a number of factors. Glasshouse temperatures in the first year were higher than those used for individuals grown for the data collected in Chapter 2.0. Combined with the higher indoor humidity, the resulting pollen appeared to be self-adhering and was not easily transferable among stigmas. Similarly, low germination and high mortality rates, and an outbreak of mildew infection, impacted the amount of plant material available for comprehensive crossing experiments. It was thus decided that the classical genetics approach was outside of the scope of the project, and focus fell upon taking the RNAseq approach described in Chapters 3.0 and 4.0 to investigate the molecular basis of heterostyly.

1.4.3 Experimental design

The work for the construction and sequencing of the RNAseq libraries (Chapter 4.0) was conducted during May and June 2014, using *L. tenue* individuals grown from wild-sampled seeds by ACB in 2013. This was prior to the fieldwork conducted by AF in June/August 2014, where seeds were sampled from more wild populations for the data generated in Chapter 2.0. For reasons outlined in section 1.4.2, the availability of healthy plants still in flowering was limited at that time; the sampling of pin and thrum treatment groups was thus determined by which individuals were suitable and which extractions yielded the highest quality RNA. This resulted in a thrum treatment group composed of three individuals from the CBT population and a pin treatment group composed of one CBT and two GRT individuals. Given that heterostyly is strongly genetically controlled, and in *Linum* likely a result of hemizygosity (Ushijima *et al.*, 2012, 2015), this asymmetrical population sampling is acceptable for broad-scale RNAseq approaches.

1.5 Project aims

1. <u>What are the characteristics of heterostyly in *L. tenue*?</u>

   First, I study the range of variation in heterostyly traits in different populations of *L. tenue*, with seeds sampled from across Andalusia in southern Spain. Specifically, I will analyse the level of reciprocity between the sexual organs using new techniques. I also investigate the development of the sexual organs and cytological mechanisms behind organ positioning. I discuss the findings in light of the evolutionary implications of ecological function.

2. Creating a high-quality reference transcriptome for *L. tenue*

   I will then create an RNAseq dataset from vegetative and floral samples of pin and thrum individuals using Illumina next-generation sequencing. I assemble a *de novo* transcriptome reference using a multiple *k*-mer approach, and demonstrate a method and automated tool I have developed to reassemble consensus unigene sequences in order to create a high-quality transcriptome reference. For the purposes of continuity of the 'biological' data chapters, the work for this is detailed in Chapter 4.0.

3. What are the differences in expression between pin and thrum individuals?

   Next, I will investigate the transcriptome profiles of pin and thrum flowers, using differential expression and analysis of global patterns of expression. I discuss the difficulties and control methods for working with highly dispersed datasets. The analyses will reveal a gene of interest to the *L. tenue S* locus.

4. What is the molecular genetic basis underlying heterostyly in *L. tenue*?

   Finally, using sequence analysis and homology techniques, I will examine the gene of interest and bring together two previously unconnected functional genetics studies to provide strong evidence that it is a potential candidate for the *G* locus constituent of the *L. tenue S* locus.

- Chapter 2 -

The developmental and functional control of distyly in *Linum tenue* (Linaceae)

Pin (top) and thrum (bottom) morphs of *Linum tenue*. Images by Stuart Brooker and Alireza Foroozani

2.0 The developmental and functional control of distyly in *Linum tenue* (Linaceae)

Ali Foroozani[1], Eleanor Desmond[1], Rocío Pérez Barrales[2], Adrian Brennan[1]

1 Department of Biosciences, Durham University, South Road, Durham DH1 3LE, UK
2 School of Biological Sciences, Portsmouth University, King Henry Building, King Henry I Street, Portsmouth, PO1 2DY, UK

2.0.1 Preamble

The following chapter is comprised of a journal manuscript aiming for submission as an original research article at the Annals of Botany. The manuscript structure is thus designed to be in line with Ann Bot's guidelines. However, for the purposes of this thesis, some elements of section/subsection formatting and word limits have been adapted to a thesis structure for continuity and coherency.

Pollinator observation data and additional floral organ measures of open flowers were collected by our collaborator Rocío Pérez-Barrales, and her work contributed to sections 2.2.1, 2.2.2, 2.2.5, 2.3.1, 2.3.2 and 2.3.3.

2.0.2 Abstract

Background and Aims: Distyly is a floral polymorphism involving reciprocal herkogamy that is shaped by selection for disassortative mating between floral morphs through improved pollen transfer efficiency by specialist pollinators. Pollen transfer efficiency can be optimized by minimizing the difference and variance in height of the stamens and pistils of each floral morph while maintaining within morph herkogamy. Distyly is typically controlled by a multiallelic superlocus (known as the S locus) containing morph-specific alleles of a small set of genes that each control different aspects of floral

morphology. We hypothesize that, consistent with their different genetic controls and functions in pollen transfer, there will be reciprocal differences in development and accuracy in different organs and morphs of distylous *Linum tenue* (Linaceae).

Methods: We measured floral organ lengths of flowers of both morphs sampled from wild *L. tenue* populations and recorded pollinators and pollinator behaviour. We grew wild sampled seeds of distylous *L. tenue* in the glasshouse and measured floral organ and cell lengths of different morphs at different developmental stages from young bud to open flower. We analysed the results to measure reciprocal inaccuracy of tall and short reproductive organs and test the factors that influence reproductive organ length and herkogamy.

Key Results: In the wild, flowers were mostly visited by three species of Bombyliidae flies. Smaller flies entered the flower corolla and likely facilitated disassortative pollen transfer. Population differences in tall and short floral organ lengths were evident both in the field and glasshouse. Short floral organs typically contributed to reciprocal inaccuracy by showing bias or height mismatches, while tall organs contributed by showing height variance. Morph-specific differences in developing buds are generated mostly by greater pin pistil lengthening compared to other reproductive organs. Reproductive organ cell length measures show that thrum style cell lengths remain short along the organ in contrast to other reproductive organs with longer cells.

Conclusions: Distyly in *L. tenue* involves multiple types of asymmetry between reproductive organs and floral morphs, indicative of the complex developmental control and different functional constraints. Differences in both cell division and cell elongation in pistil tissue contribute to morph type differences. Greater bias of short organs is probably caused by

developmental constraints, while greater variance of tall organs reflects relatively relaxed selective pressure for effective cross-pollination.

Keywords: adaptive accuracy, distyly, flower development, *Linum tenue*, pistil, stamen, reciprocal herkogamy

2.1 Introduction

The generation and maintenance of genetic diversity is of paramount importance to the survival of many plant species as it improves general resilience and adaptability to changing environmental conditions (Lin, 2011), and can allow plants to mitigate any negative effects imposed by the sessile mode of living. The majority of angiosperms are hermaphrodite, with each flower containing both male and female sexual organs (Renner, 2014). Consequently, plants have developed strategies to maximise outcrossing through wind-pollination and pollination by insects and other animals.

Heterostyly is a breeding system characterised by the presence of two (distyly) or three (tristyly) floral morphs, with distyly being generally more common among flowering plants (Lloyd and Webb, 1992; Barrett and Shore, 2008). Distylous species exhibit a thrum morph, with short styles and tall stamens (S-morph, thrum flowers), and a pin morph with tall styles and short stamens (L-morph, pin flowers) (Fig. 1.4). The reciprocal spatial displacement or herkogamy functions so that the height of the stigma on one morph corresponds to the height of the anther on the other morph to both reduce self-pollination and promote cross-pollination (Barrett, Jesson and Baker, 2000; Keller, Thomson and Conti, 2014).

The evolution of heterostyly has been addressed by two models that differ in their interpretation of the ancestral state of heterostyly and the sequence of trait acquisition. The first model assumes that the common ancestor was homostylous and self-compatible and that a mutation to a novel self-incompatible pollen type spread and established a polymorphism due to the advantage of avoiding inbreeding depression (Charlesworth and Charlesworth, 1979). In contrast, the second model assumes that heterostyly evolved from approach-herkogamous ancestors with pistils taller than stamens, whose populations were invaded by a dominant mutation shortening the style and subsequent mutations to elevate the anthers in the short-styled form to the level of the stigma in the original form. Recent studies have tended to support the approach herkogamy evolutionary model as it prioritizes pollination efficiency rather than outcrossing itself (Kissling and Barrett, 2013; Zhou *et al.*, 2015; Zhu *et al.*, 2015; Yuan *et al.*, 2017). Self-interference occurs when there is competition between the male and female sexual functions within a hermaphroditic individual (Harder and Barrett, 2006). In the case of the self-incompatible flower, if self-pollination occurs, male function is adversely affected as the pollen is wasted, and there may also be a cost to female function: there will be less space on the stigma for compatible pollen grains. Conversely, male fitness in terms of the export of pollen can be compromised if the positioning of the stigma is in such a way that it reduces contact between the anther and the visiting pollinator. In this way, the floral architecture that maximises female fitness can be different to the architecture that maximises male fitness (Johnston *et al.*, 2009). Other than opting for unisexual flowers, herkogamy provides the plant with a resolution for this sexual conflict (Barrett 2002). Heterostyly therefore functions to (i) reduce sexual interference through the spatial separation of sexual organs within the flower, (ii) promote disassortative mating through selective pollen transfer using the reciprocal positioning of the sexual organs, and (iii) prevent selfing through structural and physiological intramorph incompatibilities.

Detailed studies of the morphology and development of flower form in heterostylous species have contributed to understanding the functional significance of floral traits and their evolution (Faivre and Mcdade, 2001; Pérez, Vargas and Arroyo, 2004; Kálmán *et al.*, 2007; Sánchez *et al.*, 2010; Ferrero *et al.*, 2011; Keller, De Vos and Conti, 2012; Sá *et al.*, 2016). For example, developmental studies in the Rubiaceae (Faivre, 2000) provided support to the Lloyd and Webb (1992a,b) hypothesis that distyly evolved from approach herkogamy, a single floral morph with pistils longer than stamens. Morphological studies of distylous species with epipetalous flowers, where stamens are connected to the corolla, often show developmental constraints due to non-independence between stamen height and corolla depth (Faivre, 2000; Faivre and Mcdade, 2001; Pérez-Barrales and Arroyo, 2010; Pérez-Barrales *et al.*, 2014; Sá *et al.*, 2016). Fine tuning of reciprocal pistil-stamen length differences might also contribute to avoidance of inter-specific hybridization as observed in a morphological survey of three *Primula* species (Keller, Thomson and Conti, 2014). Knowledge about pollinators and their relative disassortative pollination efficiency can provide further insights into the function and evolution of distylyous floral traits (Pérez-Barrales and Arroyo, 2010; Simón-Porcar, Santos-Gally and Arroyo, 2014).

Heterostyly has evolved independently in at least 28 plant families and consequently, heterostylous species are a remarkable exemplar of convergent evolution across floral morphology, genetics and physiology (Ganders, 1979; Barrett and Shore, 2008). Early genetic studies in *Primula* concluded that distyly is controlled by two alleles present at a single locus (Bateson and Gregory, 1905; Gregory, de Winton and Bateson, 1923). These S alleles, termed S and s, display dominance-recessive behaviour with SS and Ss individuals producing thrum flowers and ss individuals producing pin flowers. Later, the S locus was shown to be a superlocus or cluster of at least three closely associated genes mostly inherited together by progeny with each gene controlling a different distylous trait consisting of style length (G), stamen length (A) and pollen size (P) (Lewis, 1954; Ernst,

1955). Similar genetic control appears to be the rule for distyly in other plant families (Lewis and Jones, 1992). Recent sequencing of the entire *P. vulgaris* S locus region show that the thrum morph is controlled by a cluster of five linked genes, which are missing in the reciprocal pin morph (Li *et al.*, 2016; McClure, 2016). Therefore the genetic control of distyly in *Primula* is hemizygous, being dependent on the presence or absence of a single thrum haplotype rather than a dominance interaction between two alleles. The lack of corresponding sequence at the pin morph S haplotype also explains the lack of recombination, which is required to keep S-locus genes together (McClure, 2016). The functions of two of the *Primula* S locus genes have been explored further. *GLO^T* (also referred to as *GLO2*) is a paralogue of *GLOBOSA* (GLO1) a B function MADS box gene, one of the master control ABC transcription factors that control floral organ identity (Nowak *et al.*, 2015; Li *et al.*, 2016). *GLOBOSA* is expressed in the second and third floral whorls and is important for specifying petal and stamen identity (Vandenbussche *et al.*, 2004). The expression of *GLO^T* in Primula is associated with the thrum morph (Nowak *et al.*, 2015; Li *et al.*, 2016). *CYP^T* (also referred to as *CYP734A50*) is also found within the thrum S locus region and is a cytochrome P450 type gene expressed exclusively in thrum style tissue that functions to degrade brassinolide, part of the brassinosteroid family of plant growth hormones and reduce style length (Huu *et al.*, 2016; Li *et al.*, 2016). Therefore, recent molecular genetic advances confirm classic models of distyly development for independent control of different floral tissues and traits starting with different S locus genes. Candidates for S locus genes have been identified in study systems from other plant families, including evidence for hemizygous control as some of these candidate genes are present only in the thrum morph (Ushijima *et al.*, 2012; Yasui *et al.*, 2012, 2016).

Linum has been historically important for the study of heterostyly ever since Darwin's pioneering work (Darwin, 1862, 1877) first showed the association between pollination

capacity and different floral morphs. *Linum* is a diverse genus with a wide geographical distribution (Ruiz-Martín *et al.*, 2018); it consists of approximately 180 species that exhibit wide variation in breeding systems, from self-compatible and monomorphic species to distylous species with typical heteromorphic incompatibility system (Dulberger, 1992; McDill *et al.*, 2009; Ruiz-Martín *et al.*, 2018). Distyly is common in *Linum*: it is exhibited by over 40% of the ca. 180 *Linum* species in four of the five sections of the genus and it is known to occur in other subgroups of Linaceae, notably *Reinwardtia* and several members of the Hugonioideae (Rogers, 1979). However, distyly is only found in old world species and it has been difficult to determine with high levels of confidence as to whether distyly or homostyly is the ancestral state for the family (McDill *et al.*, 2009; Ruiz-Martín *et al.*, 2018). Armbruster *et al.* (2006)'s phylogenetic study concluded that distyly could have evolved several times independently within *Linum*'s different lineages, with recent evidence arguing for an independent emergence of distyly in the South African clade of section Linopsis relative to other old world sections (Ruiz-Martín *et al.*, 2018).

The majority of the variation in species and mating systems is found in the Mediterranean area (McDill *et al.*, 2009). Distylous *Linum* species present substantial morphological variation in the degree of differentiation in traits between morphs and reciprocal herkogamy (Wolfe, 2001; Armbruster *et al.*, 2006), with some species displaying variation only in stigma height, as it is the case of *L. grandiflorum* (Darwin, 1877) and other species showing breeding system variation across the range of the species (Nicholls, 1985, 1986). Less is known in monomorphic *Linum* species, but it appears that they show a tendency towards outcrossing, possibly aided by herkogamy. For example, cultivated flax, *L. usitatissimum*, does not self-pollinate immediately because the anthers face outwards and are slightly distanced from the stigmas until after the opening of the flower (Kadam and Patel, 1938), only making contact for self-pollination if an outcrossing pollination has not

occurred. Another example is that of *L. lewisii*, which is self-compatible yet relies on insect visitation for seed production (Kearns and Inouye, 1994).

Heterostyly in *Linum* is purported to be controlled by an S locus superlocus; the evidence being that all of the self-compatible monomorphic populations of *L. tenuifolium* that have been identified resemble the pin morph (Nicholls, 1985). More recent proteomic and transcriptomic studies of style dimorphism in *L. grandiflorum* have identified a shortlist of S locus gene candidates with exclusive or enhanced expression in thrum styles (Ushijima *et al.*, 2012). One of these candidate genes, *THRUM STYLE-SPECIFIC 1* (*TSS1*), also appears to be hemizygous, present only in thrum individuals, analogous to recent findings for the genetic control of distyly in *Primula* (Ushijima *et al.*, 2012; Kappel, Huu and Lenhard, 2017).

Here, we present a detailed study of pollination, floral morphology and development in the distylous annual species, *L. tenue* with a view to separating the functional and developmental components of this floral syndrome. We observed pollinator visits in the field and measured floral organ length in open flowers in the field and the glasshouse, and floral organ cell lengths in developing flower buds in the glasshouse. Specifically, we addressed the following hypotheses: *(i)* pollinator visitors to *L. tenue* in the field facilitate disassortative mating between morph types, *(ii)* male and female floral organ height of pin and thrum morphs show differences in their contributions to reciprocal herkogamy and, *(iii)* male and female floral organs of pin and thrum morphs show differences in their development. In general we expect to find multiple sources of asymmetry in trait expression of distyly consistent with different floral organs in different morphs experiencing different functional constraints and genetic control.

2.2 Materials and Methods

2.2.1 Pollinator observations

Linum tenue is a locally frequent annual forb of grassland growing 30 to 150 cm tall that occupies meadows, olive groves, and orchards. Its native range extends through southwest Iberia and northwest Africa, although recent phylogenetic studies suggest the two lineages are separate species (Ruiz-Martín *et al.*, 2018). This study used samples from natural populations across the region of Andalusia, Spain. Observations of insect visits were conducted during July 2015 in two populations (r10, r17; Figure 2.1) in patches of grassland where 10-40 individual plants with open flowers were present. For intervals of ca. 10 minutes, between 10:00 and 17:00 CET, small areas where 3-5 plants are in close proximity were observed for approaching pollinators. Observations were carried out at different patches after two intervals. Any event where an insect was observed approaching and making contact with a flower on the plants under observation was classed as a visitation. Upon a visitation event, the flower in question was carefully approached, the morphotype of the individual plant was noted, and observations were made on the mode of contact between the insects body and the stigmas and anthers. Insects were also identified to the lowest possible taxonomic level, and the collection of a nectar or pollen reward from individual plants was recorded.

2.2.2 Plant material

Flowers or seeds were collected from 30-50 separately sampled maternal plants from 4 and 16 wild populations in 2013 and 2014 respectively from across the region of Andalusia, Spain (Fig. 2.1, Appendix I). Plants were sampled at random, with at least one-meter distance among sampled individuals, to avoid pseudoreplication and estimate the

relative presence of pin and thrum individuals. Flower or seed collection involved one newly open flower per individual plant that was preserved in separate 1.5 mL screw-top tubes with 70% ethanol. Seed collection involved several ripe fruit capsules per plant placed in separate glassine envelopes. Later, in the glasshouse, one seed per maternal plant was germinated in individual 10 cm diameter pots of two thirds John Innes no. 2 compost (ICL, Ipswich, UK) and one third Perlite (LBS horticultural supplies, Colne UK) grown to flowering in greenhouses at the Department of Biosciences, Durham University, UK, under semi-controlled growth conditions of 20°C for 16 hours of day length and 15°C for 8 hours of darkness. Upon flowering, approximately all individuals from each population were visually classified as pin or thrum floral morphs prior to more detailed floral organ measures.

2.2.3 Floral organ measures

Measurements were made of field collected flowers stored in ethanol or one to three freshly opened flowers per glasshouse grown individual. Flowers were dissected whorl-by-whorl under a dissecting microscope and the vertical height of the sepals, petals, stamens, and their component filaments and anthers, pistils and their component styles, stigmas, and ovaries were measured from the base of the ovary as reference as shown in Figure 2.1, using a combination of Vernier calipers and analysis of photos using ImageJ software (Schneider, Rasband and Eliceiri, 2012). The digital photographs of dissected flowers were made against a 1 mm ruled graph paper background using a Leica M80 light microscope (Leica Biosystems, Nussloch, Germany) set at 7.5 times magnification connected to a computer. Anther length of field collected flowers was not measured because *L. tenue* flowers usually lose anthers when stored in alcohol. Multiple investigators contributed different measures at different times, so to account for observer differences in

the glasshouse data a constant was added to each trait measure for each observer, in order to have the same mean after correction for morph frequency differences, as the samples measured using ImageJ, which were considered to be the most accurate measures.

One to three developing flowers of various bud sizes, representing different developmental stages from approximately ten to one days prior to opening, were collected from each glasshouse grown individual to describe the morphological development of stamen and pistil in pin and thrum flowers. Flowers were dissected and photos of floral organs were taken and measured using imageJ, as described for the open flowers.

2.2.4 Analysis of floral organ length and herkogamy in open flowers

Using measurements from newly open flowers, linear mixed effect models were used to test floral organ lengths and herkogamy; the difference in height between pistils and stamens for the fixed effects of flower morph (pin or thrum), flower size measured as petal length, and their interaction, while controlling for the random effects of sample individual nested within sample population. The lmerTest R package (Kuznetsova, Brockhoff and Christensen, 2017) was used to fit models and assess the significance of main effects using the restricted maximum likelihood approach. The significances of random effects were assessed by analysis of variance comparisons of nested models dropping individual effects first, followed by population effects. The proportions of variances (coefficient of determination, r^2) explained by the models were calculated using the MuMIn R package (Bartoń, 2009) that evaluates both marginal r^2 for fixed effects only and conditional r^2 for both fixed and random effects.

To identify differences in flower development, linear mixed effect models were used to test floral organ lengths and herkogamy for the fixed effects of developing bud morph type, bud developmental stage (measured as petal length), and their interaction, while controlling for the random effects of sample individual nested in sample population, using an analogous approach to the open flower analysis.

2.2.5 Analysis of morph frequency and adaptive inaccuracy

Chi-square tests were used to assess population deviations from the expected 1:1 morph ratio under complete disassortative mating. Inaccuracy in the reciprocal placement of tall and short organs was estimated using the adaptive inaccuracy method developed by Armbruster *et al.* (2017). The disassortative pollination function of distylous flowers predicts that the optimal position of tall pin stigmas is represented by the position of tall thrum anthers. Similarly, the optimum position of short thrum stigmas is represented by the position of short pin anthers. Hence, it is possible to estimate inaccuracy in reciprocity by studying the contribution of differences between population means of tall organs and short organs (bias), and the variance of organ position (imprecision) to departures from perfect matching between anthers and stigmas of pin and thrum flowers. The reciprocal inaccuracy measures were done on height of filaments and styles, excluding anthers and stigmas because anthers had been mostly shed for the wild-sampled flowers. For each sample population with measures for nine or more flowers of each morph type, floral measures were analysed using a custom script (Scott Armbruster, pers. comm.) to generate raw and mean-standardized (dividing by average organ height per population) adaptive inaccuracy estimates and confidence intervals based on equations 16 to 21 in Armbruster *et al.* (2017). Adaptive inaccuracy measured in the field informs about the effects of environmental variation on the development of reproductive organs, and its influence on

the reciprocal placement of tall and small organs. The same measure of glasshouse grown plants is a better estimator of the genetic contribution of floral developmental variation to reciprocity after controlling for environmental variation.

2.2.6 Analysis of filament and style cell length

Whole filaments and styles were separated from freshly harvested newly open flowers from glasshouse-grown individuals. Floral organs were immediately mounted on microscope slides, stained with 0.05 % toluidine blue solution, and viewed at 100x to 400x magnification using differential interference contrast Leica DMI2500 microscope (Leica Biosystems, Nussloch, Germany) with an eyepiece graticule and photographed with a Panasonic GP-US932HAE camera (Panasonic UK, Bracknell, UK). To account for localized differences in cell length at different positions along the organs, each style and stamen filament was classified into five approximately equal length regions counting from base to tip. Using appropriate planes of view that allowed clear discrimination of individual cells and the eyepiece graticule, images of each section were taken at 400x magnification. Using ImageJ software (Rasband 2017), up to 20 cells were chosen at random from each image and measured along their longest axis to the nearest 0.1 μm.

Linear mixed effect models were used to test the potential effects influencing the dependent variable, floral organ cell length. The fixed effects were organ type (pistil or stamen), flower morph (pin or thrum), and organ region treated as an ordinal variable going from the base to the tip. There were five levels of region so models with higher order linear, quadratic, cubic, and quartic relationships were fitted for this variable. The random effects were sample individual nested in sample population as for the floral organ

length analyses. The significance of random effects and both marginal and conditional r^2 coefficients were calculated as for the floral organ length analyses.

2.3 Results

2.3.1 Pollinator observations

A total of 4 and 8 hours of observations were accumulated in r10 and r17, respectively. Weather conditions were sunny as is typical of the Andalusian climate during July. In r10, flowers were visited equally by *Usia cf. pusilla* and *Bombylius cf. major* (Bombiliidae), with 17 visitation events by each visitor during the period of observation. In r17, most of the visits were paid by *Usia cf. pusilla*, and less frequently by small Halictidae (*cf. Lasioglossum*), with 16 and 3 visits respectively. In all cases, insects visited to collect nectar, but the behaviour was different. Both *Usia* and *cf. Lasioglossum* landed on the petals and crawled down to the bottom of the flower towards the nectaries to collect nectar, visiting all five nectaries and making contact with short reproductive organs more often than tall organs with the dorsal part of the body. Specifically, it was observed that pin pollen was deposited on the head and the thorax of the insects, and thrum pollen on the abdomen, and less often on the thorax. *Bombylius* visited by hovering in front of the flowers, but visits were fleeting and it was not possible to retrieve detailed information on contact rate with anthers and stigmas.

2.3.2 Open flower floral organ lengths

The results of mixed model analysis showed that the lengths of pistils and stamens and herkogamy of open flowers could be predicted by morph type as expected (Table 2.1, Appendix II). Moreover, flower size, as measured by petal length also had a significantly positive effect on pistil and stamen length, indicative of allometry between floral organs,

but not for within-flower herkogamy, indicative of selection for reciprocal accuracy of floral organ lengths. There was no evidence for an interaction effect between flower size and morph type in the expression of pistil and stamen length or herkogamy. Individual nested in population was shown to be a significant random effect in all tests, while the random effect of population was also significant in the test of pistil length.

2.3.3 Morph frequency and adaptive inaccuracy

Population morph ratio (the percentage of pin flowers) ranged between 51.3% to 59%, and in all cases Chi-square analyses showed that populations did not depart significantly from a 1:1 ratio (Table 2.2). Field sampled pin and thrum organs were longer than glasshouse grown flowers in general, with the exception of the field sample from mva that showed similar floral organ lengths to glasshouse material (Table 2.2). It might be that there is substantial genetic or environmentally-induced variation in flower size. The measures we report here are within the range given for the species in Flora Vascular de Andalucía Occidental (Valdés-Castrillón, Talavera-Lozano and Fernández-Galiano, 1987). Table 2.3 summarizes the contribution of bias and imprecision to inaccuracy of tall and short organs. Overall, these results show that the contribution of bias and imprecision to inaccuracy in reciprocity for tall and short organs is different. Measures of bias were generally greater for short organs compared to tall organs, with the exception of population snv, in which bias was similar for tall and small organs. In contrast, tall organs consistently showed greater imprecision than small organs.

2.3.4 Developing floral organ lengths

The youngest buds measured were less than 4 mm long and approximately 10 days from flowering while the longest buds were just less than 10 mm long and had fully developed petals that were about one day from opening (observations based on other tagged but non-harvested flower buds). The results of mixed model analysis showed that morph type was an important predictor of pistil, stem length and herkogamy (Table 2.4). Petal length showed significant positive relationships in all tests indicative of floral organ growth during development. There was a significant interaction effect between morph type and petal length for developing reproductive organs and herkogamy, indicating differences in the rate of growth and development of organ type in each morph. These differences are illustrated in Figure 2.2. Both pistils and stamens lengthen during flower development (petal lengthening) but they do so at different rates and to different extents in different morph types; thrum stamens have a slightly steeper slope than pin stamens and pin pistils have a steeper slope than thrum pistils (Figure 2.2). These patterns support that relative growth rate is different between tall and small organs, as expected for a distylous species. When comparing the same organs in each morph, the differences in slope is larger between pin and thrum pistils than the comparison between pin and thrum stamens. The pin pistil shows the greatest rate of growth and variance compared to the other organs, possibly reflective of less canalization and more developmental noise. The mixed model analysis results (Table 2.4) indicate that often, the random effects of individual nested in population and population make significant contributions to reproductive organ length also.

2.3.5 Floral organ cell lengths

A summary of the results of mixed model analysis of floral organ cell length are presented in Table 2.5. The random effects of both individual nested in population and population were both highly significant. Cell lengths were subject to significant interaction effects between morph, organ, and region. Both positive linear and negative quadratic ordinal organ region effects were detected. These interacting effects on cell length are visualized in Figure 2.3. Pin styles and filaments of both morphs have shorter cells about 50 μm long in region 1 at the base of each organ that increase to a constant limit of about 125 μm by region 3. In contrast, cell lengths stay consistently short at about 50 μm across all regions of thrum morph styles. Therefore, cell length seems to contribute to differences in style length between the two morphs but not for filaments. Therefore, differences in the filament length are achieved by alternative growth mechanisms, such as differences in cell division.

2.4 Discussion

Detailed measurements of floral reproductive organ and cell length revealed multiple sources of asymmetry in the expression and function of distyly in *L. tenue*. Altogether, the findings are consistent with a functional distyly breeding system that serves to promote disassortative mating between floral morphs. Our results support that functional constraints differ between each reproductive organ in each morph type. These findings also support that distyly in *L. tenue* is controlled by a superlocus consisting of multiple genes that contribute to separate floral traits that together make up the distyly floral syndrome.

2.4.1 Pollinator visitation in the field facilitate intermorph disassortative mating

Pollinator observations found that pollination by small *Usia* and *cf. Lasioglossum* flies that enter the corolla to feed on nectar tend to make most contact with short reproductive organs suggesting biases in disassortative pollen transfer. This is relevant to reproductive success as *L. tenue* is self- and intramorph-incompatible and depends on insect pollination for cross-pollination (Murray, 1986; authors pers. obs.) It is currently unclear whether more superficial visits by larger *Bombylius* flies alter this pollination dynamic.

2.4.2 Ancillary floral traits contribute to the distyly floral syndrome

Open flowers showed pistil and stamen length differences between morphs consistent with expectations for reciprocal herkogamy (Table 2.1). The lengths of most other floral organs that were not directly involved in the reciprocal herkogamy did not show significant differences between morphs, indicating that they are not part of the distyly floral syndrome nor under *S* locus control.

2.4.3 The expression of distyly is robust to environmental influences

Larger flowers produced longer reproductive organs but the degree of herkogamy between male and female organs was maintained (Tables 2.1, 2.3). This result highlights the importance of within-flower herkogamy in limiting assortative mating and the tight developmental genetic control of this key trait (Barrett, 2002). To the authors' knowledge, no study has previously considered the influence of environment on the expression of distyly. There was a trend for wild sampled flowers to be larger than glasshouse sampled flowers (Table 2.2). Unfortunately, different populations were sampled under their own

respective conditions, meaning that the extent to which flower size is environmentally or genetically controlled could not be distinguished. This issue should be investigated in future studies. Within field and glasshouse analyses, both population and individual were often significant random factors suggesting at least some genetic contribution to flower size.

Individual- and population-level variation in reproductive organ length and herkogamy was found for *L. tenue* (Table 2.3), in common with other heterostylous species (Richards and Koptur, 1993; Eckert and Barrett, 1994; Faivre and Mcdade, 2001). The presence of genetic variation in the expression of floral morphology associated with distyly could be an important source of standing variation to permit rapid and flexible breeding system responses to a changeable pollination environment (Kissling and Barrett, 2013; Jiang *et al.*, 2018; Simón-Porcar, 2018). Differences in distyly expression between populations might be driven by spatial variation in pollination efficiency or the presence of other species that might compete for shared pollinators (Kálmán *et al.*, 2007; Keller, De Vos and Conti, 2012; Kissling and Barrett, 2013). For example, fine tuning of reciprocal pistil-stamen length differences might contribute to avoidance of interspecific hybridization as observed in a morphological survey of three *Primula* species (Primulaceae) (Keller, Thomson and Conti, 2014). More extensive studies to explicitly examine individual- and population-level differences in distyly in *L. tenue* would help to better understand the fitness consequences of variation in distyly expression under local pollination conditions.

2.4.4 Male and female floral organ lengths of both morphs show different contributions to reciprocal herkogamy

Few other studies of distyly have used the new adaptive inaccuracy measure proposed by Armbruster *et al.* (2017) that allows identification of the contribution of different reproductive organs to overall inaccuracy. However, the reanalysis of floral morphology data for three *Primula* species, from the study of Keller et al. (2012), provides some examples from other distylous species against which to compare our results. Measures of adaptive inaccuracy also provide insights into the functional constraints on heterostyly in *L. tenue* by considering different sources of inaccuracy separately. Typically in this study, greater bias was found for short reproductive organs, while greater imprecision was found for tall reproductive organs. The bias, or mismatch in short floral organ heights, was due to thrum pistils generally being shorter than pin stamens. Thrum pistils also showed less imprecision or variance than pin stamens suggesting tighter regulation of growth in this specific organ. This finding matches with another observation of this study and a study of *L. grandiflorum* by Ushijima *et al.* (2015) that thrum style cells were generally shorter compared to other reproductive organs. Therefore, it is possible that limited cell expansion in this tissue leads to less imprecision in whole organ length. Only one out of the three studied *Primula* species, *P. veris*, showed also a greater bias in short organs than tall organs in the reanalysis presented by Armbruster *et al.* (2017), highlighting that sources of bias and imprecision are labile features of distyly that can differ between species.

Greater imprecision was observed for tall floral organs relative to short floral organs consistent with findings of Armbruster *et al.* (2017). Our study of developing flower buds found also that tall organs showed greater variance than short organs, particularly in the later stages of flower development (Figure 2.2). In addition, there was greater variation in

the longer cells of reproductive organs (Figure 2.3). These results might reflect the common observation that tall stamens and pistils are more effective than short organs at disassortative pollen transfer as they tend to make more frequent contact with pollinators bodies (e.g. Wolfe, 2001; Lau and Bosque, 2003; García-Robledo, 2008; Pérez-Barrales and Arroyo, 2010; Zhu *et al.*, 2015).

Preferential cross pollination between tall reproductive organs can lead to further breeding system evolution to dioecy where the less efficient male and female functions of short organs in thrum and pin morph types, respectively, are lost (Barrett, Morgan and Husband, 1989; Eckert and Barrett, 1994). Alternatively, cases of breeding system shifts to selfing are frequently characterised by loss of the thrum morph type due to its inferior female fitness (Pannell, Dorken and Eppley, 2005). Relatively relaxed selection for accuracy in tall organs relative to short organs might allow the persistence of greater standing variation in populations and/or more relaxed developmental control of these traits (Sanchez, Ferrero and Navarro, 2008; Keller, De Vos and Conti, 2012; Armbruster *et al.*, 2017). There is evidence from this study that floral development contributes to the greater imprecision observed for tall organs in the form of increasing variance in pin pistil height during flower development and greater variation in the longer cells of reproductive organs (Figures 2.2 and 2.3).

2.4.5 Male and female floral organs of both morphs show differences in their development

Developing floral organs of each morph showed consistent differences in growth rates, primarily driven by the different pistil growth rates, from a relatively early stage from one week prior to flower opening (Figure 2.2). These developmental differences are most likely

completed just before flower opening, as has been noted for *L. grandiflorum* (Ushijima *et al.*, 2015). Analysis of floral morph development and cell lengths identified at least two distinct mechanisms to achieve reproductive organ height differences. Developing pin flowers showed enhanced growth of the tall pin pistil during floral development (Figure 2.2). Since pin style cell lengths of mature flowers are not significantly different from the cell lengths of pin and thrum filament (Figure 2.3), this additional length has probably been achieved through increasing cell number by extra cell division in this organ. The second developmental mechanism to achieve morph differences is reduced cell elongation in short thrum styles (Figure 2.3). Therefore, pistil tissues appear to employ two different developmental mechanisms to control height in each floral morph. The developmental control of height differences between stamens between pin and thrum morphs is not apparent from this study. These observations of organ-specific developmental mechanisms support the model for genetic control of distyly by a superlocus consisting of multiple physically linked genes, each contributing to a distinct floral trait (Lewis and Jones, 1992).

2.5 Concluding remarks

Detailed morphological and developmental analysis of reproductive organ height in distylous *L. tenue* has revealed interesting sources of asymmetry and inaccuracy between male and female reproductive organ height in different floral morphs. Morph-specific differences are driven by both arrested cell elongation in short thrum pistils and enhanced cell division in long pin pistils, highlighting the importance of pistil height differences for the expression of distyly. In terms of adaptive inaccuracy, short reproductive organs show a greater bias (mismatch) in organ heights than tall organs, while tall organs show greater imprecision (variance) in organ height. In particular, thrum pistils show the least variance but are consistently shorter than their matching pin stamens. Finally, the expression of

distyly is influenced by both genetic and environmental factors. These fine-scale morphological and developmental details raise many further questions about the potential evolutionary and functional constraints on distyly in this species and more generally. Further understanding will require detailed ecologically and phylogenetically informed studies of *S* locus genetics, floral morphology, and pollination biology in more *Linum* species and other distylous groups.

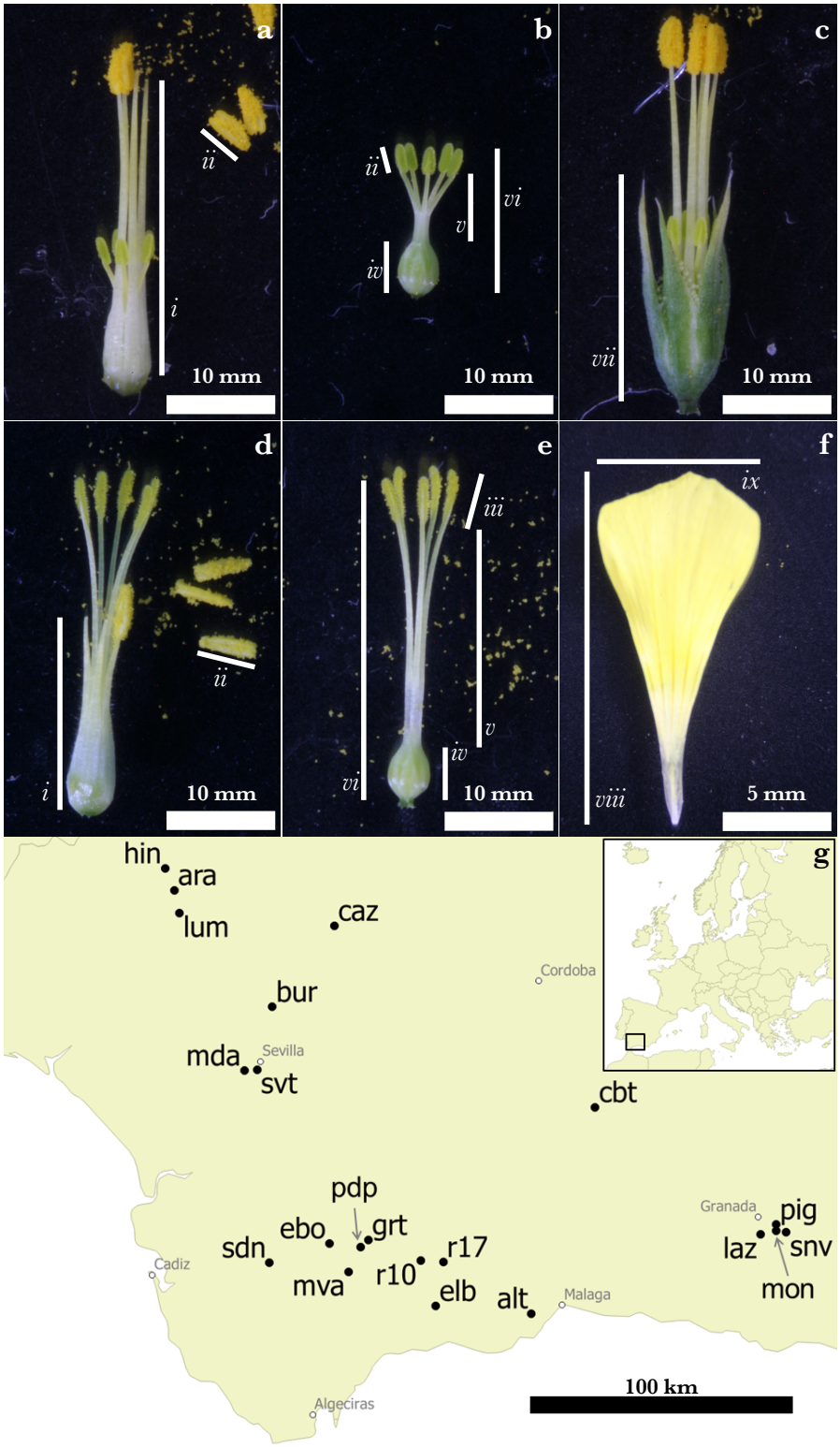


Figure 2.1 Dissected *Linum tenue* flowers showing floral organ measures and map of sample populations. (a) Thrum morph with outer two whorls removed, (b) Thrum morph with outer three whorls removed, (c) Thrum morph with only the second whorl (petals) removed, (d) Pin morph with outer two whorls removed, (e) Pin morph with outer three whorls removed, (f) a removed petal. The lengths measured are: i = filament, ii = anther, iii = stigma, iv = ovary, v = style, vi = pistil, vii = sepal, viii = petal height, ix = petal width. (g) Map of sampled region. The inset shows the sampled region within the context of Europe. Populations were sampled during the summers of 2013, 2014, and, 2015 by ACB, AF, and RPB.

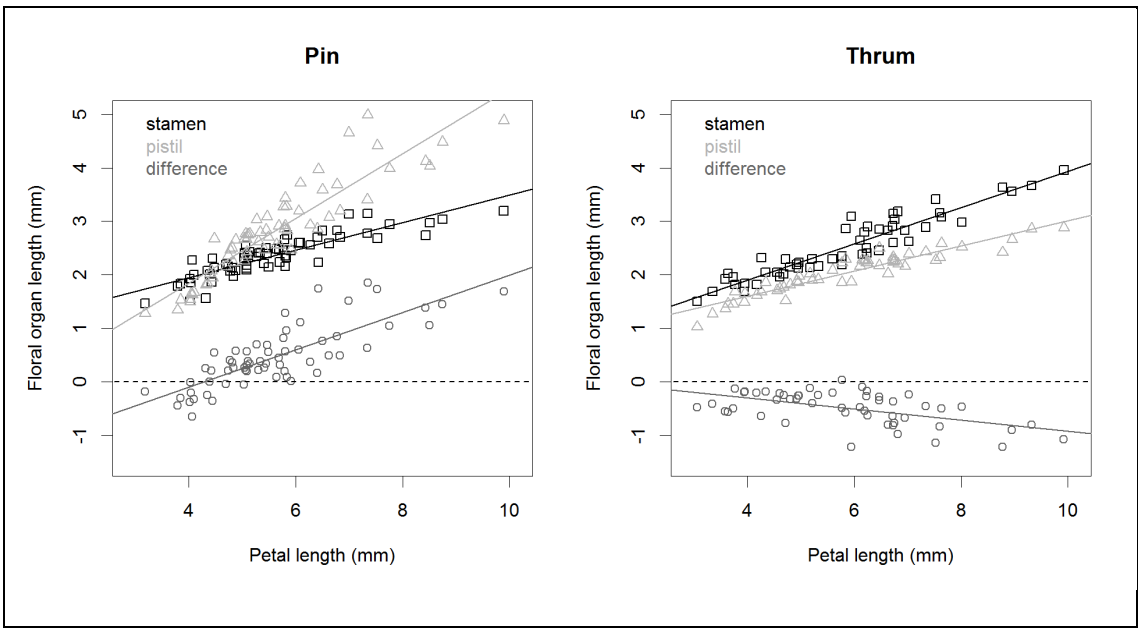


Figure 2.2 Relationships between Linum tenue floral organ length and petal length and floral morph in developing flower buds. A 1:1 petal length to floral organ length aspect ratio was used to plot the values. Lines indicate best fit linear models between the plotted variables.

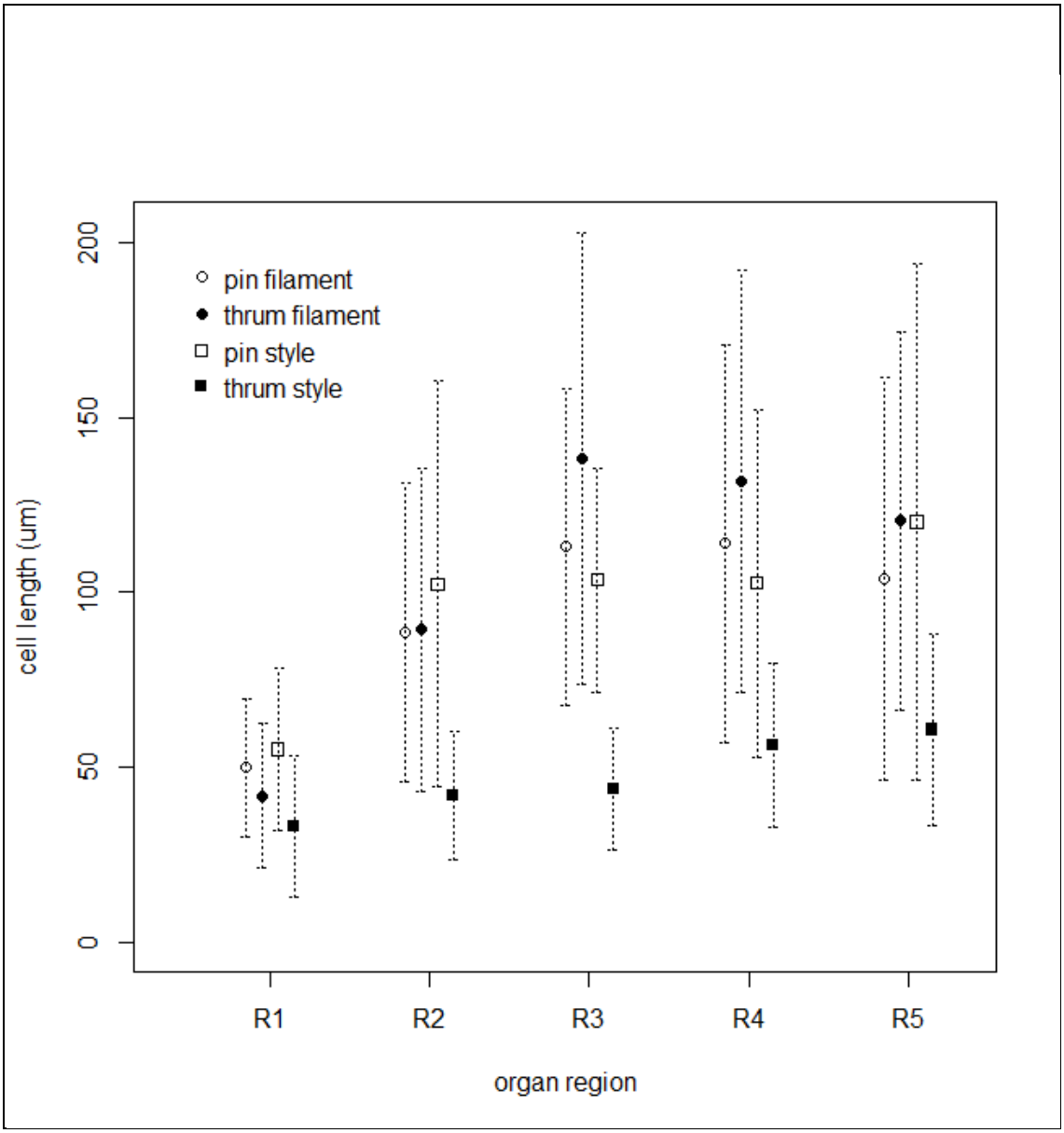


Figure 2.3 Cell length means and standard deviations for *Linum tenue* floral morph, organ type, and region of organ. Organ regions (from open flowers) divide the total length of each organ into five, starting from the base (R1) to the tip of the organ (R5).

Table 2.1 Mixed model analysis results for *Linum tenue* open flower measurements. Mixed models were performed on non-transformed data using the lmer REML fit function of the R lmerTest. The p values of mixed effects were evaluated using t-tests with Satterthwaite degrees of freedom approximations, while the p values of random effects were evaluated by sequentially dropping random effects from the model and comparing the prior model using the anova function with likelihood ratio tests. R2 values were calculated using the r.squaredGLMM function of the R MuMIn package and are either conditional for the full mixed model or marginal for fixed effects only.

Response	Random effects	No. obs.	Variance	SD	*P* value	R^2 cond.	Fixed effects	Estimate	SE	*p* value	R^2 marg.
Pistil length	individual x population	150	0.117	0.343	**9.66^{e-15}**	0.955	intercept	4.822	0.241	**$<2.00^{e-16}$**	0.916
							morph	-2.009	0.392	**4.94^{e-07}**	
	population	16	0.001	0.024	**1.50^{e-02}**		petal length	0.191	0.017	**$<2.00^{e-16}$**	
	residual		0.140	0.374			morph*petal length	-0.099	0.027	**3.52^{e-04}**	
Stamen length	individual x population	115	0.106	0.325	**2.65^{e-07}**	0.905	intercept	3.418	0.334	**$<2.00^{e-16}$**	0.841
							morph	1.868	0.483	**1.35^{e-04}**	
	population	16	0.004	0.061	**2.92^{e-02}**		petal length	0.142	0.024	**1.19^{e-08}**	
	residual		0.159	0.399			morph*petal length	0.024	0.034	4.76^{e-01}	
Herkogamy (pistil – stamen)	individual x population	115	0.141	0.375	**$<2.00^{e-16}$**	0.986	intercept	1.658	0.307	**1.39^{e-07}**	0.968
							morph	-4.143	0.443	**$<2.00^{e-16}$**	
	population	16	<0.001	<0.001	7.71^{e-02}		petal length	0.030	0.022	**1.85^{e-01}**	
	residual		0.119	0.345			morph*petal length	-0.105	0.031	**9.44^{e-04}**	

Table 2.2 Floral morph frequencies and reproductive organ heights of *Linum tenue* sample populations. Population sample size for the two floral morphs (P and T), mean organ length for each tall and small organ (P style S; T filament A; T style a; P filament a), mean organ length across all organ types, variance of tall organs and short organs, population morph ratio and Chi-square statistic (in all comparisons, d.f.=1; in all cases, population morph ratio did not depart from the 1:1 morph ratio).

Population	Sample size		Mean length of tall organs (mm)		Mean length of small organs (mm)		Mean organ length (mm)	Variance of tall and short organs (mm^2)				Population morph ratio	Chi-square
Population	P	T	S	A	s	a		Var (S)	Var (A)	Var (s)	Var (a)	% P flowers	
snv	36	25	12.0	11.406	6.125	6.869	9.165	0.985	1.940	0.302	0.696	59.02	1.9836 (n.s.)
pdp	45	34	11.390	12.004	5.836	7.028	9.084	0.867	1.052	0.212	0.564	56.96	1.5316 (n.s.
r10	42	37	12.150	12.185	6.187	7.435	9.509	0.891	0.831	0.347	0.658	53.16	0.3165 (n.s.)
r17	40	37	12.766	12.962	6.492	7.704	9.991	0.415	1.323	0.205	0.505	51.95	0.1168 (n.s.)
mva	39	37	7.231	7.131	3.900	4.406	5.671	0.258	0.244	0.107	0.126	51.32	0.0526 (n.s.)

Table 2.3 Adaptive inaccuracy of floral organ height of Linum tenue sample populations. Legend same as for Table 2.2. F and G indicated plants from the field or glasshouse respectively.

Popn. (seed source)	Popn. Type	Sample size		Mean length of tall organs (mm)		Mean length of small organs (mm)		Mean organ length (mm)	Variance of tall and short organs (mm^2)				Popn. morph ratio	Chi-square
		P	T	S	A	s	a		Var (S)	Var (A)	Var (s)	Var (a)		
Mva	F	39	37	7.231	7.131	3.900	4.406	5.671	0.258	0.244	0.107	0.126	51.32	0.053 (n.s.)
Pdp	F	45	34	11.390	12.004	5.836	7.028	9.084	0.867	1.052	0.212	0.564	56.96	1.532 (n.s.
R10	F	42	37	12.150	12.185	6.187	7.435	9.509	0.891	0.831	0.347	0.658	53.16	0.317 (n.s.)
R17	F	40	37	12.766	12.962	6.492	7.704	9.991	0.415	1.323	0.205	0.505	51.95	0.117 (n.s.)
Sdn	F	36	25	12.000	11.406	6.125	6.869	9.165	0.985	1.940	0.302	0.696	59.02	1.984 (n.s.)
Ara	G	23	12	6.763	6.436	3.422	4.677	5.449	0.391	0.135	0.123	0.292	-	-
Bur	G	9	11	6.181	7.267	3.573	4.646	5.417	0.446	0.801	0.176	0.216	-	-
Cbt	G	10	10	6.359	6.981	3.466	4.605	5.352	0.334	0.189	0.040	0.083	-	-
Ebo	G	10	9	7.110	6.772	3.486	4.550	5.498	0.137	0.395	0.056	0.118	-	-
Hin	G	9	8	6.368	7.009	3.302	3.946	5.156	0.295	0.512	0.085	0.141	-	-
Lum	G	9	15	6.316	6.772	3.595	4.101	5.193	0.175	0.366	0.094	0.062	-	-
Mon	G	12	14	5.946	6.778	3.540	3.854	5.039	0.357	0.979	0.239	0.200	-	-
Snv	G	10	14	6.151	6.897	3.538	4.415	5.245	0.304	0.317	0.073	0.030	-	-
Svt	G	16	18	6.463	7.138	3.764	4.363	5.433	0.469	0.212	0.141	0.120	-	-

Table 2.4 Mixed model analysis results for *Linum tenue* developing bud measurements. Legend as for Table 2.2.

Response	Random effects	No. obs.	Variance	SD	*p* value	R^2 cond.	Fixed effects	Estimate	SE	*p* value	R^2 marg.
Pistil length	individual x population	56	0.040	0.20	**1.16e-02**	0.926	intercept	-0.151	0.166	3.65e-01	0.860
	population	9	0.016	0.126	**1.93e-03**		morph	1.099	0.224	**3.88e-06**	
	residual		0.062	0.249			petal length	0.677	0.027	**<2.16e-16**	
							morph*petal length	-0.361	0.037	**5.33e-15**	
Stamen length	individual x population	56	0.013	0.116	**5.05e-03**	0.904	intercept	1.010	0.097	**<2.16e-16**	0.840
	population	9	0.002	0.043	2.32e-01		morph	-0.508	0.134	**2.58e-04**	
	residual		0.023	0.151			petal length	0.242	0.016	**<2.16e-16**	
							morph*petal length	0.106	0.022	**7.97e-06**	
Herkogamy (pistil − stamen)	individual x population	56	0.588	0.243	**1.32e-05**	0.911	intercept	-1.133	0.156	**2.11e-10**	0.781
	population	9	0.136	0.117	**1.01 e-03**		morph	1.610	0.210	**1.72e-11**	
	residual		0.050	0.223			petal length	0.430	0.025	**<2.00e-16**	
							morph*petal length	-0.467	0.034	**<2.00e-16**	

Table 2.5 Mixed model analysis for *Linum tenue* floral organ cell measurements. Legend as for Table 2.2 except that the fixed effect, region, was treated as an ordinal factor with five levels permitting tests of linear (L), quadratic (Q), cubic (C) and quartic (^4) models of this factor.

Response	Random effects	No. obs.	Variance	SD	*p* value	R^2 cond.	Fixed effects	Estimate	SE	*p* value	R^2 marg.
cell length	individual x population	9	0.06002	0.24500	**<2.16e-16**	0.579	intercept	4.39078	0.11137	**1.22e-09**	0.444
	population	2	0.00000	0.00000			organstyle	0.07409	0.03059	**1.56 e-02**	
	residual		0.18785	0.43340	**6.02 e-08**		typethrum	0.05059	0.16716	7.71 e-01	
							region.L	0.51854	0.04361	**<2.16e-16**	
							region.Q	-0.38458	0.04526		
							organstyle:typethrum	-0.81119	0.04514	**<2.16e-16**	
							organstyle:region.L	0.02969	0.06618		
							organstyle:region.Q	0.10935	0.0683	**<2.16e-16**	
							typethrum:region.L	0.22670	0.06717		
							typethrum:region.Q	-0.19306	0.06794	6.54 e-01	
							organstyle:typethrum:region.L	-0.26319	0.09963	1.10 e-01	
							organstyle:typethrum:region.Q	0.45374	0.10046	**7.57 e-04**	
										4.55 e-03	
										8.33 e-03	
										6.76 e-06	

- Chapter 3 -

Expression analysis of long- and short-styled heterostylous morphs of *Linum tenue* (Linaceae)

Pin-morph *Linum tenue*. Image by Stuart Brooker and Alireza Foroozani

3.0 Expression analysis of long- and short-styled heterostylous morphs of *Linum tenue* (Linaceae)

3.0.1 Preamble

The following chapter is composed as a traditional thesis chapter. The methods used to assemble the reference transcriptome of *Linum tenue* are detailed in Chapter 4.0. I would like to highlight here for the reader that the terms 'contig' and 'feature' are at times used interchangeably. In genomics, features are usually genes or transcripts present on a particular stretch of sequence in a genome reference FASTA file, and are defined by their start and stop coordinates on that sequence. These sequences are known as contigs, and, depending on the resolution of the reference genome, can represent whole chromosomes or fragmented chromosomes (in the form of scaffolds). The use of ancillary annotation files are required in various formats for the downstream analyses that are discussed in the following chapter, even though we do not have the genomic information to create them. Given that our reference sequence has been constructed as a *de novo* transcriptome comprised of consensus unigene sequences (Chapter 4.0), each contig largely resembles a transcript and thus the entire contig can be considered a feature in and of itself. In-house annotation models have been created reflecting this structure. Therefore, at points in the following chapter where I discuss elements of expression, I have appropriately done so in terms of *features*. However, when discussing specific sequences in the *L. tenue* reference transcriptome, I have done so in terms of *contigs*.

3.0.2 Abstract

Linum tenue (Linaceae) displays heterostyly, a particular angiosperm mating system trait where reciprocal polymorphisms in stigma and anther positioning produce populations with two floral morphotypes: pin individuals where stigmas are positioned above the

anthers, and thrum individuals where anthers are positioned above the stigmas. This functions as an elegant structural mechanism to promote outcrossing through segregated deposition of morph-specific pollen on opposing ends of the insect pollinator(s), which is associated with the molecular mechanisms of self-incompatibility. Recent studies of heterostyly in *Primula* (Primulaceae) have made huge advances in our understanding of the underlying genetics, finding candidate genomic regions for the *S* locus. These studies reveal hemizygous determination of the thrum morph in *Primula* arising from duplication events at the *S* locus. Foreshadowing work in *Linum grandiflorum* found evidence for hemizygous expression of thrum-specific alleles, though the genomic architecture of the *Linum S* locus, and the function of its constituent alleles, remains an area of ongoing research. Here, we present a mRNAseq analysis of *L. tenue* pin and thrum floral transcriptomes and an examination of genes of interest from *L. grandiflorum*. We find strong evidence for the presence in our transcriptome reference of a thrum flower-specific candidate for the *G* locus, the style length and female incompatibility type determinant feature of the *L. tenue S* locus. Our findings provide a missing link between *THRUM STYLE-SPECIFIC1* (*TSS1*), a *S* locus candidate in *L. grandiflorum*, and *VASCULAR-RELATED UNKNOWN PROTEIN1* (*VUP1*), a gene from a small, obscure family of highly-divergent homologs in *Arabidopsis thaliana*.

3.1 Introduction

High-throughput cDNA sequencing (RNAseq) allows genome-wide analysis of gene expression to be carried out in a single experiment (Mortazavi *et al.*, 2008), and even smaller scale-experiments with a few samples can produce vast amounts of data. The depth of sequencing provides data than can be used to combine both transcript discovery and quantification, and it is little surprise that RNAseq has eclipsed former transcriptomic

analysis techniques, such as microarrays, across the biological sciences (Conesa *et al.*, 2016).

The depth of sequence coverage that Illumina methods provide is highly appealing, but, by its nature, dealing with the large volumes of fragmented short-read information comes with significant challenges and caveats. These short reads are generally 100-150 bp in length, though the chemistry is constantly improving and more modern machines can generate reads of up to 300 bp (https://emea.illumina.com/systems/sequencing-platforms.html); in our dataset a majority of the reads ranged from 100-125 bp. Assuming the availability of a high quality reference, in the form of either a genome or a *de novo* assembled transcriptome (Chapter 4.0), matching these reads to their parent transcripts is a major issue inherent when dealing with short-read data; particularly when attempting to allocate reads between alternate-splice variants with shared exon sequences, or between transcripts of closely-related genes. There are various stages in the experimental and data analysis pipelines where these issues can be addressed.

Experimental designs should aim to observe multiple biological replicates per treatment; gene expression is noisy, particularly in diverse natural populations, and the inclusion of even a few extra replicates is valuable in terms of increasing the power of the experiment (Conesa *et al.*, 2016). The experimental design should also take the expected read depth for each sample into account: here, the investigator should consider the size of the genome and the expected fold change expression differences for the genes of interest; both of these factors will affect the power of the experiment. The power of the experiment to avoid type II (false negative) errors is largely determined by the number of observations that can be made for a given transcript: the larger the genome, the lower the relative read depth will be for a given locus (assuming an absence of RNA composition bias); similarly, for very

small differences in expression (<2 fold change), the ability to detect significant changes decreases rapidly for samples with read totals below 10 million (Conesa *et al.*, 2016).

Once sequencing has been conducted, the next stage of the analysis is to align the reads to a reference in order to determine the true origin of each read and acquire the information to calculate transcript abundance. This is one of the most crucial stages of the analysis and is where a majority of the data can be lost if performed incorrectly. If aligning the reads back to a genomic reference, where reads may span splice junctions, a 'splice-aware' alignment tool must be used (Martin and Wang, 2011). Other important software parameters that should be adjusted are the strandedness options to match the library preparation of the samples, whether reads are single- or paired-end, the size of the sequenced fragment (known as the 'insert' between the Illumina adapters), and the threshold for mismatches between the reads and the reference. This last parameter is of particular relevance to our *L. tenue* dataset as the reference is comprised of consensus unigene sequences, and the read data is taken from wild populations of outcrossing (and likely highly heterozygous) individuals. When aligning reads to a reference that is genetically divergent, parameters should be changed to allow higher mismatching rates and have lower gap penalties. Many mapping tools are unable to make combined use of single- and paired-end data together and single-ended read libraries often need to be dropped from paired-end datasets. However, the number of single-ended reads is usually a negligible fraction of the paired-end read total for that library. One major cause of uncertainty in the alignment process is presence of reads that map to multiple locations on the reference. Assuming pains have been taken to generate as high quality a reference as possible, the main ways to minimise multi-mapping at this stage is to improve the accuracy of the alignment by appropriate employment of the mapping parameters outlined above. There can be a trade-off when changing the parameters of the alignment to be more accommodating to divergences between the reads and the reference: allowing

for more mismatches and lower gap penalties may increase the overall mapping rate at the cost of an increase in reads mapping to multiple locations. It is important to find an acceptable balance between the two through experimentation with the alignment parameters. The user then has the choice of either dropping multi-mapped reads from the dataset, selecting the single locus with the best alignment result, or using downstream tools that can estimate abundance that includes multi-mapped data (Warden, Yuan and Wu, 2013).

Normalisation of the raw read counts to control for transcript length, total library size, and RNA composition bias, is the next important step that precedes cross-library comparisons of gene expression. As with many aspects of RNAseq, there are a multitude of approaches and tools available, each with their own idiosyncrasies, that can impact the estimation of differential expression. Below is an outline of some of the most prominent approaches, some of which can be used in tandem according to the experimental design and the intrinsic qualities of the data. Early RNAseq studies used reads per kilobase of exon model per million reads (RPKM) (Mortazavi *et al.*, 2008), or its paired-end equivalent fragments per kilobase of exon model per million reads mapped (FPKM), normalisation methods, which are used as within-sample scaling factors to control for feature length. However, correcting for feature length is only necessary when ranking levels of expression within samples, and has been demonstrated to introduce heavy biases when used for cross-sample comparisons (Wagner, Kin and Lynch, 2012). The modified method of transcripts per million (TPM) was suggested to make samples more comparable, though is still subject to some bias when comparing samples from different tissues or different experiments (Conesa *et al.*, 2016). Examples of other methods in this suite of normalisation approaches that focus on distribution adjustments of the read counts are upper quartile and quantile normalisation (Bullard *et al.*, 2010). Upper quartile normalisation first removes all features from the dataset with universal counts of 0 and

ranks the feature counts within each sample and uses the 75th percentile value as a divisor for all feature counts within that sample. This can be performed by itself or following RPKM normalisation. The resulting normalised counts, that tend to be very small, can then be optionally scaled up by multiplying each value by the mean of the 75th percentile values for all samples. In a similar vein, quantile normalisation ranks the feature counts for each sample and calculates the mean value for each feature in a given quantile (rank position). The quantile-normalised value is thus the mean of all features in that quantile. This method functions to completely equalise quantiles across samples while maintaining the original feature order, but is most appropriate when variation in global properties are due to technical inconsistencies (e.g. different sequencing machines) and unrelated to the underlying biology (Hicks and Irizarry, 2014). In this way, distribution-adjustment normalisation approaches rely on assumptions of comparable expression distributions and RNA composition across samples, making them less appropriate when library sizes vary dramatically, or when experimental designs comprise different tissues.

Another branch of normalisation methods use library size as a scaling factor for each sample; the most popular examples are trimmed mean of *M*-values (TMM) (Robinson and Oshlack, 2010) and DESeq2 (Anders and Huber, 2010; Love, Huber and Anders, 2014). Both methods operate on the assumption that most genes are not differentially expressed, and are used in the edgeR and DESeq2 Bioconductor packages for differential expression respectively. These methods are most suitable for use on our dataset, as they adjust for both sequencing depth (differences in library size) and library composition (different tissue transcriptomes) together.

For TMM, features with read counts of 0 in all samples are removed from the dataset and each sample is scaled by its total read count; i.e. ratio between the count value for each feature and the total count value for that sample is calculated. The sample with a 75th

quantile closest to the mean of the 75th quantile across all samples is then selected as the 'reference sample'. Within each sample, a set of unbiased features are used to determine the scaling factors for that library; the ramification here is that different features are used for each sample to derive their scaling factors. Biased features are filtered out sample-by-sample through the removal of features above a defined threshold of log2 fold change differences between that sample and the reference sample. When the features to be used for the scaling factor are determined, weighted averages for their log2 ratios (reference feature count divided by sample feature count) are calculated; this weighted average ensures small changes in features with low read counts, which can result in large log2 fold changes, don't skew the data. These weighted averages are then raised to the power 2 to give the scaling factors for each library that are finally centred around 1, which does not change the results of differential expression but gives the data more 'agreeable' mathematical properties, according to Robinson & Oshlack (2010). Normalised reads are obtained by dividing the feature counts in each sample these 1-centered scaling factors.

Libraries with large differences in library size have been demonstrated to be less skewed by outlier samples when normalised using DESeq2 (Dillies *et al.*, 2013), whereby the geometric mean is calculated for each feature across all samples and used as a divisor for each sample count of this feature. The size factor for each sample is then determined by the median of these sample-feature-count to feature-geometric-mean ratios within each sample (library). This is a further method for avoiding extreme expression differences in features from skewing the size factors, and in practise tends to give more influence to 'house-keeping' genes with moderate expression differences between tissues and individuals. DESeq2 then employs shrinkage estimation, using dispersion values for each feature across the replicates estimated through Bayesian approaches. Normalised expression counts for each feature are plotted against their variance and modelled to a negative binomial distribution using generalised linear models. The theory here is that

variance in expression is highest the more lowly expressed a feature is, but decreases with increasing expression. Significant differences in count data are then determined using the Wald test.

Once differential expression analyses have been conducted, adjustment of the *p*-values for Type I error false discovery rate (FDR) is a key step in the process. The nature of genomic experiments involves multiple testing over many features, ~120,000 features in our dataset for *L. tenue*. Correspondingly, at a *p*-value threshold of 0.05, if all features were differentially expressed, ~6000 false positives are expected in the differentially expressed data. The *p*-value alone is therefore not sufficient when conducting differential expression; it is necessary to adjust for FDR. This is usually done using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) that overlies the expected *p*-value distributions from true positive differentially expressed features, which will form a 0-skewed distribution, and non-differentially expressed features, which will form a uniform distribution across *p*-values. The mathematics behind the adjustment is relatively simple: *p*-values are ordered and ranked, then for a specific feature (*f*), the *p*-value for *f* is multiplied by the quotient of the highest rank position in the dataset and the rank position for *f*. This limits the number of false positives that are reported as significant by effectively shifting up the *p*-values to 'correct' the distribution to a 0-skew.

In recent years, huge advances have been made in our understanding of the genetics of heterostyly. Intensive genomic studies in *Primula* (Nowak *et al.*, 2015; Li *et al.*, 2016; Cocker *et al.*, 2018) have successfully characterised the architecture of a 450 kb region, where duplication events in the thrum alleles appear to assert hemizygous determination of the short-styled phenotype. This is consistent with the commonly accepted model of a dominant *GPA/gpa* thrum genotype, and a recessive *gpa/gpa* pin genotype: where the *G* locus (from the German 'griffel': *'style'*) determines the style length and female

incompatibility type, the *P* locus determines the pollen morphology and male incompatibility type, and the *A* locus determines the positioning of the anthers. Together, these loci comprise a superlocus known as the *S* locus.

Work in *Linum grandiflorum* (Ushijima *et al.*, 2012) has identified five genes, *TSS1*, *TPP1*, *LgMYB21*, *LgAP1* and *LgGLX1*, with thrum-specific expression. *TSS1* is of particular interest as it is expressed in the tissues of the style, and has since been found to be absent from pin genomic DNA (Ushijima *et al.*, 2015), suggesting a hemizygous genotypic determinant of thrum individuals in *L. grandiflorum*, similar to what is observed in *Primula*.

Here we present an analysis of expression differences between floral transcriptomes of pin and thrum individuals of *L. tenue*. As outlined in Section 4.2, the experiment was designed to capture as broad a range of developmental stages as possible, whilst also providing a vegetative control growth stage. Young bud, immature flower, open flower and leaf growth stages were sampled from three pin individuals and three thrum individuals, providing a total of 24 samples. These samples were sequenced using an Illumina HiSeq 2500 over two lanes, to provide technical replicates; and given the expected number of generated reads per lane, we aimed to sequence to a depth of 20 million reads per sample. We will pay particular attention to *TSS1*, and demonstrate the homolog (putative ortholog) in *L. tenue* to be a strong candidate for the *G* locus.

3.2 Materials and Methods

3.2.1 Comparison of mapping tools and parameters

To demonstrate the behaviour of different read mapping tools, and to inform the alignment settings to be used for transcript quantification, the STAR v2.3.5a (Dobin *et al.*,

2013a) and TopHat2 v2.1.0 (Kim *et al.*, 2013) aligners were used to align the reads for each library individually to the *L. tenue* transcriptome reference using three preset groups of customized options: 'loose', 'default', and 'stringent'. For STAR: loose parameters were defined as 'outFilterMismatchNmax 15', 'scoreDelOpen -1', 'scoreDelBase -1', 'scoreInsOpen -1', and 'scoreInsBase -1'; and stringent parameters were defined as 'outFilterMismatchNmax 5', 'scoreDelOpen -3', 'scoreDelBase -3', 'scoreInsOpen -3', and 'scoreInsBase -3'. For TopHat2: loose parameters were defined as 'bt2-D 20', 'bt2-R 3', 'bt2-N 1', 'bt2-L 20', 'bt2-rdg 4,2', and 'bt2-rfg 4,2'; and stringent parameters were defined as 'bt2-D 10', 'bt2-R 1', 'bt2-N 0', 'bt2-L 24', 'bt2-rdg 6,4', and 'bt2-rfg 6,4'. The appropriate strandedness options were used for both aligners in all presettings: for STAR this was 'fr-firststrand'; and for TopHat2 this was also 'fr-firststrand'.

3.2.2 Subsetting of the libraries

In order to control for large variations in library size, the 18 libraries over 5,978,885 reads were subsetted to this number by randomly selecting reads from the FASTQ files, along with their reciprocal paired mate. The same methods of mapping, quantification, differential expression and downstream analyses were applied as to the rest of the dataset.

3.2.3 Mapping and quantification of *L. tenue* libraries

For transcript quantification, paired reads were aligned to the *L. tenue* transcriptome reference sequence, produced in Chapter 4.0, with STAR v2.3.5a using the loose parameter presetting, in order to allow mismatches between reads and multi-allelic consensus sequences in the reference. Sample expression was quantified according to an *L. tenue* annotated transcriptome model, a custom-made Browser Extensible Data (BED)

file, using Partek E/M, from the Partek Flow software suite (Partek Inc., 2018), using strict paired end compatibility and a minimum read count of 10 across all samples. Partek E/M employs an expectation-maximum algorithm to allocate multi-mapped reads among their matching loci using a maximum likelihood approach.

3.2.4 Differential expression analysis

Differential expression was performed using DESeq2 (Love, Huber and Anders, 2014). For analysis of differential expression, the biological triplicate sets for all three floral growth stages were compared by morphotype. The list of differentially expressed genes were filtered for FDR, by using a FDR step-up threshold of 0.05 and by excluding features that showed less than a two-fold change in expression (i.e. a fold change between -2 and 2). In order to determine the features most likely to be associated with flower development in the full and subsetted datasts, the feature list was also filtered to create a list of features exclusively expressed in the floral growth stages using in-house python scripts.

3.2.5 Exploratory analyses

In order to understand relationships between the samples beyond differential expression, a number of approaches were taken to observe other interactions in expression patterns. These analyses were conducted using the tools in the Partek Flow software suite. A principal components analysis (PCA) was performed on the count data prior to and post differential expression analysis, using a log2 transformation and allowing features to contribute by variance. A t-distributed stochastic neighbour embedding (t-SNE) analysis was performed on the data with the following parameters: perplexity 30, a measure of the effective number of neighbours for each point on the graph; random generator seed 1;

initialise output values at random; number of iterations 1000; distance metric between data points Euclidian; features contribute to the graph equally; count values transformed by log2. Hierarchical clustering was also performed on the dataset using Partek Flow with default parameters to cluster samples according to overall similarity of expression.

3.2.6 Genomic clustering

Genomic clustering analysis was performed on the filtered list of differentially expressed features from the subsetted dataset in the hope of detecting signals of an *S* supergene locus of linked genes. The differentially expressed features (contigs) were extracted from the *L. tenue* transcriptome sequences and their putative corresponding loci were found in the genomes of *Arabidopsis thaliana* (TAIR10), *Populus trichocarpa* and *Primula veris* through local TBLASTX searches, with default parameters including the selection of only the top BLAST hit. *A. thaliana* was selected as a reference due its level of development as a model and high characterisation of the *S* locus, *P. trichocarpa* was selected as it was the closest reference species to *L. tenue* with a chromosomal-level resolution of the genome at the time of writing, and *P. veris* was selected as it was the best developed heterostyly model species with candidate regions for the *S* locus at the time of writing. The sequences used for *A. thaliana* and *P. trichocarpa* were at the chromosomal level, and the *P. veris* sequences used were the best-resolved scaffold set. A subset of scaffolds from the *P. veris* genome representing candidate contigs putatively linked to the *S* locus (Nowak *et al.*, 2015) was also searched for matches to the differentially expressed *L. tenue* contigs. The genome of *Linum usitatissimum* was not used during this exercise as the length of assembled scaffolds (at the time of writing) were not long enough to display the data in the desired manner. The build of the *P. veris* genome was used despite this due to the characterisation of its *S* locus region.

BLAST results were then organised by subject (genome reference sequence) chromosome/scaffold, and the start positions of each alignment in the subject were plotted to visualise their clustering patterns on chromosomes of these reference species.

3.2.7 *L. grandiflorum S* locus candidates in *L. tenue*

Amino acid sequences of the five genes (TSS1, LgAP1, LgMYB21, TPP1 and LgGLX1) found to display thrum-specific expression in *L. grandiflorum* by Ushijima et al. (2012) were downloaded from the DDJ/EMBL/GenBank (http://getentry.ddbj.nig.ac.jp/) using accessions AB617824-AB617828 respectively. Homologs were searched for in the *L. tenue* reference transcriptome using TBLASTN with default parameters, limiting the maximum number of target sequences to 10. The DESeq2 normalised count data for 43 resultant matching *L. tenue* contigs were extracted, and each of the contigs were searched for in the list of *L. tenue* differentially expressed features (Section 3.3.3).

The cDNA sequence of the single *L. tenue* transcript contig, Contig_141165 (a homolog, and possible ortholog, of *L. grandiflorum* TSS1), found to match to one of the *L. grandiflorum* genes of interest *and* found to be differentially expressed between *L. tenue* pin and thrum morphs was translated *in silico* into amino acid sequences with all six possible reading frames using the online ExPASy Translate Tool (Appel, Bairoch and Hochstrasser, 1994; Gasteiger *et al.*, 2003; Artimo *et al.*, 2012) (https://web.expasy.org/translate/). To select the correct translational reading frame for Contig_141165, the amino acid sequence for TSS1 was used to query the six theoretical Contig_141165 amino acid sequences using BLASTP. To confirm selection of the correct reading frame, the six theoretical Contig_141165 amino acid sequences were also aligned to TSS1 alongside those of *A.*

thaliana homologs VUP1, VUP2, VUP3 and VUP4 (Grienenberger and Douglas, 2014), using MUSCLE v3.8.31 (Edgar, 2004) with default settings; alignments were viewed in Jalview v2.9 (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009) and a UPGMA tree of average distance was calculated using percentage sequence similarity.

The amino acid sequence for Contig_141165, translated in the sense (5' to 3') orientation using reading frame 3, was aligned to VUP1, TSS1 and a range of *VUP1* homologs across the angiosperm phylogeny including *Sellaginella moellendorffii* (Embryophyta) using MUSCLE v3.8.31 with default parameters, visualised using Jalview v2.9. Through TBLASTN of *VUP1-4* and *VUP1* homologs to the *L. tenue* reference transcripts, three contigs, Contig_051339 (matching to *VUP1*, *VUP3*, *VUP4*, Populus_trichocarpa1, Populus_trichocarpa2 and Populus_trichocarpa3), Contig_051340 (matching to Populus_trichocarpa1 and Populus_trichocarpa2) and Contig_051341 (matching to Populus_trichocarpa1 and Populus_trichocarpa2), representing putative *VUP1* homologs (and thus putative paralogs of *L. tenue* Contig_141165) were discovered. For these *L. tenue* contigs, the predicted amino acid sequences from the correct reading frames (using the same techniques as outlined for Contig_141165) were included in the alignment of VUP1 homologs.

Phylogenetic analyses were conducted on the alignment using IQ-TREE v1.6.9 (Nguyen *et al.*, 2015). Both Bayesian and maximum likelihood inference analyses were performed and, using the Bayesian information criterion (BIC) and the Akaike information criterion (AIC), the best-fit substitution model was selected with ModelFinder (Kalyaanamoorthy *et al.*, 2017). Bootstrap values were generated with the VT+G4 model of substitution for 1000 ultrafast bootstraps (Minh, Nguyen and von Haeseler, 2013; Hoang *et al.*, 2018). The phylogenetic analysis was repeated using an alignment excluding Contig_051339, Contig_051340 and Contig_051341, the putative *VUP1* homologs of *L. tenue*.

3.3 Results

3.3.1 Comparison of mapping tools and parameters

The STAR aligner generated the highest mapping results overall when compared to TopHat2 (Table 3.1). STAR generated an overall alignment rate of 82.32%, 81.14% and 78.39%, whereas TopHat2 produced overall alignment rates of 70.22%, 70.12% and 69.75%, for the loose, default and strict parameter presettings respectively. Also of note is the significant improvement in unique mapping rates, which are consistently over 20% higher with STAR than with TopHat2. There was little variation between the mapping results for the individual libraries at each presetting, though with increasing stringency, this variation increases very slightly as demonstrated by coefficients of variation of 0.023, 0.024 and 0.025 for STAR and 0.074, 0.075 and 0.077 for TopHat2, for the overall mapping rates at loose, default and stringent presettings respectively. These coefficients of variation also demonstrate a comparable, but slight increase in, variation with the TopHat2 aligner, with STAR giving more consistent results overall.

Whilst not used as a comparative measure, it should also be noted through personal observations that STAR consistently ran much faster than TopHat2, sometimes by a difference of 5 hours.

Table 3.1 Mapping results for the full *L. tenue* data set with the STAR and TopHat2 aligners. Each library was mapped individually, and averages and standard deviations are shown for each alignment presetting, as described in Section 3.2.1.

	STAR loose	STAR default	STAR stringent	TopHat2 loose	TopHat2 default	TopHat2 stringent
Total aligned reads (%) ± SD	82.32 ± 1.92	81.14 ± 1.95	78.39 ± 1.95	70.22 ± 5.22	70.12 ± 5.24	69.75 ± 5.34
Total unique paired alignments (%) ± SD	67.17 ± 1.13	66.09 ± 1.16	63.51 ± 1.27	42.06 ± 2.51	41.92 ± 2.52	41.28 ± 2.66
Total non-unique paired alignments (%) ± SD	15.15 ± 1.09	15.05 ± 1.08	14.88 ± 1.07	11.30 ± 1.39	11.31 ± 1.41	11.17 ± 1.42
Total unaligned reads (%) ± SD	17.68 ± 1.92	18.86 ± 1.95	21.61 ± 1.95	29.78 ± 5.22	29.88 ± 5.24	30.25 ± 5.34

Table 3.2 Read counts and number of alignments for each RNAseq library, ranked in size order. The morph and growth stages have been given as abbreviations for each sample: pin (_p), thrum (_t), leaf (L), young bud (YB), immature flower (IF) and open flower (OF). Alignments were performed using STAR with the loose parameter setting.

Sample individual and name	Growth stage/morph	Total reads	Total alignments (paired-end) to the *L. tenue* reference
CBT8_OpenFlower	OF_t	810,556	1,558,002
CBT6b_Leaf	L_p	1,652,967	3,297,533
CBT7b_YoungBud	YB_t	2,136,112	4,093,598
GRT9c_Leaf	L_p	2,184,258	4,580,721
GRT1a_YoungBud	YB_p	3,585,868	6,917,477
GRT9c_YoungBud	YB_p	5,817,359	11,454,051
CBT8_YoungBud	YB_t	6,463,463	12,493,752
GRT9c_OpenFlower	OF_p	7,892,092	16,931,168
CBT10_ImmatureFlower	IF_t	9,427,315	19,133,866
CBT10_YoungBud	YB_t	9,926,960	20,159,619
CBT8_ImmatureFlower	IF_t	10,191,126	19,796,823
CBT10_OpenFlower	OF_t	10,994,604	22,569,737
GRT1a_ImmatureFlower	IF_p	11,559,619	23,758,819
CBT7b_Leaf	L_t	11,578,817	24,409,763
GRT1a_Leaf	L_p	13,115,010	26,906,364
CBT6b_OpenFlower	OF_p	13,725,994	29,313,366
GRT1a_OpenFlower	OF_p	20,535,561	42,099,893
CBT6b_ImmatureFlower	IF_p	22,318,996	47,060,478
CBT7b_OpenFlower	OF_t	22,408,546	47,008,698
CBT10_Leaf	L_t	24,394,478	50,505,081
GRT9c_ImmatureFlower	IF_p	26,071,618	53,914,515
CBT8_Leaf	L_t	43,399,923	91,306,528
CBT6b_YoungBud	YB_p	47,158,940	98,519,708
CBT7b_ImmatureFlower	IF_t	74,227,468	154,852,114

3.3.2 Mapping and quantification of libraries

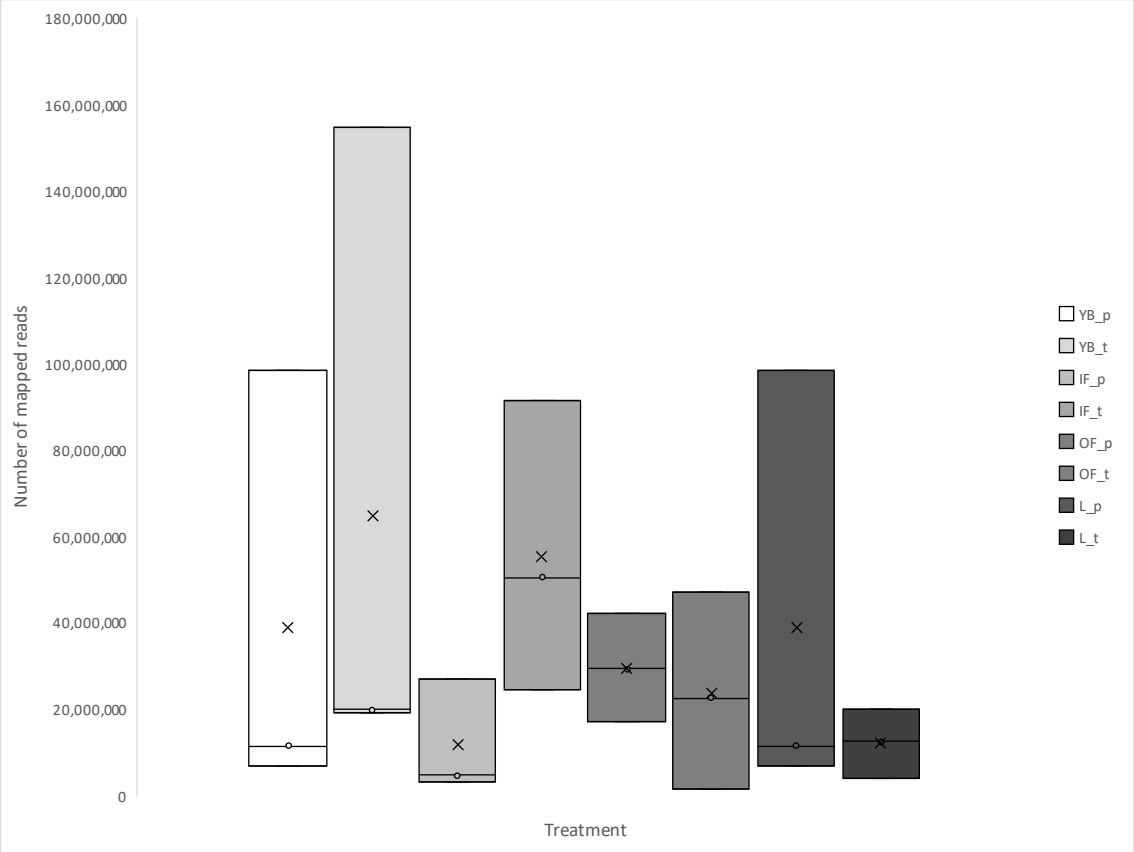


Figure 3.1 Boxplots of the alignment data for each treatment. The morph and growth stage have been given as abbreviations for each sample: pin (_p), thrum (_t), leaf (L), young bud (YB), immature flower (IF) and open flower (OF).

Total read counts were highly variable between samples due to a large range in size of RNAseq read libraries (Table 3.2). The number of aligned reads ranged from 1,558,002 in sample CBT8_OpenFlower, to 154,852,114 in sample CBT7b_ImmatureFlower. Variation in library size was on a sample-by-sample basis and was not a product of treatment; combined data distributions for each treatment triplicate are shown in Figure 3.1. This indicates the variation in library size was randomly distributed across the dataset, however the mean and median values for these raw counts by triplicate varied greatly.

The mean library size was 16,732,402 with a median of 11,277,111. Total reads mapping to pin morph individuals in floral stages was 329,969,475, and to thrum morph individuals in floral stages was 301,666,209 (Figure 3.2). The mean number of alignments

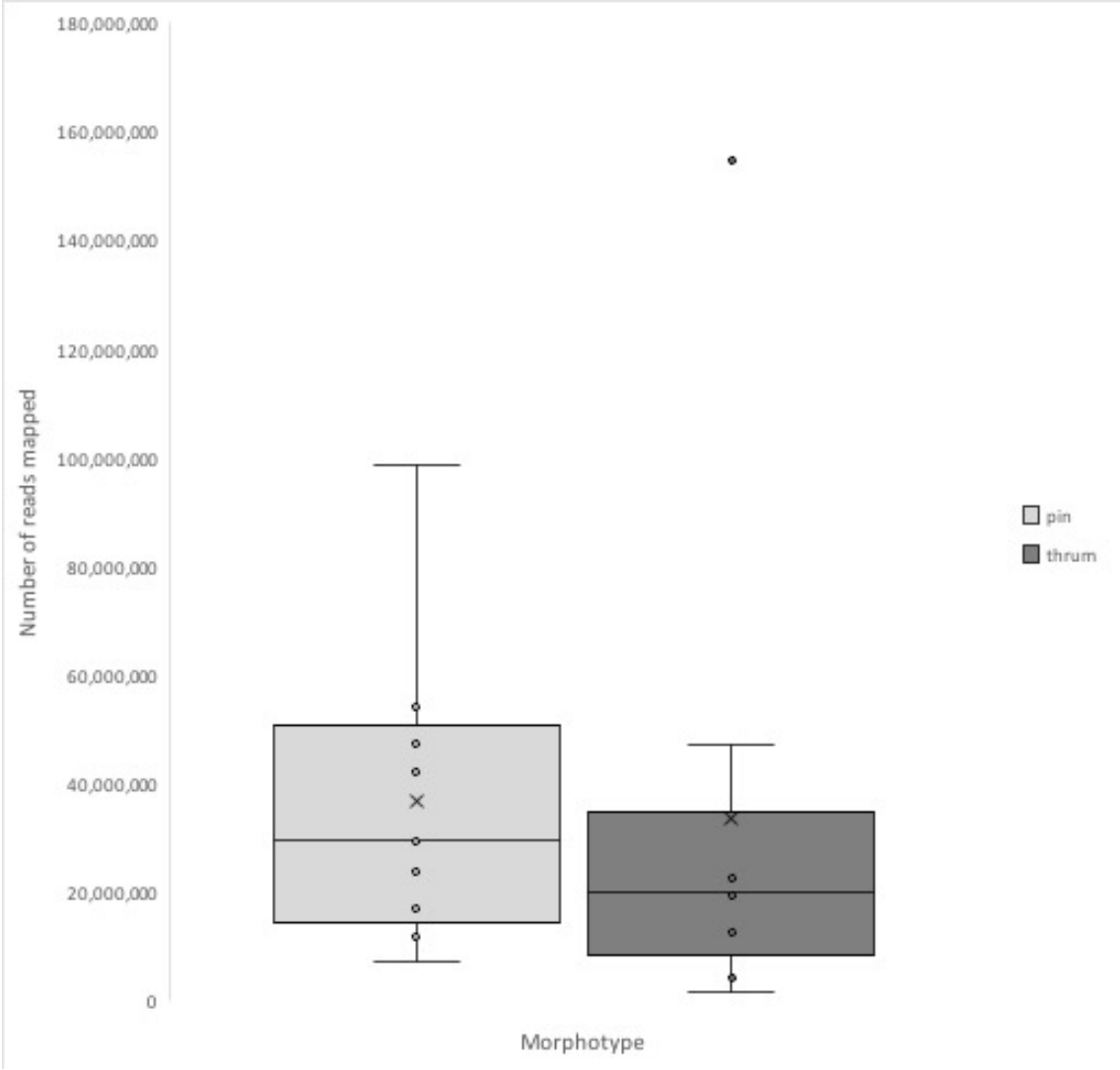


Figure 3.2 Boxplots of the alignment data for floral developmental stages of each morph. Sample CBT7b_IF has been designated as an outlier comparative to other thrum samples.

for both groups are very similar (36,663,275 pin, 35,518,468 thrum), but the pin dataset displays larger interquartile ranges and contains more of the largest libraries.

3.3.3 Differential expression analyses

All differential expression results were filtered primarily by false discovery rate (FDR) adjustment, using a threshold of 0.05, and a minimum fold change of 2. For the full dataset this yielded a total of 2363 (Appendix III), and for the subset dataset 1586, differentially expressed features between all pin and all thrum floral growth stages. The *p*-value distribution of the datasets were shown to be heavily skewed towards 0 as opposed

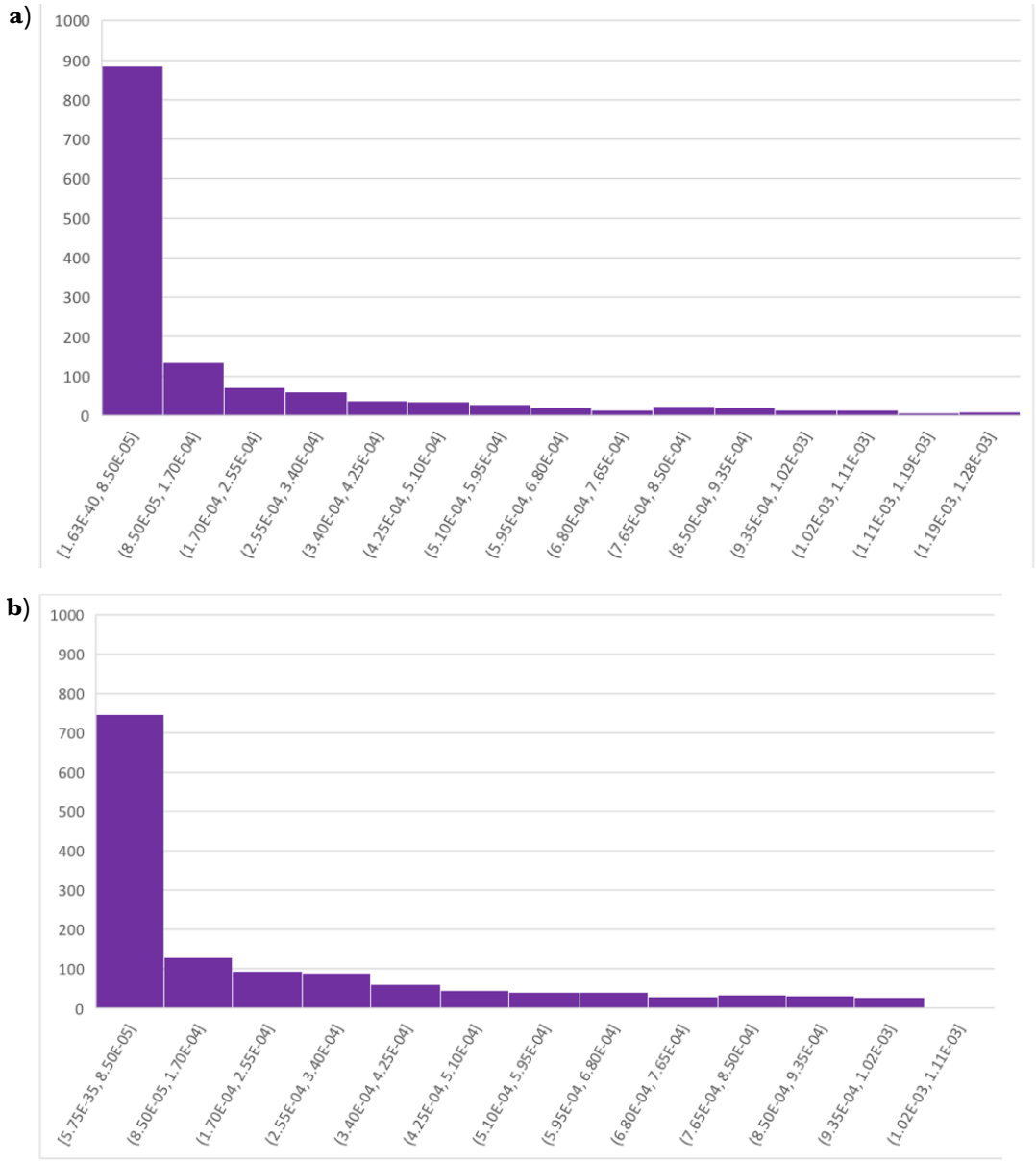


Figure 3.3 *p*-value distributions for differentially expressed features shared between the full dataset (**a**) and the subsetted dataset (**b**) after filtering for fold change and FDR-adjusted *p*-values.

to uniform from 0 to 1, as would be expected in the absence of differential expression, indicating the variance model used by DESeq2 was an appropriate fit for the data. Between both sets of differentially expressed features, 1371 features are shared, with the *p*-values tending to be lower in the full dataset (Figure 4.3), with 140 more features having a *p*-value less than or equal to 0.000085 in the full dataset.

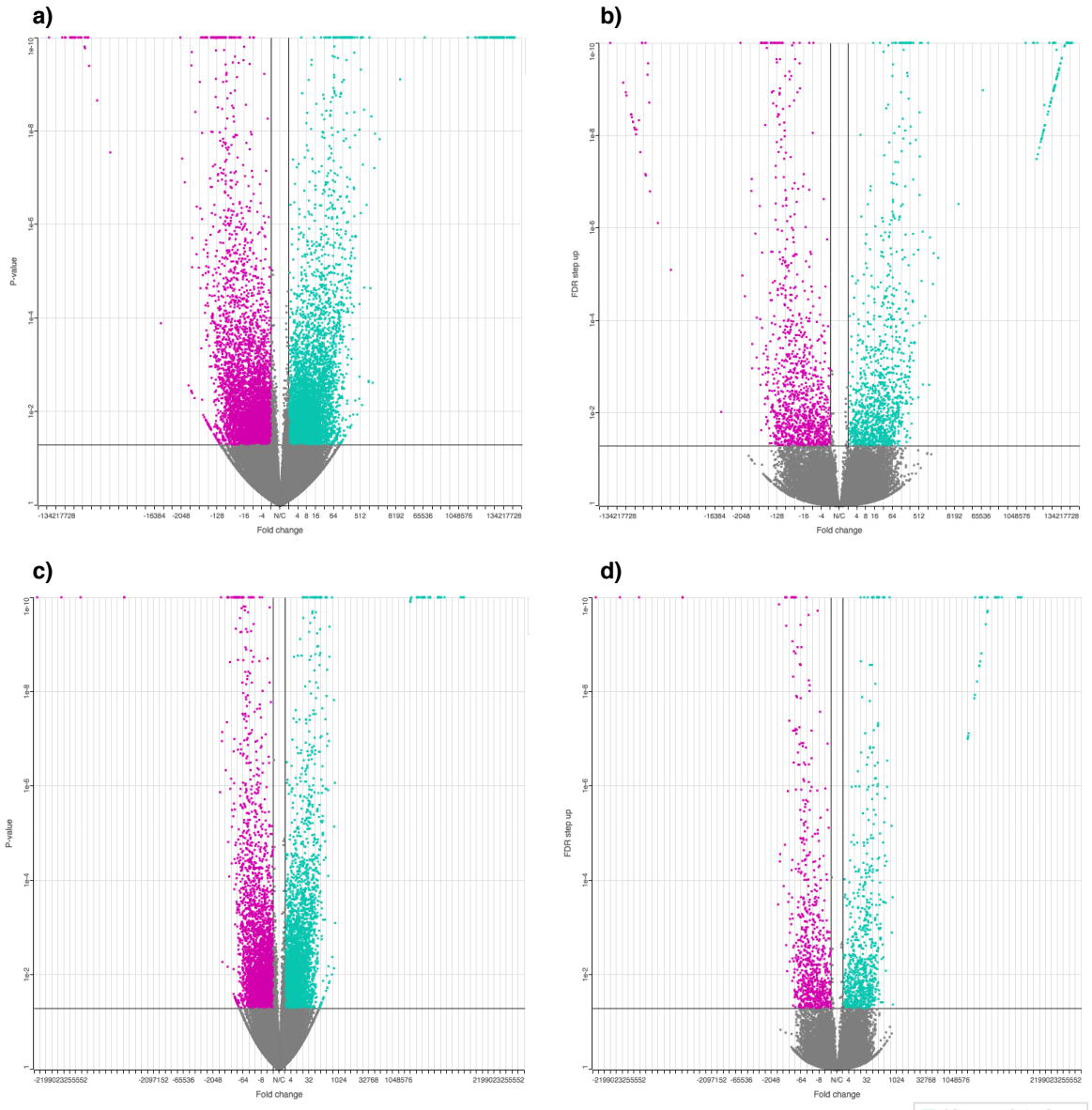


Figure 3.4 Volcano plots of *L. tenue* differential expression analyses with DESeq2. Fold change values are displayed on the x-axis and probability values on the y-axis. Differential expression results in terms of fold change are displayed for the full dataset, a) and b), and the subset dataset, c) and d), are shown against both *p* value, a) and c), and FDR adjustment, b) and d).

Figure 3.4 displays the differential expression results for the full- (3.4a and 3.4b) and subsetted (3.4c and 3.4d) datasets in volcano plots, with log2 fold change plotted against *p*-values (3.4a and 3.4c) and FDR-adjusted *p*-values (3.4b and 3.4d). Significance values of 0.05 and minimum fold change of 2 have been selected. Both sets of plots follow the expected pattern of increasing significance (lower *p*-value) as fold changes move further away from 0. The number of significantly differentially expressed features greatly decreases after FDR-adjustment. Of interest between the full- and subsetted datasets is the presence of far fewer outliers in the subsetted dataset, as demonstrated by the reduction of the 'arms' of the volcano at the extremes of the fold change scale, particularly for down-expressed features.

3.3.4 Exploratory analyses

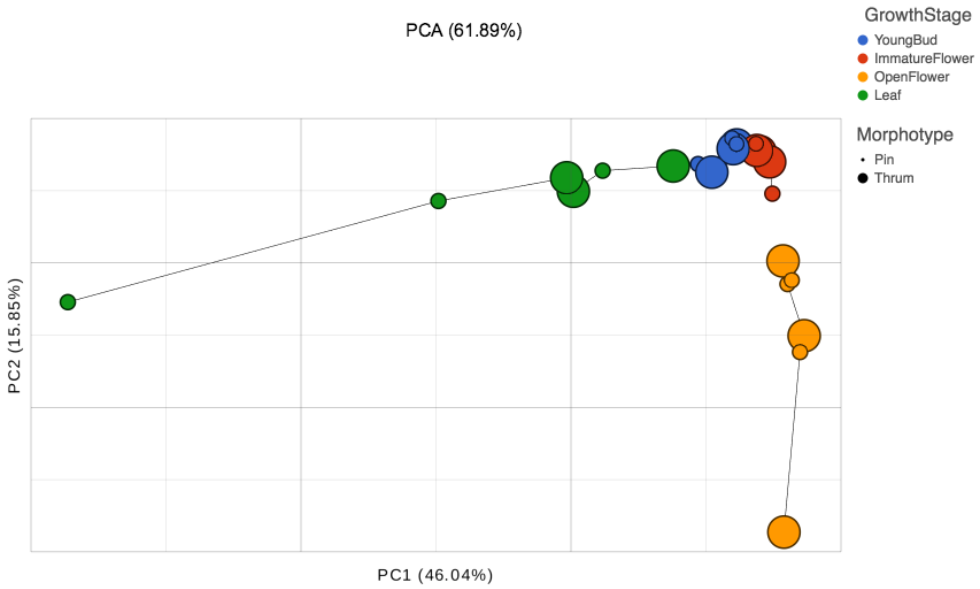


Figure 3.5 Principal components analysis of the full dataset prior to differential expression analysis. Samples have been coloured and connected by growth stage, and node sizes denote the style phenotype (morphotype). The first two principal components are displayed.

PCA revealed that 61.89% of the full dataset was explained by PC1 and PC2 (Figure 3.5), and further inclusion of PC3 explained 70.98% of the data. Figure 3.5 demonstrates sample clustering largely by developmental stage, and 3-dimensional representations of

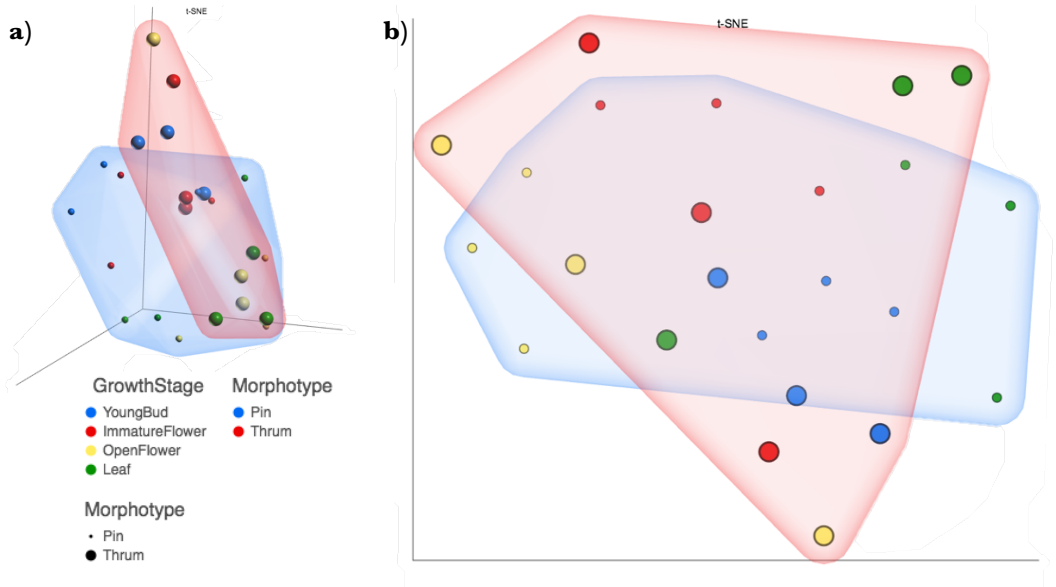


Figure 3.6 t-SNE analysis of the full dataset prior to differential expression analysis. Samples have been coloured growth stage, and the style phenotype (morphotype) is displayed by size and highlighted by colour. The first three (**a**) and the first two (**b**) principal components are displayed.

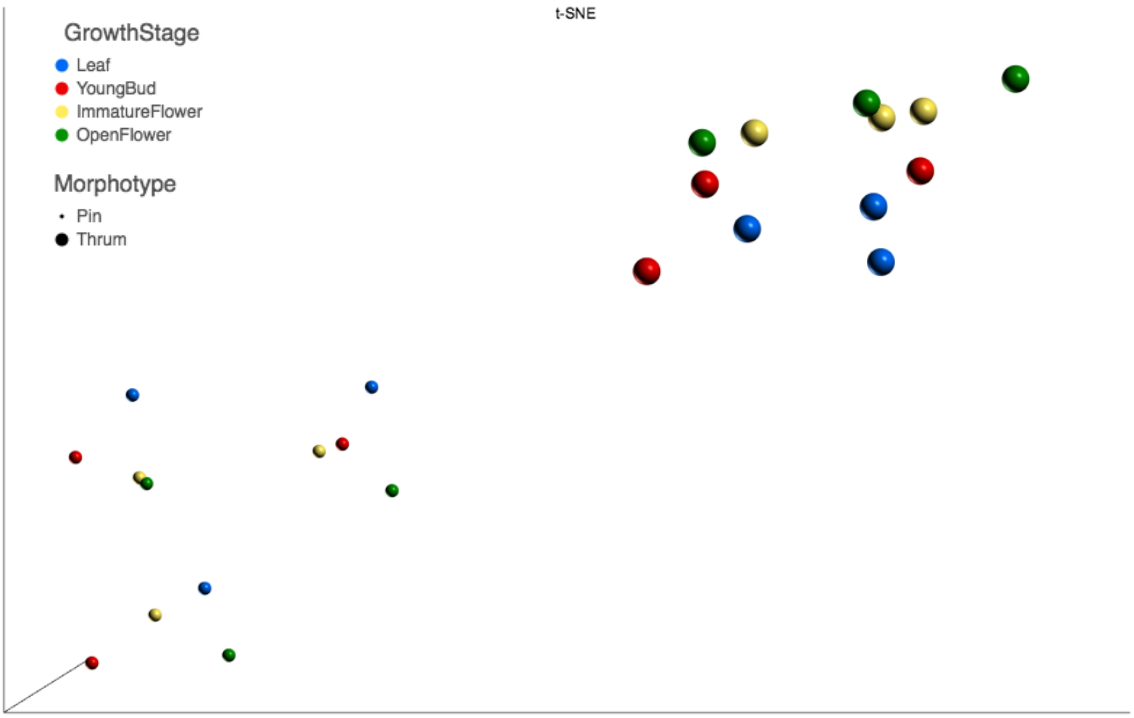


Figure 3.7 t-SNE analysis of the subset dataset after differential expression analysis. Samples have been coloured by growth stage, the style phenotype (morphotype) is displayed by size. The first three principal components are displayed.

the data with the first 3 principal components added further dimensionally to this clustering. At this stage of the analysis, no pattern by morphotype is displayed. As can be seen in Figure 3.5, the open flower and leaf clusters each show a single large outlier; these individuals corresponded to library sample CBT8_OpenFlower and CBT6b_Leaf, which were the smallest samples with 1,558,002 and 3,297,533 total aligned reads respectively.

PCA analysis conducted on the subsetted dataset prior to differential expression analysis revealed a similar clustering pattern, with PC1 and PC2 together explaining 62.40% of the data.

When conducted on the datasets prior to differential expression, t-SNE analysis revealed loose but overlapping expression patterns between developmental stages in both the full dataset (Fig. 3.6) and the subsetted dataset for the first 2 principal components. When viewed 3-dimensionally with use of the third principal component (Fig. 3.6a), some separation can be seen between pin and thrum morphotypes. However, when conducted on the datasets post differential expression analyses on the feature list filtered by FDR-adjusted *p*-value and minimum fold change (Fig. 3.7), distinct clustering by morphotype is evident. Of particular interest here is that distinct sub-clustering can be seen within the pin-morph cluster, where each sub-cluster was comprised of all the developmental stages for a single individual. Similar proximity positioning of samples by individual can be observed to an extent within the thrum cluster, but this is much less distinct.

Hierarchical clustering of both datasets yielded very similar results to the PCA and t-SNE analysis; the resultant heat map for floral growth stage samples in the subsetted dataset is shown in Fig. 3.8. As is clear from the heatmap, samples show distinct clustering by morphotype, with floral growth stages for each individual falling into exclusive pin or thrum clades. The heatmap displays four large blocks of expression patterning, two areas

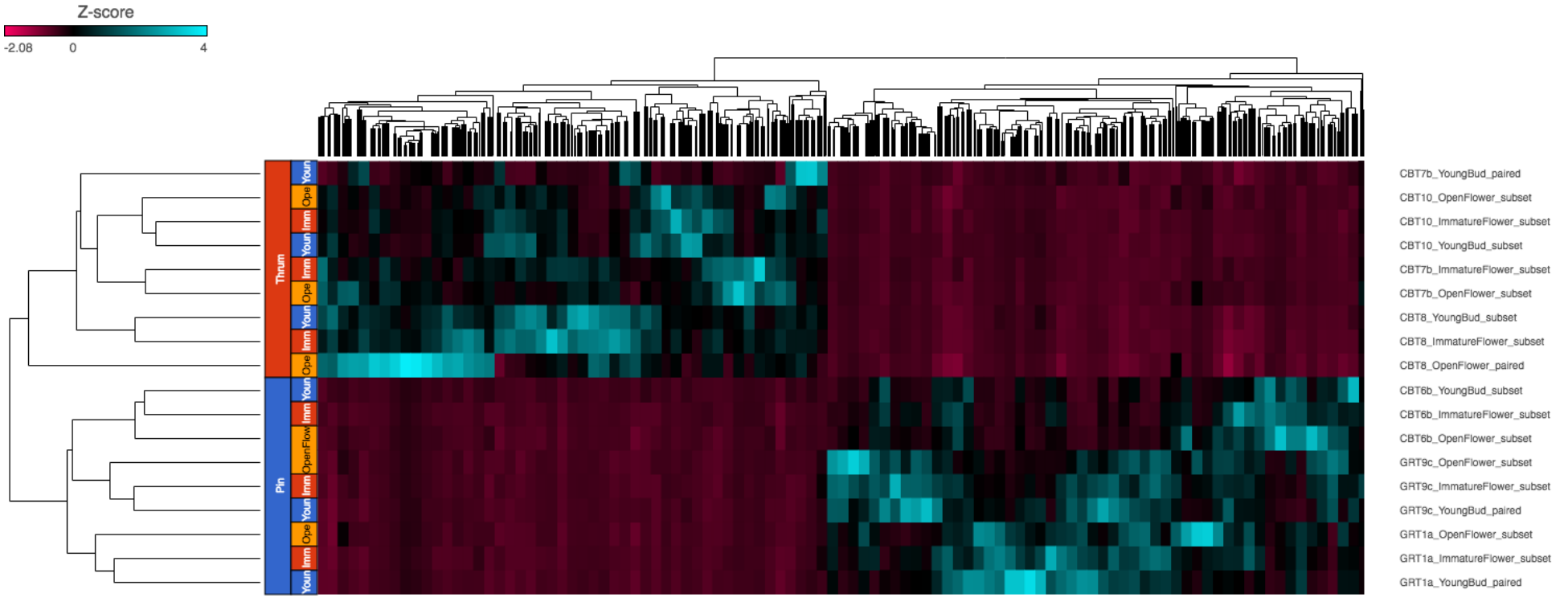


Figure 3.8 Heat map displaying the results of hierarchical clustering analysis of floral developmental stages in the subset dataset. The dendrogram for the samples is displayed on the left of the heatmap, with sample names displayed on the right. Conversely, the dendrogram of features is displayed on the top. Each cell in the heatmap represents the z-score (expression value) for a single feature (contig) within an individual.

of universal down-regulated expression within morphotypes (displayed in cherry), and two areas of up-regulated expression (sky blue) within morphotypes. Areas of up-expression are not as uniform in terms of expression patterns between individuals, and display sub-blocks of up-regulated features that largely correspond to within-individual patterns of expression.

Again of interest in the hierarchical clustering is the pattern of sample clustering by individual rather than by developmental stage. As can be seen clearly in Figure 3.8, the 9 pin morph samples are assembled into 3 clades towards the bottom of the hierarchy, with each clade consisting of all the samples for a single individual. This pattern is present in the thrums, with the samples for individual CBT10 forming a single clade, but overall the pattern is weaker, with some growth stages forming their own 'outgroups'.

3.3.5 Genomic clustering

Of the 1586 contigs from the subsetted dataset found to be differentially expressed between pin and thrum floral developmental stages after filtering for significance and fold change: 560 found matches in *A. thaliana*, 585 found matches in *P. trichocarpa*, and 510 found matches in *P. veris*. In the *P. veris* candidate *S* locus subset (comprised of 8 scaffolded contigs), 35 of the *L. tenue* contigs found BLAST hits.

In each of the *A. thaliana* and *P. trichocarpa* reference genomes, BLAST hits were distributed across the chromosomes. In *P. trichocarpa* the most frequently matched chromosome was Chromosome 1 with 67 (11.97%) hits, and the remaining Chromosomes 2-19 saw an average of 28.78 hits (SD 7.39). In *A. thaliana*, the hits were more unevenly distributed between its 5 chromosomes: Chromosomes 1, 3 and 5 saw an average of 132 (SD 19.97),

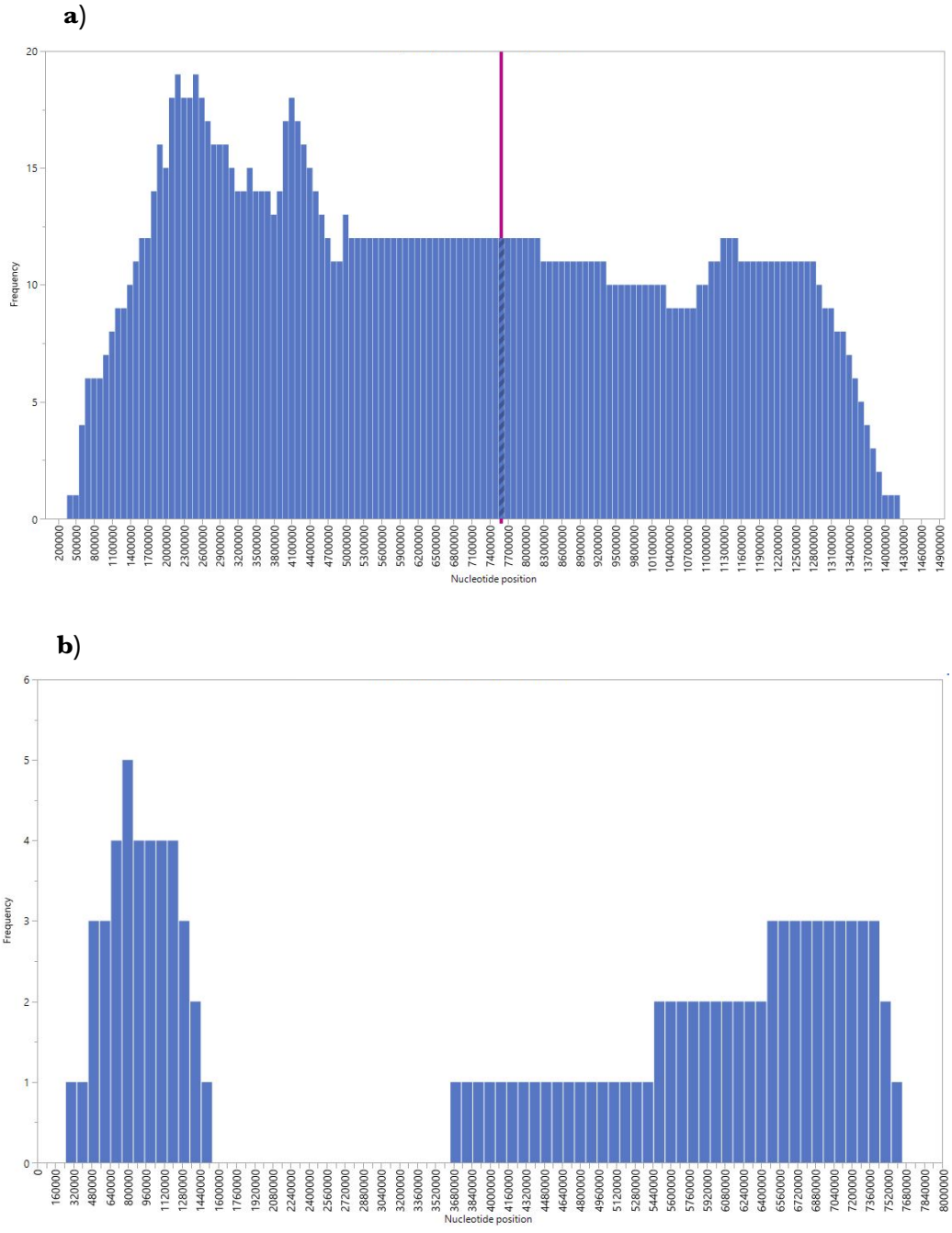
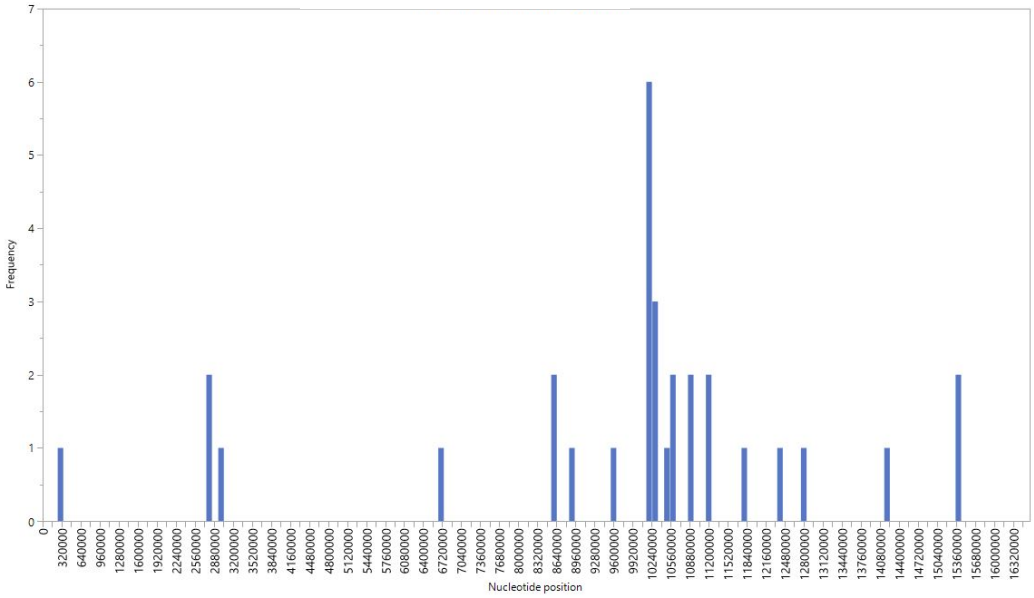


Figure 3.9 Plots of differentially expressed contigs and the start positions of their pairwise alignments to genomic reference sequences: (**a**) *A. thaliana* chromosome 3, with the genomic location of VUP1 indicted by the red marker; (**b**) *A. thaliana* chromosome 4; (**c**) *P. trichocarpa* chromosome 19; and (**d**) *P. veris* Contig927. The count frequencies are represented on the y-axes and x-axes represent starting nucleotide coordinates of the alignments on the chromosomes.

c)



d)

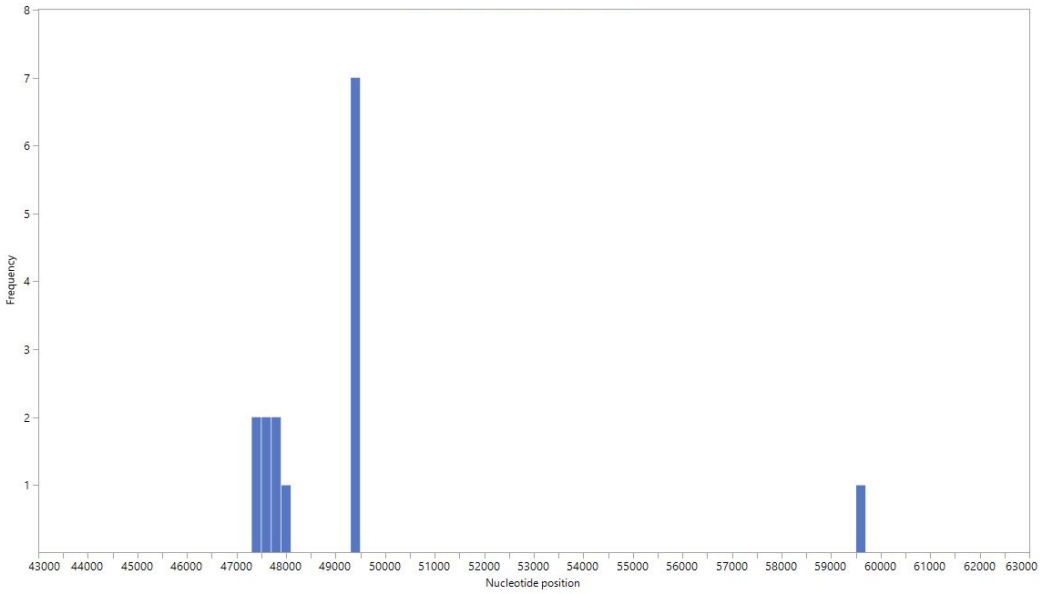


Figure 3.9 Continued.

Chromosome 2 received 84 hits and Chromosome 4 received 76 hits; regarding the organellar genomes, 3 contigs matched to the chloroplast and 1 to the mitochondrial genome. In *P. veris*, the 510 BLAST matches were found to be distributed across 330 scaffolded contigs, with 224 of these contigs containing only a single BLAST match from the *L. tenue* query. Of the 35 *L. tenue* contigs that found a match in the *S* locus reference, a majority of these were matched to JTKG01000926.1 (15 hits) and JTKG01000478.1 (10 hits).

The results of the differentially expressed contig mapping for selected chromosomes/contigs are displayed in Figure 3.9. Chromosomes were selected as they are known to contain genes linked to the *S* locus of that species (*A. thaliana* Chromosome 4, *P. trichocarpa* Chromosome 19 and *P. veris* JTKG01000926.1/Contig927) and floral development (*A. thaliana* Chromosome 3). Clustering appears particularly distinct for *A. thaliana* Chromosome 3 (Figure 3.9a) and *P. veris* Contig927 (Figure 3.9d). BLAST hits appear more sparsely distributed on *A. thaliana* Chromosome 5 (Figure 3.9b) and *P. trichocarpa* Chromosome 19 (Figure 3.9c), with a majority of genomic regions matching to only a single contig.

3.3.6 *L. grandiflorum S* locus candidates in *L. tenue*

TBLASTN of the amino acid sequences of five genes with thrum-specific expression in *L. grandiflorum* yielded the following number of matching *L. tenue* transcript sequences: TSS1, 1; LgAP1, 8; LgMYB21, 9; TPP1, 10; LgGLX1, 10. Of these 43 *L. tenue* contigs, only Contig_141165, the single match for *L. grandiflorum* TSS1, was found to be present in the DESeq2 filtered list of differentially expressed features. The normalised count data for Contig_141165 exhibits clear thrum-specific expression in *L. tenue*. With the exception of

sample CBT6b_YoungBud (count of 4.5), all pin samples displayed counts of 0.0 for feature Contig_141165. For thrum samples of individuals CBT10, CBT7b and CBT8 respectively, feature Contig_141165 displayed counts of 121.55, 15.2428 and 29.6401 for immature flower samples; 111.104, 29.6401 and 69.6678 for open flower samples; and counts of 0.0 for all leaf and young bud samples.

From the functional annotation of the *L. tenue* reference transcriptome (Section 4.2.5), a predicted peptide translated from nucleotide coordinates 102-539[+] on Contig_141165 was found to be a putative homolog of *A. thaliana* VUP1. Of the six theoretical amino acid sequences that could be derived from Contig_141165, TSS1 found a BLASTP match only with Ltenue_Contig41165_5'3'_Frame3. Further alignment of Contig_141165 theoretical amino acid sequences to TSS1, VUP1 and *A. thaliana VUP1* paralogs VUP2-4 (Appendix IV) showed Ltenue_Contig41165_5'3'_Frame3 represented the only reading frame that shared four conserved VUP motifs designated M1-M4 (Figure 3.10) by Grienenberger & Douglas (2014). Clustering of these sequences by percentage sequence similarity (Appendix V) demonstrated clear exclusion of the other *L. tenue* theoretical peptides, with Ltenue_Contig41165_5'3'_Frame3, TSS1 and VUP1-4 forming a distinct clade and other *L. tenue* theoretical peptides falling as outgroups.

Alignment of Contig_141165, TSS1, VUP1-4 and *VUP1* homologs showed overall low levels of sequence similarity among homologs (only 28% of residues were constant across all sites), with the exception of four highly conserved motifs, M1-4 (Figure 3.10). In accordance with Grienenberger & Douglas (2014), we found motifs M1-3 to be present in all *L. tenue* homologs but inconsistent presence of motif M4, which was only retained in Contig_051339 and Contig_141165. Of special interest was the dissimilarity between *L. tenue* Contig_141165 and Contig_051339-41 sequences; contigs Contig_051339-41 share many residues that were not present or conserved in Contig_141165. Contig_141165

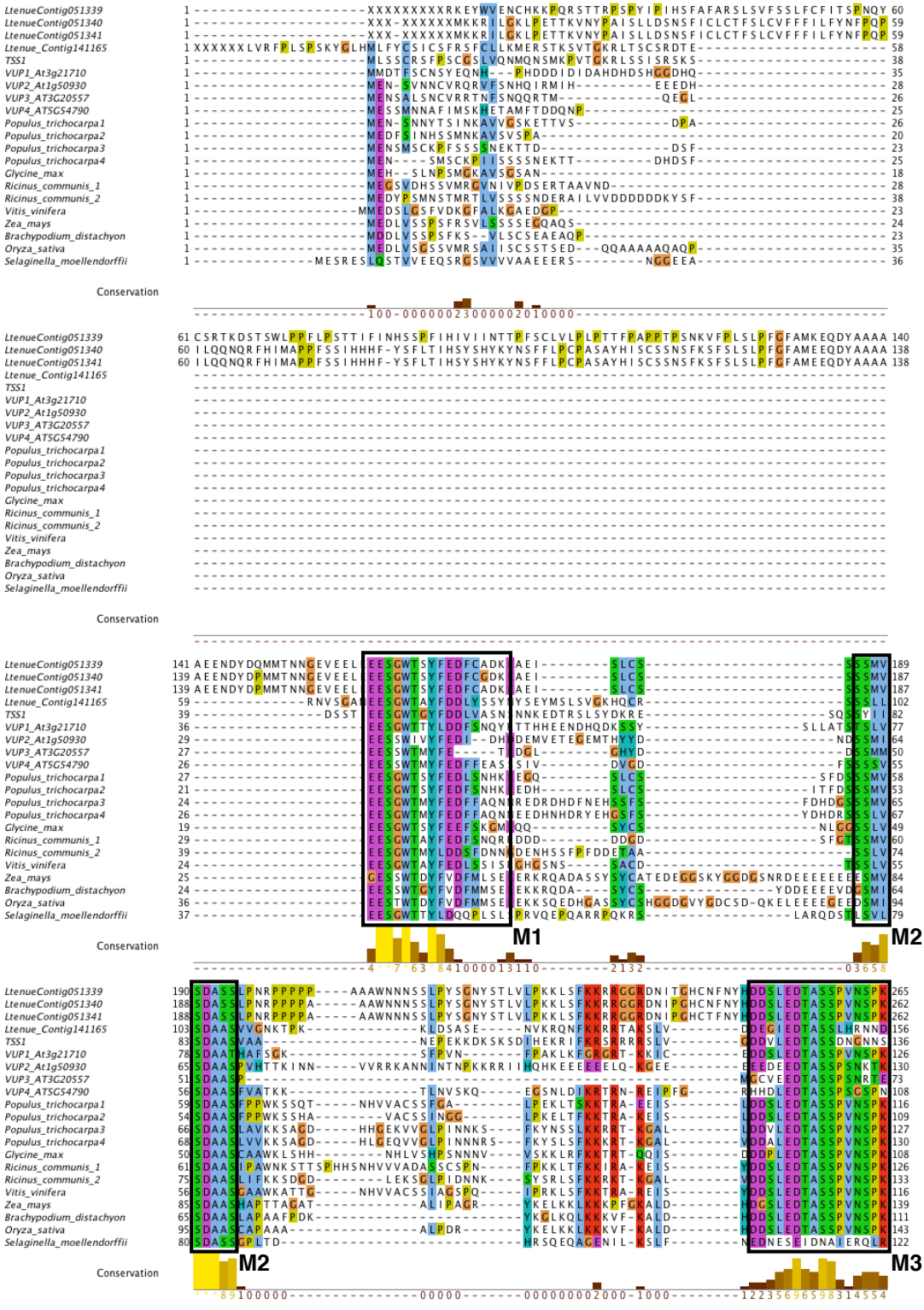


Figure 3.10 Alignment of full-length Ltenue_Contig141165_5'3'_Frame3 and its putative paralogs, TSS1, VUP1-4 and putative *VUP1* homologs from across the angiosperm phylogeny and *Selaginella moellendorffii* (Lycopodista). Alignments were made using MUSCLE v3.8.31, and amino acid residues coloured according to their physicochemical properties (Clustalx colouring system). M1-4 indicate conserved sequence motifs, as defined by Grienenberger & Douglas (2014), among VUP1 homologs.

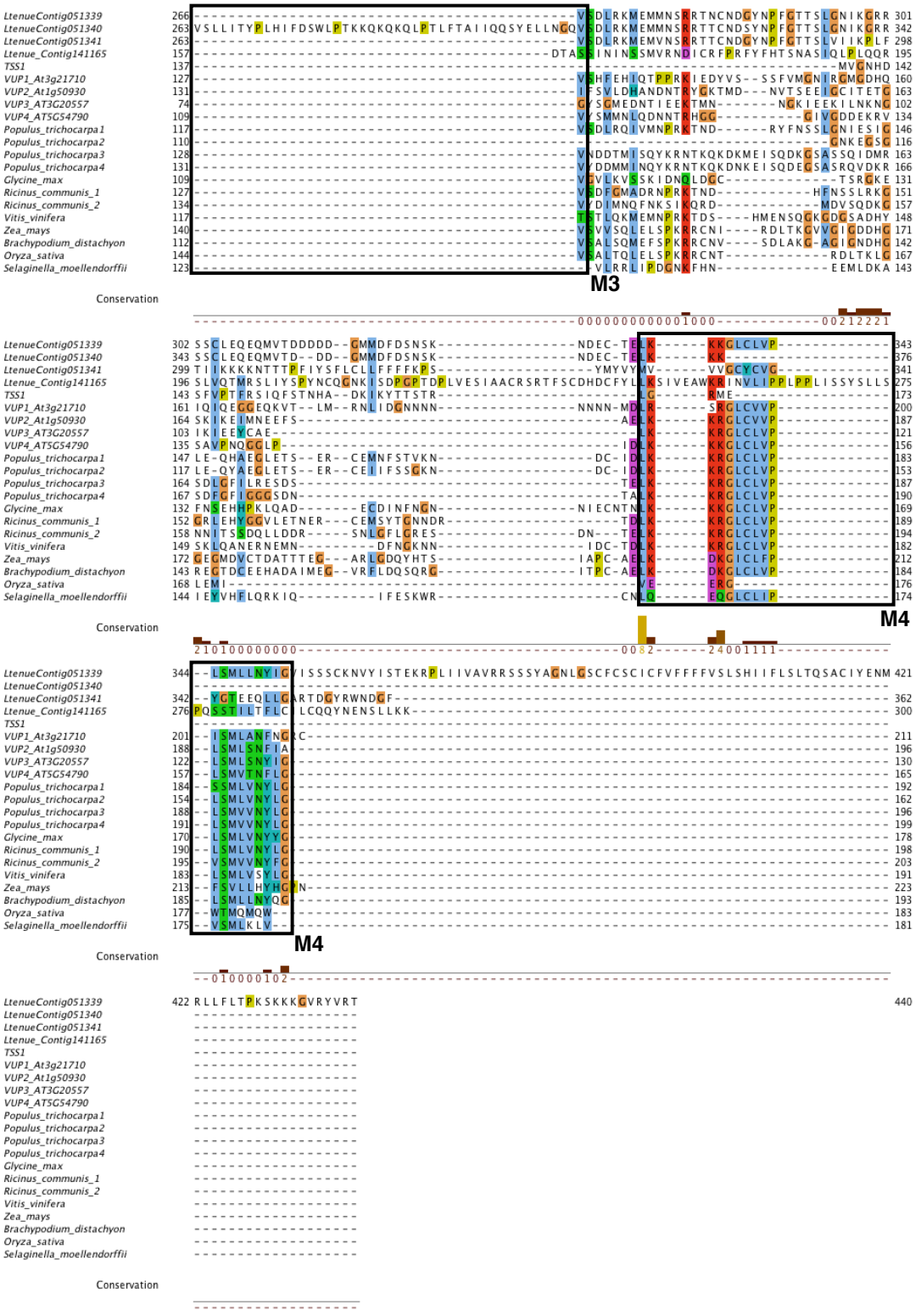
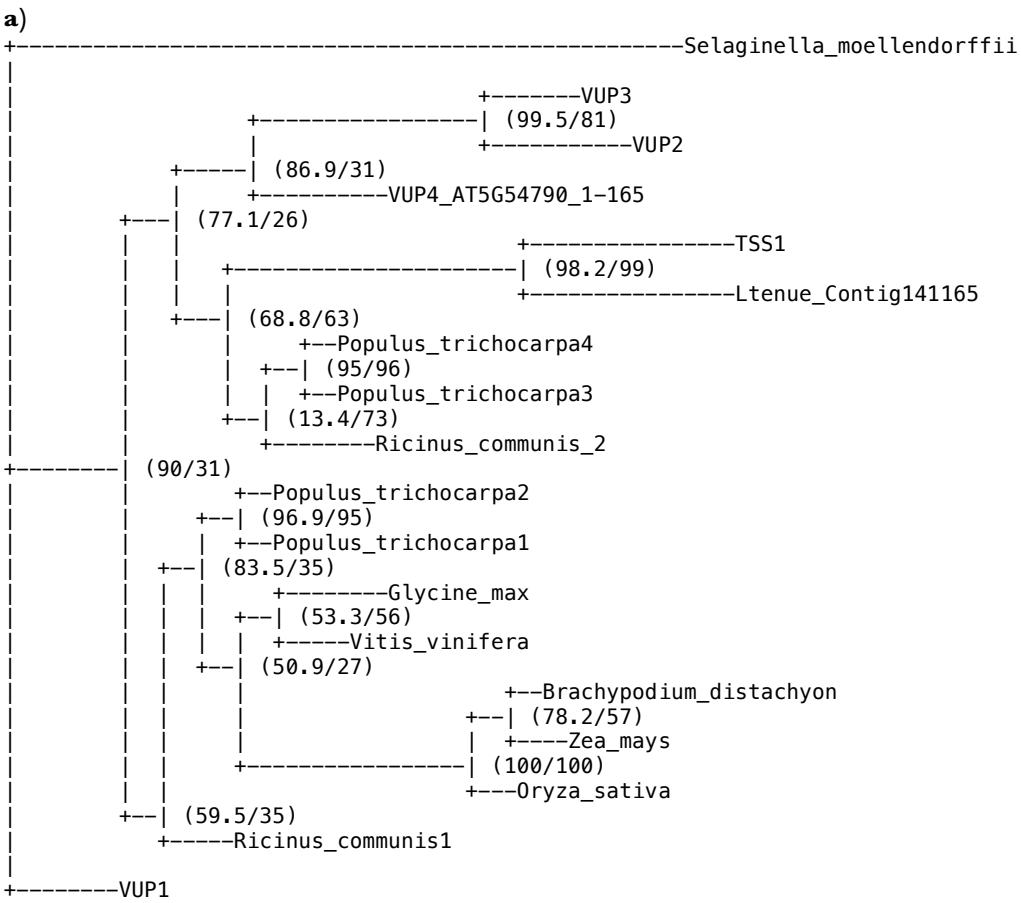


Figure 3.10 continued.

appeared to share greater similarity with TSS1 and VUP1-4 than its conspecific (putative) paralogs. The amino acid sequences of *L. tenue* homologs Contig_051339-41 were all longer than the other sequences in the alignment set, with the presence of ~140-150 leading amino acid residues. Also apparent from our alignment was the truncation of TSS1 relative to other homologs, resulting in an absence of the M4 motif; the alignment in Appendix IV shows the absence of ~100 residues on the trailing end of the sequence relative to L. tenue Contig_141165.

The alignment set of the twenty-one putative homologs of Contig_141165 and TSS1 were used to create two sets of trees (Figure 3.11), excluding (Figure 3.11a,b) and including (Figure 3.11c) putative *L. tenue* paralogs of Contig_141165. Considering both sets of trees, support for the nodes were generally highest at the leaves of the tree, and decreased for most groups as nodes approached the base. The most evident exceptions to this were for the subclades of Poaceae (monocots) species *Zea mays*, *Brachypodium distachyon* and *Oryza sativa*. In all trees, Contig_141165 and TSS1 were shown to be closest to each other, forming their own monophyletic groups with high support. For the exclusive analysis, *Selaginella moellendorffii* consistently formed a clear outgroup at the base of the trees relative to the other angiosperm taxa. The inclusive analysis exhibits the paralogs of Contig_141165 and contigs Contig_051339-41, forming distinct outgroups to all other taxa. In further contrast to the exclusive analysis, the topology of the inclusive maximum likelihood and consensus trees was equivalent, with a Robinson-Foulds distance of 0 (Robinson-Foulds distance of 8 between exclusive trees).

a)



b)

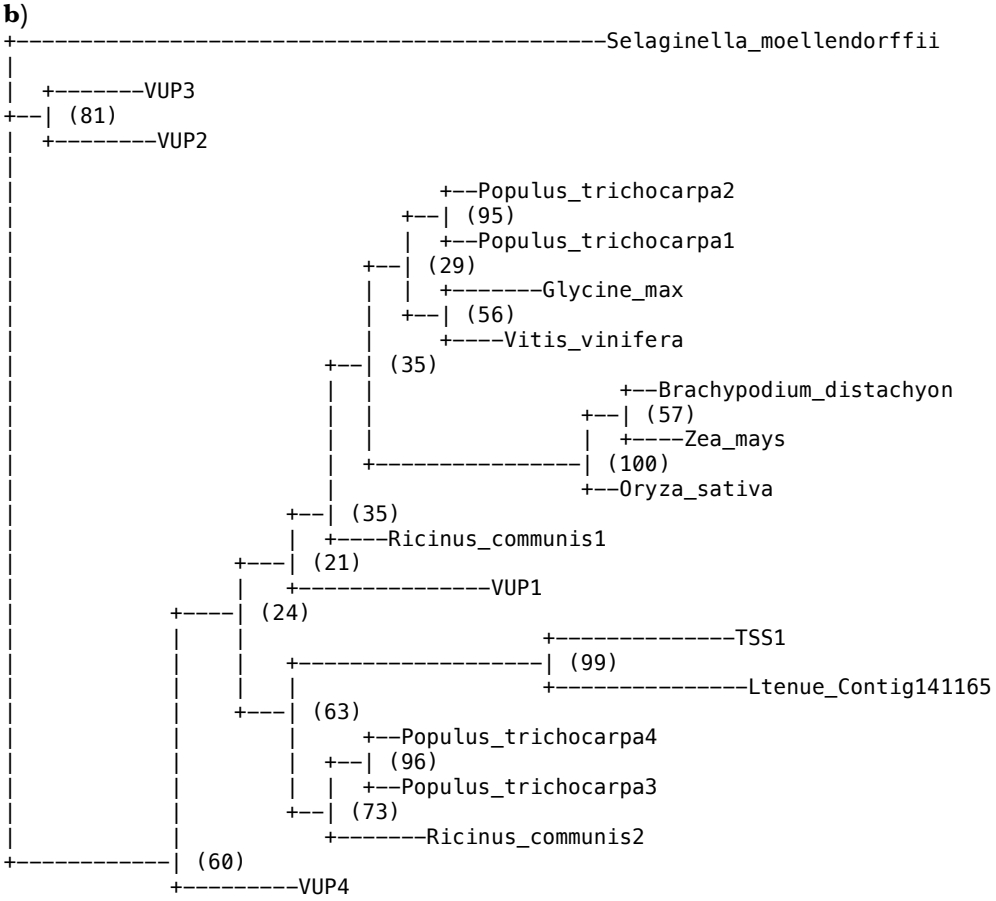


Figure 3.11 Legend on continued figure below.

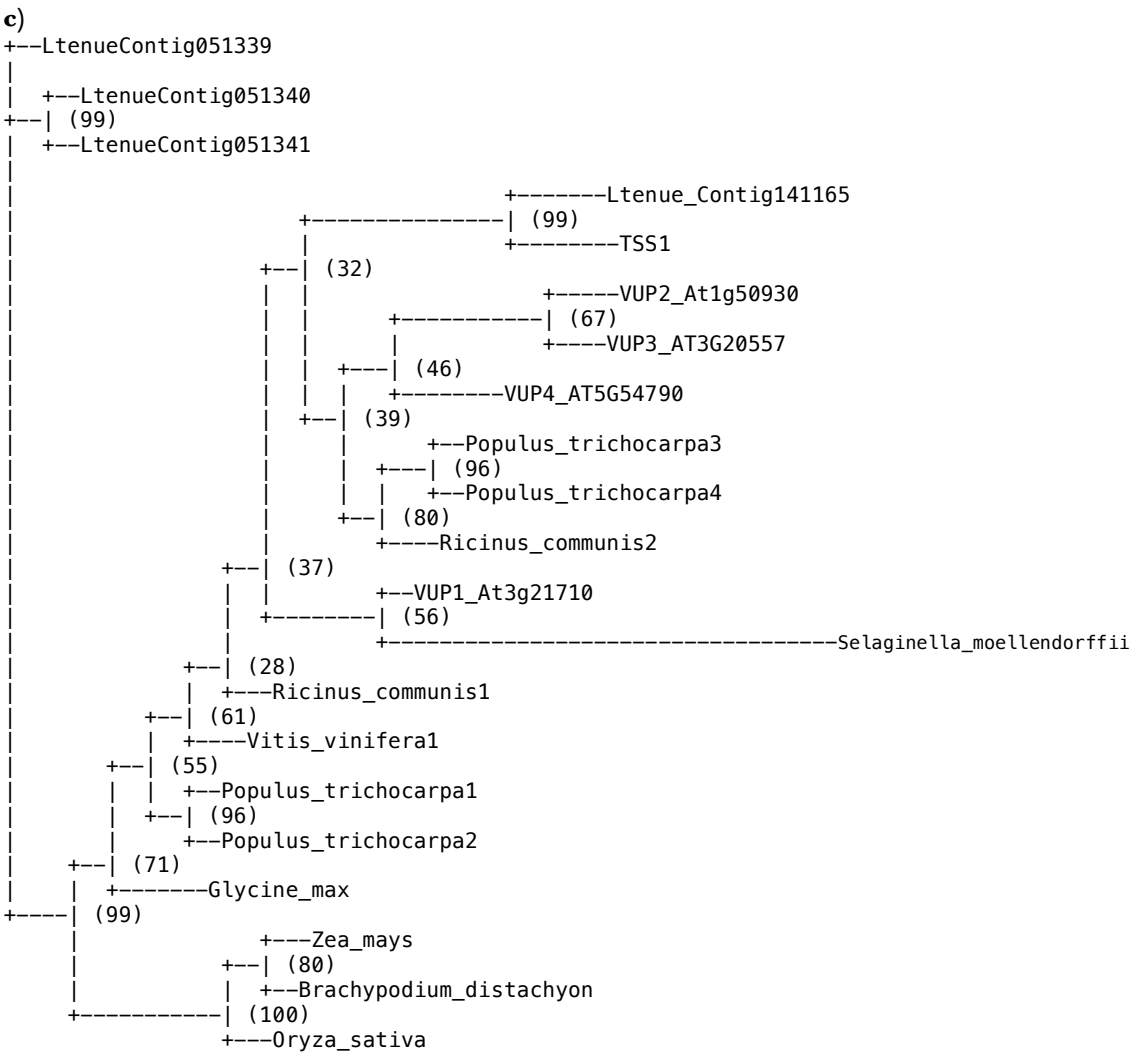


Figure 3.11 Unrooted trees of *L. tenue* Contig_141165, its putative paralogs (Contig_051339-41), its putative ortholog in *L. grandiflorum* (TSS1) and its putative homologs from across the tracheophyta. **a)** Maximum likelihood tree of the protein set excluding *L. tenue* contigs Contig_051339-41, with percentage SH-aLRT support and ultrafast bootstrap support respectively shown for each node. **b)** Consensus tree of the protein set excluding *L. tenue* contigs Contig051339-41 derived from 1000 ultrafast bootstraps, with percentage ultrafast bootstrap support shown for each node. **c)** Consensus tree of the whole protein set, displaying ultrafast bootstrap support as in b).

3.4 Discussion

3.4.1 Comparison of mapping tools and parameters

Alignment of the reads to a genomic or transcriptomic reference is a crucial step in any RNAseq pipeline and is often the step where a majority of the data can be lost if not performed correctly. If cDNA reads are being aligned to a genomic reference sequence, then a 'splice aware' aligner should always be selected for this task. The two most commonly used and highly regarded tools, both in the literature and online forums,

specifically designed for this purpose are STAR and TopHat2. Our findings that STAR generates higher overall and unique alignment rates, whilst also being much quicker and less computationally burdensome, are of little surprise and support the reported literature (Dobin *et al.*, 2013; Engström *et al.*, 2013). It is however an important step for any RNAseq study to experiment with the choice of tools and parameters, as different datasets may be better suited to different tools. Our data demonstrates little variation between the presetting parameters for each software tool, accounting for only small differences in overall results, whereas the majority of variance is between the tools used. However, efforts should be taken to maximise the rates of mapping as even small percentages can represent hundreds of thousands of reads, and as the distributions of reads mapping to each feature in the genome is rarely uniform, this could lead to disproportionate changes in transcript counts or detection across the genome. Our data (Table 3.1) demonstrate that STAR is more suitable for mapping of the *L. tenue* dataset to the reference transcriptome; with a consistent ~10% increase in total alignment rates and ~20% increase in unique alignment rates. Our data also demonstrate lower variation between samples when aligning with STAR, with standard deviations ranging from 1.92 to 1.95, whereas with TopHat2 the standard deviation of samples ranged from 5.22 to 5.43.

It should be noted that the presetting values were not designed to be directly comparable between the alignment tools used. The algorithms governing both are different, particularly with regards to seeding (matches between parts of the read and the target sequence) and extending (the dynamic process by which the rest of the read is matched), and for example changes to the seed parameters in STAR will have larger consequences downstream in the alignment process. For this reason, seed options were only adjusted in the TopHat2 presettings, though gap and insertion penalty scores were changed for both. The presettings thus aimed to demonstrate the differences in alignment that can be derived from the same tool. Given the nature of the *L. tenue* reference, which as explained

in Chapter 4.0, had been created from consensus unigene sequences, it was expected that reference loci may be comprised of allelic and transcript isoform variants. Thus the loose presetting was selected for the STAR aligner when aligning for the purpose of transcript quantification.

3.4.2 Library quantification

The quantification demonstrated that the libraries were of highly variable size, with there being an over one hundred-fold difference between the smallest and largest libraries in terms of reads aligning to the reference. Various control measures were taken during the sample and library preparation procedures in an attempt to avoid such variation effects between samples. All plants were grown in the same glasshouse and samples were harvested at the approximately the same time of day and season: between 11:00 and 13:00 over a two-week period in June 2014 (see Section 4.2.1). Care was also taken to normalise RNA concentrations using fluorimetry post-extraction and post-DNAse treatment, and equal volumes (and thus equal total quantities of RNA) were used as input for the library preparation protocols. After construction, libraries were then quantified using qPCR and concentrations were again normalised prior to sequencing. The variation in sample sizes raises various issues and challenges for differential expression studies and downstream applications for specific contigs of interest. There is an increased chance of false positives being present in the dataset due to the variable libraries, as results could be heavily skewed by outlier counts in a single sample. As the power of differential expression experiments are sensitive to the number of observations for each feature, the samples with the lowest number of reads are likely to be most adversely affected. Similarly, there is an increased risk of false negatives in the dataset, as the detection of lowly expressed transcripts might become lost in the overall noise (Finotello and Di Camillo, 2015).

Normalisation of the read counts was attempted on a per sample basis using CPM, quantile normalisation and upper quartile scaling, as these methods can be effective at removing library size effects (Evans, Hardin and Stoebel, 2018). Upper quartile normalisation did yield highly comparable counts between the libraries, but given the disparate range in library sizes within and between triplicate sets, normalisation through these methods are unlikely to be entirely sufficient. Given the amount of variance in interquartile ranges between samples and treatments, normalisation methods that use quartile information are particularly unreliable for smaller libraries. If libraries within a dataset are comparable and have globally high read counts (>25 million), variance in counts will naturally lower as highly expressed features tend to deviate less from the mean. However, with globally fewer read counts, the variance in expression on a sample-by-sample basis will be much higher, thus counts in the top quartile will be subject to greater stochasticity. Zyprych-Walczak et al. (2015) examined the impacts normalisation methods can have on the count data, particularly highlighting the bias that can arise in upper quartile normalisation. We thus opted to employ the DESeq2 model of normalisation that uses library scaling factors based on median values, which is less susceptible to outliers than distribution adjustment approaches.

To provide some measure of control over the differences in library sizes seen in the full dataset, large samples were subsetted to 5,978,885 randomly selected paired reads (the first quartile limit). This subsetted dataset aimed to reduce the magnitude of variance between library sizes and was used alongside the full dataset for all downstream expression analyses as a control for variable sample size. Though the differential expression testing in DESeq2 uses Cook's distance to remove count outliers and features with low counts (where the mean of normalised counts falls below a threshold), this was used as a comparative measure to assess the overall behaviour of the full dataset. A caveat, however,

of subsetting the data is that it can reduce the power of the experiment in three main ways. With fewer observations, the shrinkage estimations used to match the data to the generalised linear model may be affected and result in the exclusion of more features for differential expression; and lower gene counts overall can increase the number of features falling below the mean-of-normalised-counts-threshold that are discarded from the analysis. The presence of lower counts overall also tends to increase the severity of the FDR *p*-value adjustment, resulting in more features falling above the user-defined significance threshold, that we set to 0.05.

This creates a trade-off that can be difficult to reconcile, as use of the subsetted dataset increases the likelihood of false negatives, yet the range of variation in orders of magnitude between samples throws the reliability of results for the full dataset into doubt. The expectation is that the subsetted dataset will detect fewer differentially expressed features, but that if a majority of these features are also found to be comparably differentially expressed in the full dataset we can be more confident regarding the impacts of sample bias.

3.4.3 Global expression patterns

The PCA of the full dataset prior to normalisation and differential expression analyses (Figure 3.5) revealed that the developmental transcriptomes are on the whole distinct, with discrete clustering of samples by growth stage. This provides a strong indication that the data are biologically sound, as we would expect different flowering stages to be specifically characterised by differential expression at many loci (Gao *et al.*, 2014; Klepikova *et al.*, 2015). We know that the *S* locus is a multiallelic supergene, but the differences between morphs is likely to be driven by the cumulative effects of many more

differentially expressed genes that are regulated by the *S* locus or earlier stages of the morph-specific developmental pathways (Cocker *et al.*, 2018).

Of interest in the PCA results is the high level of variation that the first few principal components explain. This analysis is sensitive to the variance contributed by each feature. It is therefore unsurprising that the open flower and leaf clusters each display extreme outlier samples as these outliers are the two smallest libraries (Table 3.2), and represent the samples with the largest deviation from the mean and median library size. In contrast, the larger-sized libraries were less prone to be outliers. With increased observations of count data, the variance in counts tends to decrease (hence the negative binomial model of DESeq2); reducing the magnitude of principal components for large samples. It is also possible that, given both coverage of the reference and average read depth of covered regions is proportional to library size, when counts are log2 transformed, the larger libraries are closer to the mean and median values. That the difference between PCA of the full dataset and PCA of the subsetted dataset is near negligible is further evidence that the smaller libraries are producing the largest skew in the data, as opposed to generalised variations in library size overall. This suggests that libraries above a minimum sequencing depth are more import for reliable expression analyses, which is a view echoed in the literature (ENCODE consortium RNAseq guidelines: https://bit.ly/2CXRQmI).

A t-SNE analysis was also performed on the data prior to differential expression testing (Figure 3.6). t-SNE is a similar technique to PCA in many ways in that it performs dimensionality reduction for a multivariate and high-dimensional dataset. However, while PCA functions purely to explain the variance observed in a dataset, t-SNE attempts to take the underlying structure of the data into account by giving more value in a similarity matrix to neighbouring nodes (Van Der Maaten and Hinton, 2008). When t-SNE was performed on the data and allowed the features to contribute by variance,

discrete separate clusters were generated based on developmental stage, similar to the PCA. However, by allowing features to contribute equally, the t-SNE presented here reduces the effect that more variable features have on the graph, which may result from differences in library size. As can be seen from Figure 3.6b, the first two t-SNE principal components group samples largely by developmental stage. In comparison to the PCA, the floral growth stages are overall grouped closer together in the t-SNE, as opposed to the open flower samples forming a discrete cluster in the PCA graph, and leaf samples are mostly confined to the right-hand edge of the graph. It appears also, based on t-SNE PC1 and PC2 alone, that pin-morph samples within a developmental stage tend to cluster closer than the equivalent developmental stages of thrum-morph samples. However, when PC3 is taken into account (Figure 3.6a) this pattern is not apparent. Nonetheless, this could provide some support for differences between pin and thrum expression patterns that will be discussed later.

In contrast to the PCA, the t-SNE plot of principal components 1-3 does suggest a minor degree of distinction between the pin and thrum transcriptomes. This difference is very slight, and only becomes more significant once the t-SNE is conducted on the differentially expressed data (Figure 3.7), though does give credence to the argument that morph differences in expression are derived from the regulation of a (relatively) small number of genes.

3.4.4 Differential expression with DESeq2

Differential expression was conducted to compare all pin floral growth stages and all thrum floral growth stages as, given the variance in library sizes, a sample size of nine provides greater power than pairwise comparisons of individual floral growth stages in

triplicate. Differential expression analysis found 2363 significantly differentially expressed features in the full dataset and 1885 significantly differentially expressed features in the subsetted dataset. As observed in Figure 3.3, the 1371 differentially expressed features that are shared between the two datasets were found to be more significant in the full dataset, as demonstrated by the greater skew towards zero *p*-values. This differential was to be expected, as discussed in section 3.4.2. High overlap of the differentially expressed features in the subsetted and full datasets was promising, suggesting that the large variance in library size did not have too severe an impact on the ability of DESeq2 to remove false positives from data with greater levels of dispersion.

As shown in Figure 3.4, other than the number of differentially expressed features detected, there was little difference in the dispersion of the overall sample expression between the full and subsetted datasets. The largest difference was the reduction of features showing extreme changes in expression levels, particularly for down-expressed features. This disproportional reduction in extreme values is likely due to features, already shown to be lowly expressed, failing to meet the count value threshold for significance in the subsetted dataset. This is further evidence that the full dataset has increased power to avoid type II false negative errors.

When conducted on the differentially expressed feature set (Figure 3.7), t-SNE analysis revealed clear distinction between pin and thrum individuals. This result is to be expected, given the data is comprised solely of differentially expressed features, but allows the nature of the distinction between the two morphotypes to be examined in greater detail. A t-SNE analysis prior to differential expression suggested slightly closer neighbouring of pin individuals than thrum individuals within the developmental stage groups. However, the neighbouring pattern changed when conducted on the differentially expressed dataset. Figure 3.7 clearly shows sub-clusters of pin samples that subcluster strongly by individual

rather than by growth stage. This subclustering by individual was also reflected in the thrum samples, but was subject to greater overlap and was therefore less distinctive. This observation is further reinforced by the results of hierarchical clustering (Figure 3.8), where the floral developmental stage samples for pin individuals sort into three distinct clades, yet the thrum samples, whilst broadly sorting by individual, displayed weaker grouping by individual and was more subject to outlier samples. This pattern could be due to two reasons: true representation of biological activity or sample bias between pin and thrum datasets.

Assuming that there was unbiased sampling of the natural diversity seen in *L. tenue* populations among the six sampled individuals, and that there was no sample bias between pin and thrum libraries, this result could be indicative that there is more variation in overall gene expression between pin individuals than different floral stages within individuals, but that this pattern is weaker for thrum individuals. This result supports recent findings in the heterostyly literature that suggests that the *S* locus controlling distyly shows a hemizygous determination of thrum individuals as a result of a single thrum supergene not present in pin individuals. Younger thrum loci that have arisen more recently in evolutionary time will have accumulated fewer allelic variants than the older pin loci, and thus thrum individuals may display a smaller degree of genotypic diversity. Li *et al.*, (2016) indicates that the thrum morphotype evolved from a proto-pin (approach herkogamy) progenitor, with the proto-pin present 100-125 million years ago (mya) and the duplication event leading to the present-day thrum morphotype occurring 33.1-72.1 mya. Another common occurrence during duplication events is that gene regulatory mechanisms can be disrupted, causing effects similar to what is known as 'transcriptomic shock' (Hegarty *et al.*, 2006). This could be a factor in explaining the tighter regulation observed in pin individuals. The dynamics of population genetics between the distylous morphs may also provide a speculative explanation for this pattern of subclustering.

Meeus *et al.* (2012) present similar finding when investigating the effects of morph frequency bias in populations of *Pulmonaria officinalis* (Boraginaceae). When populations showed a frequency bias towards pin individuals, the F_{ST} (a genetic measure of population differentiation) values tend to decrease, especially when populations are located in close proximity to each other. When F_{ST} values are low (<0.1) due to gene flow between populations, pairwise genetic distances are higher for thrum individuals between populations than for pin individuals, indicating lower levels of per population diversity in thrums. However, this is dependent on specific traits observable in *P. officinalis* and may have limited implications for the pattern we see in *L. tenue*. Primarily, *P. officinalis* pin individuals produce double the amount of pollen grains and this results in higher amounts of pin-pin gene flow between populations. The pin frequency bias reduces the effective thrum population size, and thus can lead to lower levels of thrum diversity.

Another possible contributing factor to the morph-specific structural subgrouping lies in the dominant inhibitory effect of the candidate Contig_141165, a homolog of *VUP1*. Due to the interplay between genetics, environment and chance, nature displays inherent variation, which commonly manifest in measurements for traits showing Gaussian distributions. If left to elongate uninhibited, the cells in the pin styles and filaments of different individuals, and their underlying gene expression profiles, are likely to display natural variation. However, in thrum styles, if the *VUP1* homlog, which is known in *Arabidopsis thaliana* to have an inhibitory effect on the expression of many genes, is acting to supress cell elongation before it has begun, we might expect the level of variance between thrum individuals (in both gene expression and cell length) to be greatly reduced. This pattern is observable in our data for cell lengths (Figure 2.3).

Non-biological explanations are also possible. The power of the data should be taken into consideration as these patterns of expression could also be artefacts of experimental design

and library bias. First, when considering the results of the hierarchical clustering analysis (Figure 3.8), the groupings should be treated with caution in the absence of bootstrapping support, as the strength of the emergent pattern could be weak. The outlier samples within the thrum clade also deserve closer examination. The two outlying samples were CBT8_OpenFlower and CBT7b_YoungBud, which were the first and third smallest libraries with ~1.5 million and ~4 million aligned reads respectively (Table 3.2). For CBT7b_YoungBud, the main expression differences from the other samples for individual CBT7b appear to be within the block of upregulated thrum features (top left of Figure 3.8), where a stretch of features appear to be relatively downregulated in CBT7b_YoungBud compared to other CBT7b samples. While we would hope for DESeq2 to effectively remove skewed feature counts in relatively small libraries, the possibility of unequal coverage in the smaller libraries cannot be ruled out. Similarly, in the case of CBT8_OpenFlower, the main expression differences from other CBT8 samples was in the same block of upregulated thrum features, with relative up- and down-expression for some features. Given CBT8_OpenFlower was the smallest library, similar effects of library RNA composition bias cannot be ruled out. If expression patterns were more uniform, we could see a restructuring of the thrum clade reveal similar sub-clades grouped by individual, similar to the pin clade.

3.4.5 Experimental power

A post-hoc power analysis (Appendix VII) revealed that, for our triplicate experimental design, our experiment could have power as low as 60%, indicating our analyses could have failed to detect as many as 40% of the true positives in the dataset. Both at the time of the experiment and now at the time of writing, our experimental design was in line with the ENCODE guidelines and the literature, though future experiments should

always use a pilot dataset to conduct a power analysis prior to running a full investigation. The inclusion of even two extra biological replicates per treatment could increase the power of the experiment to 80-90%. It is possible therefore that our dataset does not at present contain the full complement of loci that show significant differential expression between pin and thrum morphs. That the expression results for the full and subsetted datasets were very close is indicative that the full dataset is, overall, reliable to use for differential expression with DESeq2. The reliability of the results is largest when large scale patterns in gene expression are examined overall, instead of examining individual features of interest from the differentially expressed list. Coefficients of variation for feature expression counts in each triplicate set have also been created, and could be used as a further diagnostic confidence measure when considering each of the differentially expressed features separately.

3.4.6 Genomic clustering

Genome clustering analysis (Figure 3.9) revealed some interesting co-localisation of differentially expressed contigs on selected chromosomes in reference species. Initially conceived as a further way of inferring gene function beyond the transcriptome annotation, the clustering of *L. tenue* contigs in the same genomic regions, particularly on *A. thaliana* Chromosome 4, *P. trichocarpa* Chromosome 19 and *P. veris* Contig927, is of strong interest. Given that the *S* locus is a tightly-linked genomic region, predicted to be located near a centromere (Pamela and Dowrick, 1956; Li *et al.*, 2015), finding how contigs co-localise on a reference genome could indicate candidate regions for the *S* locus in *L. tenue*. The reference genomes used for clustering were selected carefully: the closest possible species to *L. tenue* with the highest possible resolution, at the chromosome-level or at least with a contig-scaffold N50 in the megabases.

Of great interest in our results is the co-localisation of 142 differentially expressed *L. tenue* features on a ~13 Mb region of *A. thaliana* Chromosome 3 and surrounding the genomic location of *VUP1*. If Contig_141165 is a homolog of *VUP1*, sequences of the features falling in this location could represent loci of the *L. tenue S* locus, or neighbouring loci linked to the *S* locus through reduced recombination. This provides another avenue for further exploration to track down the genomic location of the *L. tenue S* locus. However, this method is dependent on many assumptions of synteny and shared genetic mechanisms for self-incompatibility and heterostyly, many of which may not hold. It is possible that many of the differentially expressed features not finding BLAST matches are novel loci, or unique to *L. tenue*. Further refinements of the method are thus necessary to bring these findings forward, repeating the technique with the full-dataset in particular.

3.4.7 Contig_141165 represents a candidate *G* locus allele of the *L. tenue S* locus

The discovery of Contig_141165 as a possible ortholog of *TSS1* and its possible relationship (as a putative homolog) with *VUP1* is a significant finding. To the author's current knowledge, no connection in the literature has yet been made in heterostylous taxa, in *Linum* or otherwise, between putative *S* locus candidates and *VUP1* in *A. thaliana*. Proteomic and transcriptomic approaches in *L. grandiflorum* have demonstrated thrum-specific expression of TSS1 in the tissues of the (short) style (Ushijima *et al.*, 2012), and subsequent investigations have demonstrated that *TSS1* is absent from pin genomic DNA (Ushijima *et al.*, 2015); this is strongly suggestive of hemizygous dominant control of the thrum phenotype, as is evident in *Primula* (Li *et al.*, 2015, 2016; Cocker *et al.*, 2018). The expression signature for Contig_141165 is concordant with this finding: Contig_141165 is only expressed in thrum individuals and is absent from vegetative tissue. We found that

expression of Contig_141165 only begins in the later stages of flower development, which complements our findings in Chapter 2.0 that differences between male and female organ lengths accelerate in the latter stages of floral development (Figure 2.2). Another key finding of Ushijima *et al.* (2015) is the reduction of cell lengths in the tissues of the thrum style, which again is echoed in our findings. These findings together provide strong evidence that *(i)* the physiological developmental mechanisms, and *(ii)* their causative genetic control, determining style height in the thrum flowers of *L. grandiflorum* and *L. tenue* are equivalent. Our discovery that Contig_141165 is a potential homolog of *VUP1* advances our knowledge of the genetic basis of heterostyly in *Linum* as VUP1 has been shown to be strongly correlated with reduced cell elongation when expressed in the tissues of *A. thaliana* (Grienenberger and Douglas, 2014).

Through an extensive study by Grienenberger and Douglas (2014), *VUP1* (*VASCULAR-RELATED UNKNOWN PROTEIN1*) in *A. thaliana* was found to encode a predicted protein of 24 kD, whose expression was detected in various organs and tissues, particularly in vascular tissues and the tissues of floral organs, namely the sepals, petals, and stamen filaments. Constitutive overexpression resulted in a range of substantial defects, namely: shorter floral organs, severe dwarfism, and a 75% reduction in epidermal cell lengths of hypocotyls. Transcriptomic analyses also revealed the effects of *VUP1* overexpression to be surprisingly pleiotropic, repressing the expression of many genes involved in the brassinosteroid, gibberellic acid, and auxin-response pathways, which are known to be involved in the regulatory control of cell elongation and floral development (Goda *et al.*, 2004, 2008; Cao *et al.*, 2006; Sun *et al.*, 2010). Based on the strong dominant inhibitory effects of VUP1, the thrum style-specific expression of TSS1, and our patterns of thrum-specific expression, we propose Contig_141165 is a strong candidate for the *G* locus of *L. tenue*; the style length and female incompatibility type determinant of the *S* locus.

Interestingly, despite its substantial and far-reaching regulatory effects, the function of *VUP1* remains unknown, and the report of Grienenberger & Douglas (2014) appears to be its only occurrence in the literature. Discovery of the sequence of TSS1 preceded the discovery of the function of VUP1-4, and neither VUP1, TSS1 or Contig_141165 contain recognisable functional, structural or conserved domains. This perhaps explains why the link between VUP1 and TSS1 has, until now, remained undiscovered. Intriguingly, in their sequence analysis of TSS1, Ushijima *et al.* (2012) attempted to find homologs in a range of species, and independently found similar conservation of the diagnostic residue motifs later documented by Grienenberger & Douglas (2014). Both groups used VUP2-4 and many of the same species in their analyses, yet the connection between *TSS1* and *VUP1* was not made. This is most likely a result of the overall low sequence identity between homologs (35-40%), and that the M4 domain is absent in the truncated TSS1 protein. The construction of gene models for peptide searches, and PSI-BLAST approaches (which are designed to specifically search for conserved domains and motifs), used during our annotation pipeline (Chapter 4.0) has likely been a key factor in our discovery.

This high level of overall divergence observed in Contig_141165 and its homologs is another key line of evidence for it being a *S* locus candidate. As described by Grienenberger and Douglas (2014), *VUP1* is possibly unique to the vascular plant (tracheophyta) lineage. *VUP1* and its homologs appear to be part of a small gene family, with no more than four members. In this way, contigs Contig_051339-41 may represent the full complement of paralogs in *L. tenue*. Our phylogenetic analysis of Contig_141165 inclusive of its putative paralogs (Figure 3.11c) are suggestive of a high level of divergence within the *L. tenue* genome. That Contig_141165 is more closely related to TSS1 and divergent from its paralogs provides evidence for reduced recombination at the Contig_141165 locus and an equivalent function to TSS1.

3.5 Concluding remarks

The study described in this chapter has demonstrated distinct differential expression between the pin and thrum morphs of distylous *L. tenue*. Despite potential issues with variations in library size within the dataset, we have demonstrated that our data are still effective for discerning overall differences in gene expression between morphs and floral developmental stages. The discovery of Contig_141165 as a candidate for the *G* locus is of interest and may contribute to our understanding of heterostyly genetics and development in *Linum*.

Further work investigating the biological processes, such as pathway enrichment studies of this and similar high-throughput transcriptomic datasets, would be extremely helpful and could further our understanding of the underlying genetic mechanisms of heterostyly in *Linum*.

- Chapter 4 -

High-quality transcriptome condensation into consensus unigene sequences with BALLISTA: a case study with the *de novo* transcriptome assembly of *Linum tenue*

***Linum tenue*.** Image by Alireza Foroozani

4.0 High-quality transcriptome condensation into consensus unigene sequences with BALLISTA: a case study with the *de novo* transcriptome assembly of *Linum tenue*

Alireza Foroozani[1] and Adrian C. Brennan[1]

1 Department of Biosciences, Durham University, South Road, Durham DH1 3LE, UK

4.0.1 Preamble

The following chapter is comprised of a journal article manuscript aiming for submission as a methods paper at Genome Research. The manuscript structure is thus designed to be in line with Genome Research's guidelines. For this reason, many aspects of the work have been condensed. However, for the purposes of this thesis, some features of section/subsection formatting, word limits, additional explanations, and the inclusion of extra figures, have been appropriated for continuity and coherency.

I would like to define here for the reader the term '*k*-mer', which is used in this chapter and in the General Discussion (Chapter 5.0). Modern genome and transcriptome assembly tools assemble sequencing reads into longer sequences, which are representative of their parent genomic or transcript sequences, using de Bruijn graphs. All possible substrings of a defined length (*k*) present in the input sequencing reads are extracted, which are known as *k*-mers. A de Bruijn graph is then constructed by assigning the *k*-mer sequences to the vertices, and (*k*-1)-mers (the *k*-mer minus the first or last base) are assigned to the nodes. Assembly algorithms function to solve the de Bruijn graph by joining each overlapping node in what is known as a Eulerian pathway, thus using *k*-mers as the windows of overlap between the reads to assemble longer sequences.

4.0.2 Abstract

High-throughput transcriptome analyses (RNAseq) are being increasingly applied to non-model organisms due to affordable costs and the development of accessible tools to conduct *de novo* transcriptome assembly. However, assembling a high-quality transcriptome reference to conduct downstream expression analyses without the use of a genomic reference comes with its own unique set of challenges. Modern transcriptome assembly tools are highly efficient at handling high volumes of read data with variable coverage and are understandably in high demand. However, the initial *de novo* transcriptomes that are assembled are often hundreds of thousands of contigs long and mostly composed of fragmented transcripts, allelic variants, and alternate-splice isoforms. This isoform variation can hinder contig elongation in the assembly process, and can be a greater impediment if the non-model species has a complex (i.e. large, eukaryotic and heterozygous) genome and RNA is extracted from multiple individuals grown from wild-sampled seeds. We present the BALLISTA (aBstraction of ALLelic and ISoform-level variation for Transcriptome Analyses) pipeline as an additional post-assembly processing method, designed specifically to improve the quality of *de novo* transcriptome assemblies derived from mRNA data. Through a quick, automated, user-friendly procedure, BALLISTA uses a close- or distantly-related reference proteome to sort the transcripts of a *de novo* assembly into unigenes, which are then condensed into consensus sequences. We demonstrate the efficacy of BALLISTA, through the *de novo* transcriptome assembly of *Linum tenue* and *Arabidopsis thaliana* from short-read mRNAseq data, to construct high quality reference transcriptomes with reduced ambiguity. This improves downstream applications such as read alignment for transcript quantification and subsequent expression analyses.

4.1 Introduction

Vast improvements in and greater accessibility to sequencing technologies over the last decade have had a resounding impact in the experimental approaches used in many fields of research, such that 'next-generation sequencing' (NGS) is now common vernacular in biology (Buermans and den Dunnen, 2014). The application of NGS for transcriptomic studies, or RNA sequencing (RNAseq), in particular has allowed developments in our understanding gene expression profiling, alternative splicing and of allele-specific expression (Hrdlickova, Toloue and Tian, 2017).

The most popular modern NGS platforms tend to be based on Illumina sequencing chemistry, which generate hundreds of gigabytes of short-length (~100-300 bp) (https://emea.illumina.com/systems/sequencing-platforms.html) sequence data, as they provide the highest value in terms of cost per base sequencing and low error rate. The drastically reduced costs of sequencing have allowed the emergence of genome-scale studies in non-model organisms; where important resources, such as genome sequences and transcript maps, are lacking (Ekblom and Galindo, 2011). As RNAseq requires the mapping of sequencing reads to a genome or transcriptome reference sequence, studies with non-model organisms need to create a draft reference using *de novo* assembly if a close relative with a sequenced genome is not available (Paszkiewicz and Studholme, 2010). There is now a wide array of open access *de novo* assemblers that can efficiently handle the volume and complexity of these shorter reads to construct draft transcriptomes directly from RNAseq data. However, dealing with RNA sequence data provides a unique set of challenges. A variety of factors, such as variable sequencing depth reflecting expression differences across the genome, a lack of information regarding underlying exon-intron structure, and the presence of transcript isoforms such as alternate-splice and allelic variants, make transcriptome assembly problematic in different ways than genome

assembly (Martin and Wang, 2011). RNAseq data also contain inherent errors as a result of PCR-based steps involved in library preparation (Oshlack and Wakefield, 2009; Hansen, Brenner and Dudoit, 2010) or sequencing errors, which can be as high as 3-4% in Illumina data (Dohm *et al.*, 2008).

The most commonly used *de novo* transcriptome assembly tools are Cufflinks (Roberts *et al.*, 2011), Oases (Schulz *et al.*, 2012), and Trinity (Grabherr *et al.*, 2011), the last of which increasingly appears to be the 'assembler of choice' in the literature and online forums. Whilst immensely valuable, these assembly programs are highly sensitive to the errors and polymorphisms present in the sequence data, which often generate assemblies of many hundreds of thousands of contigs relative to the few tens of thousands of expected genes. The inflated sizes of the assemblies are a result of fragmented/incomplete transcripts, and the ineffective collapsing of alternate-splice or allelic variants into loci, producing high levels of redundant transcripts. While obtaining isoform-level information is a key goal for some RNAseq experiments, conducting differential expression on these raw assemblies with unrealistically high numbers of contigs can be problematic for downstream analyses (Finotello and Di Camillo, 2015). Creating a good reference is therefore of utmost importance, and researchers wishing to conduct *de novo* transcriptome assembly should *i)* perform appropriate pre-processing steps of the sequence data, such as error correction and trimming; *ii)* optimise the assembly strategy, through experimentation with various assembly programs and parameters; and *iii)* consider post-processing of the assembly, such as scaffolding or unigene reconstruction.

Scaffolding is a technique commonly applied in genome sequencing to link together contigs produced by the initial assembly. The longer sequence fragments that can be obtained from genomic DNA allow a sequencing strategy to combine paired-end (PE) sequence data from sequencing libraries consisting of varied fragment size distributions,

allowing assembly algorithms to bridge together contigs by gaps of known length (Wajid and Serpedin, 2012). These extended contigs can be also be referred to as 'super-contigs'. The gaps derived from PE sequencing of longer reads may not be supported by sufficient read depth to allow base calling but are still informative to elucidate the relationships between contigs, such as orientation and distances between them. The gaps are represented in the sequence data by strings of undetermined nucleotides, typically denoted as 'N'. Attempts have been made to apply scaffolding techniques to post-assembly *de novo* transcriptomes derived from mRNA data by using mRNA to protein translational information, such as Scaffolding using Translational Mapping (STM) (Surget-Groba and Montoya-Burgos, 2010), and Transcriptome Post-Scaffolding (TransPS) (Liu *et al.*, 2014). The core premise of such techniques is to use translation BLAST programs (i.e. BLASTX) to compare the non-model transcriptome against the proteome of a model or previously sequenced related species, and then to use the obtained BLAST coordinates to guide the scaffolding process. As amino acid polymorphisms between species tend to be under stronger selective constraints and accumulate much more slowly over evolutionary time than non-synonymous nucleotide differences, this can theoretically allow even distantly related organisms to be used as proteome references (Surget-Groba and Montoya-Burgos, 2010). However, it remains to be seen how effective these programs can be in practice for improving non-model transcriptome assemblies. For example, STM only reassembles partially overlapping contigs, and further rejects this reassembly if the output contains more than one 'super-contig'. This can result in relatively limited assignment of contigs into scaffolds, and low levels of condensation of redundant isoforms.

We present BALLISTA (aBstraction of ALLelic and Isoform-level variation for Transcriptome Analyses) as an alternative and improved pipeline for the post-assembly processing of *de novo* transcriptomes. The effectiveness of this new pipeline was demonstrated by our *de novo* assembly of the *Linum tenue* and *Arabidopsis thaliana*

transcriptomes, each derived from short-read mRNAseq data. In contrast to existing scaffolding pipelines, while BALLISTA makes use of translational information to create unigenes, it does not use BLAST coordinates to create a scaffold, and it condenses all overlapping regions between contigs. We thus argue that, as the contig-extension process in our pipeline is unconstrained by the assumption that protein structure and organisation must match the reference, BALLISTA displays greater effectiveness at condensing isoforms into a representative consensus sequence. The quality of a *de novo* transcriptome lies in the proportion of the original sequence reads it explains and its completeness in terms of universal single-copy orthologs. We further aim to generate an assembly with reduced isoform-level ambiguity to aid the accuracy of read alignment for downstream analyses. BALLISTA is also easily run through the Linux or MacOS command line and can be run on single or multiple cores making efficient use of system resources.

4.2 Materials and Methods

4.2.1 Sample collection, library construction and sequencing

The target species, *Linum tenue* (Linaceae), is an annual meadow-growing wild flower occurring in southwest Iberia. Being insect-pollinated, the transfer of pollen between viable individuals is ensured through a species polymorphism known as distyly. Distyly involves two floral morphs (long-styled, LS and short-styled, SS), where male (anthers) and female (stigmas) reproductive organs are spatially separated within the flower but heights extend to complementary positions between morphs to allow reciprocal inter-morph pollen transfer (Barrett and Shore, 2008). A reference transcriptome for *L. tenue* was desired in order to better understand the differences in floral gene expression that lead to the differential floral organ development observed in distyly. Seeds from wild-

Figure 4.1 Qualitative floral developmental stages of *Linum tenue* used to create RNAseq libraries. From left to right: young bud – at this stage the growing bud is seen and encased by sepals, there is no visible floral tissue; immature flower – the floral structure is starting to grow but has not yet fully dehisced, yellow petals are visible behind the sepals; mature (open) flower – the flower is now fully open and all floral organs are exposed and visible.

sampled plants were grown to flowering under partially-controlled growth conditions of 15-20°C and 16 hour supplementary light day length.

Plant tissue was collected from leaves and three qualitatively defined stages of flower development, to capture as wide a range of flowering development as possible, and snap frozen in liquid nitrogen; all samples were collected and RNA extracted over a two-week period in June 2014. Tissues were all harvested within a two-hour window of the day from 11:00 to 13:00, before the petals begin to close in the late afternoon. Total RNA was extracted from young bud, immature flower, open flower, and leaf tissues (Figure 4.1) from 3 LS and 3 SS individuals using TRIzol Reagent (ThermoFisher Scientific, Waltham, US) and DNAse treated with TURBO DNA-free (ThermoFisher Scientific, Waltham, US). RNA samples were quantified using Qubit fluorometry (ThermoFisher Scientific, Waltham, US) before and after DNAse treatment, checked for purity using NanoDrop spectrophotometry, and RNA integrity was assessed by agarose gel electrophoresis. A total of 24 RNAseq libraries were constructed using TruSeq Stranded mRNA kits (Illumina, San Diego, US), using 2 µg of input RNA per sample. Libraries were then quantified using the qPCR NGS Library Quantification Kit (Agilent

Technologies, Santa Clara, US), using the following thermal cycle: 3 minutes at 95°C followed by 30 cycles of 15 seconds at 95°C and 30 seconds at 60°C. Primers provided by the qPCR NGS Library Quantification Kit anneal to the Illumina adapters, indicating whether adapter ligation has been successful. Fragment size distributions were verified using Tapestation (Agilent Technologies, Santa Clara, US). Paired-end (PE) stranded sequencing was performed by pooling all libraries together and running over two lanes on an Illumina HiSeq 2500 platform at the DBS Genomics Facility (Durham, UK) in July 2014.

4.2.2 Data pre-processing

The raw FASTQ reads were preprocessed to exclude low quality or suspicious reads from the assembly. The quality of the raw sequence reads was gauged before and after pre-processing using FastQC (Andrews, 2010). Rcorrecter (Song and Florea, 2015) was used at default settings to correct errors in the reads; any reads containing non-correctable errors were dropped from the dataset. Reads were then trimmed using Trim Galore! v0.4.4 (https://github.com/FelixKrueger/TrimGalore) with the following options: paired mode, retain unpaired reads, minimum sequence length of 36, quality 5 (trimming ends of reads below threshold quality). Sequence reads for libraries run on separate lanes of the Illumina 2500 were concatenated together, generating a single dataset for each orientation and read type for each sequenced library: forward-paired, reverse-paired, forward-unpaired and reverse-unpaired. Ribosomal reads were then filtered out through alignment to a dataset comprising the SILVA LSU, SSU (Pruesse *et al.*, 2007; Quast *et al.*, 2012) and 5SRNAdb (Szymanski *et al.*, 2016) ribosomal datasets, after uracils in these datasets had been converted to thymines using in-house python scripts. Alignments for

ribosomal filtration were performed with Bowtie2 v2.3.3.1 (Langmead and Salzberg, 2012) using the *'very-sensitive-local'* parameter option.

4.2.3 Initial *de novo* transcriptome assemblies

The preprocessed reads were then assembled by combining different assemblies produced using the Trinity (Haas *et al.*, 2013) and Velvet/Oases (Zerbino and Birney, 2008; Schulz *et al.*, 2012) assembly platforms. A single assembly was generated with Trinity v2.5.1, using the default *k*-mer length of 25, minimum contig length of 250, minimum *k*-mer coverage of 2, and path reinforcement distance of 20. The *in silico* read reduction (normalisation) and appropriate strandedness options (RF) were also utilized to maximize computational efficiency and reduce assembly errors. Four assemblies of varying *k*-mer lengths (21, 23, 35, 40) were generated using the Velvet/Oases pipeline (Velvet v1.2.10 and Oases v0.2.08). Prior to assembly with Velvet/Oases, an *in-silico* normalisation of reads was performed separately on paired and unpaired reads using ORNA (Durai and Schulz, 2017), with a base value setting of 1.3.

The multiple assemblies were merged using EvidentialGene (http://arthropods.eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html), to reduce contig redundancy. The EvidentialGene algorithm first pools transcript contigs together based on predicted coding sequence (CDS) and amino acid sequences, then using the amino acid information selects the highest quality 'protein' and removes all other perfectly matching (redundant) sequences. Contigs failing to reach minimum protein coding requirements are subsequently dropped from the assembly. EvidentialGene then performs a BLAST of the new assembly against itself; pairwise alignment data and the CDS/amino acid models are used to select 'main' (primary

transcripts) and 'alt' (alternate isoform variants) datasets. The main and alt datasets were combined into a reduced redundancy (RR) assembly for downstream analyses.

4.2.4 Reconstruction of unigenes with BALLISTA

The BALLISTA pipeline (Figure 4.2) was used to condense the outputs of the *L. tenue* RR-assembly into unigenes in order to reconstruct a more reliable reference against which mapping and gene expression analyses can be performed. The following steps were implemented through our Ballista pipeline (https://github.com/durhamuniversitybioinformatics). The RR-assembly was aligned through DIAMOND-BLASTX (Buchfink, Xie and Huson, 2015) against the most recent *Arabidopsis thaliana* (TAIR10) proteome; for contigs that found a single or multiple matches the best hit was selected and referred to as *BLAST-hit*, and contigs with no match as *no-hit*. The *BLAST-hit* contigs were then sorted into clusters according to their corresponding locus in the reference proteome, and each cluster was independently fed though the CAP3 Sequence Assembly Program (Huang and Madan, 1999) with the following parameters: end clipping flag 0, gap penalty factor 1, mismatch score factor -1, overlap length cut-off 20, max gap length in any overlap 600, max overhang percent length 500. This reassembled all *BLAST-hit* contigs that match to same locus on the reference sequence, and the resultant groups of contigs ('loci'), generated from reassembled *BLAST-hit* contigs, were then collected together to create a primary 'scaffold'. The *no-hit* contigs were then searched against this primary 'scaffold' using BLASTN in order to incorporate contigs that previously had not found matches due to larger differences from the reference TAIR10 loci. Contigs with new BLAST hits were similarly sorted into clusters with their primary scaffold contigs, and all clusters were again fed independently through CAP3.

The reassembled 'loci' were then merged with all remaining *no-hit* contigs to create a post Ballista assembly composed of reconstructed unigenes.

The completeness of all assemblies were quantitatively assessed by analysing the universal single-copy orthologs contained in the data with BUSCO v3 (Simão *et al.*, 2015), using the embryophyta lineage dataset (downloaded from https://busco.ezlab.org/) and transcriptome run mode.

4.2.5 Reassembly of *A. thaliana* transcriptome

The performance of the BALLISTA pipeline when used with the increasingly distantly-related proteome references was explored by testing *de novo* reassembly of the *Arabidopsis thaliana* transcriptome. *A. thaliana* mRNAseq sequence reads from libraries prepared by Wan et al. (2015) (accessions GSM1868687, GSM1868688, GSM1868689, GSM1868696, GSM1868697 and GSM1868698) were downloaded from the NCBI Gene Expression Omnibus (GEO) (Barrett *et al.*, 2012; Wos and Willi, 2018). Reads were preprocessed and assembled into a single-*k* Trinity assembly using the steps and parameters outlined above.

The initial *A. thaliana* assembly was then reassembled with the BALLISTA pipeline using six angiosperm species with increasing phylogenetic distance as proteome references. These reference species were *Arabidopsis lyrata* (Rawat *et al.*, 2015), *Brassica oleracea* (Liu *et al.*, 2014), *Carica papaya* (Ming *et al.*, 2008), *Populus trichocarpa* (Tuskan *et al.*, 2006), *Aquilegia coerulea* (Aquilegia coerulea Genome Sequencing Project, http://phytozome.jgi.doe.gov/) and *Amborella trichopoda* (Albert *et al.*, 2013). These species represent the angiosperm phylogeny in relation to *A. thaliana* by providing proteome references at the same genus

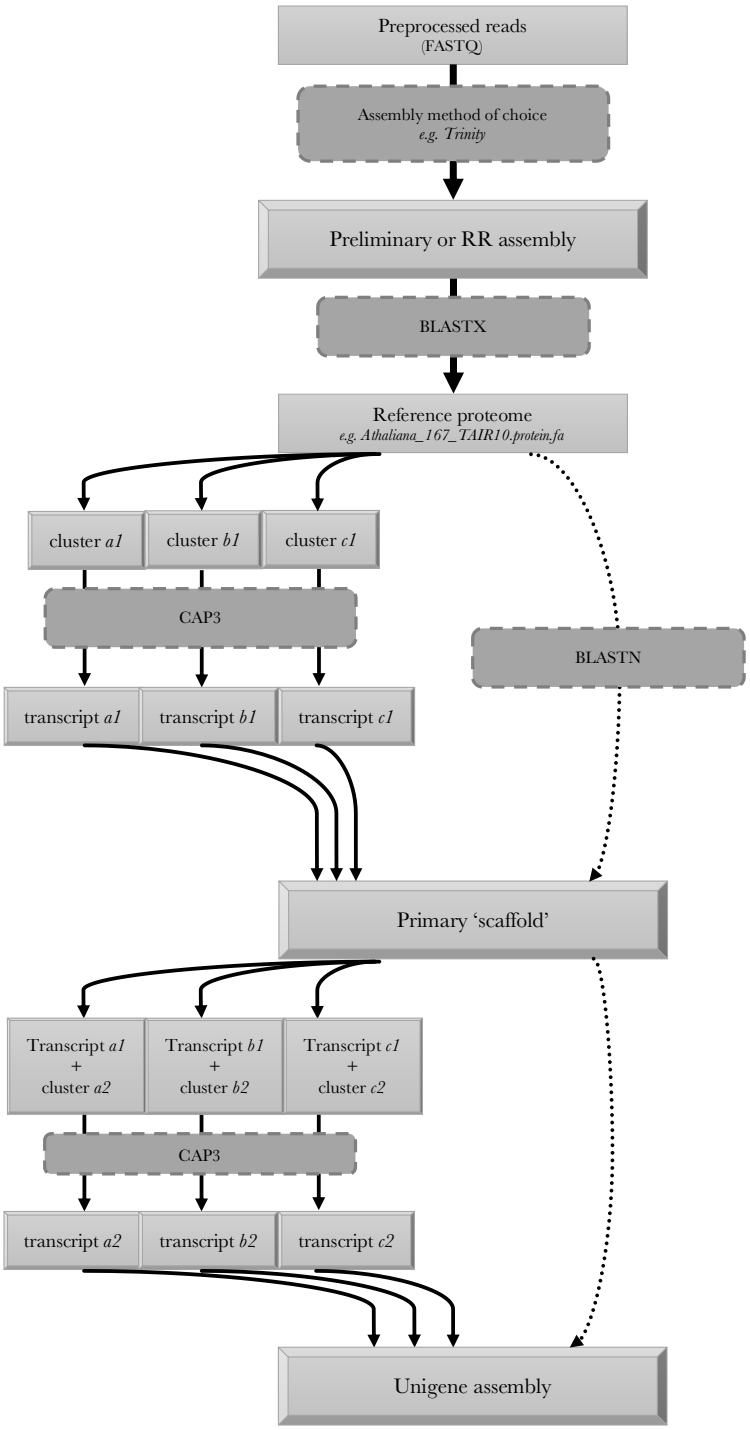


Figure 4.2 A graphical representation of the Ballista pipeline, as automated through Python. Arrows represent the flow of sequence data: from the reference proteome onwards, solid arrows represent contigs from the assembly with a BLASTX/BLASTN match to the reference proteome/primary 'scaffold' (BLAST hit) and dotted arrows represent those without (BLAST no-hit). The user generates an assembly through any method of their own choosing and selects a suitable reference proteome, which are both used as inputs for Ballista. The script then conducts a BLASTX of the assembly against the reference proteome to generate clusters composed of contigs aligning to the same locus. The contigs in each cluster are then assembled as independent groups using CAP3 to produce preliminary 'scaffolds' of unigenes. This scaffold is then used to further merge BLAST no-hit contigs into the assembly through BLASTN to generate new sets of clusters for CAP3 assembly. The final set of unigenes can be merged with any remaing BLAST no-hit contigs to produce the final assembly.

(*Arabidopsis*), family (Brassicaceae), order (Brassicales), super-ordinal clade (Rosids), super-ordinal clade (Eudicots) and super-ordinal clade (Angiosperma) taxonomic hierarchical levels, respectively.

4.2.6 Functional annotation

The post-BALLISTA transcriptome was annotated using a number of functional annotation approaches to construct gene models, using the suite of programs implemented by the Trinotate v3.1.1 pipeline (Bryant *et al.*, 2017). Transdecoder v5.3.0 (http://transdecoder.sourceforge.net/) was used to predict the longest ORFs and these were used to conduct BLASTP searches to the UniProt database (Bateman *et al.*, 2017) for homology searches. Subsequently, protein domains were identified using HMMER v3.2.1 (Eddy, 2011) and signal peptides were predicted using signalP v4.1 (Petersen *et al.*, 2011). The transcriptome was also searched for homology to transcription factor loci by BLASTX searching the *Arabidopsis thaliana* and more closely related *Linum usitatissimum*, *Salix purpurea* and *Poplulus trichocarpa* transcription factor databases downloaded from PlantTFDB v4.0 (Jin *et al.*, 2017). Bidirectional best-hit BLAST searches were also conducted to *Linum usitatissimum*, *Arabidopsis thaliana*, *Salix purpurea* and *Poplar trichocarpa* proteomes using our in-house reciprocal best-hits tools written in python (https://github.com/durhamuniversitybioinformatics). Finally, the NCBI Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2015) was searched for homologous domains using RPSTBLASTN. These custom BLAST searches were merged into final the Trinotate annotation report, and used to create a general feature format (GFF3) ancillary annotation file. Gene Ontology (GO) annotation results were plotted using WEGO v2.0 (Ye *et al.*, 2006, 2018), according to their hierarchical GO level.

4.3 Results

4.3.1 *Linum tenue* sequencing and initial *de novo* transcriptome assemblies

The 24 *L. tenue* cDNA libraries of different tissue types (leaf, young bud, immature flower, open flower) generated a total of 803,516,244 pre-processed reads (401,622,055 forward and 401,894,189 reverse), of which 803,155,300 (99.96%) were paired. Following pre-processing, reads had an average Phred quality score of ~36 and ranged in size from 35-126 bp, with a mean of 91.42 bp.

The summary statistics of the initial assemblies using a single-*k* Trinity and multiple-*k* Velvet/Oases approach are shown in Table 4.1. The reduced redundancy (RR) assembly was generated through pooling and condensing the 946,900 contigs produced from the separate initial assemblies using EvidentialGene. The RR assembly was 157.80 Mbp in length with an N50 of 708 bp, comprised of 249,320 contigs ranging from 209 to 16,770 bp in length.

Mapping the preprocessed reads back to the initial assemblies (Table 4.1) using STAR v2.3.0 (Dobin *et al.*, 2013) yielded mapping rates ranging from 43-90%, with the proportions of reads mapping to multiple contigs ranging from 21-55%. The Trinity assembly had the highest rate of reads mapping to unique transcriptomic locations (35%). The RR assembly showed demonstrably improved mapping statistics, with a 56% mapping rate for reads aligning to unique locations.

The BUSCO analyses demonstrated assembly completeness ranging from 23-27% in the Velvet/Oases assemblies to 87% in the Trinity assembly (Table 4.1). Of the universal single-copy orthologs present in the Trinity assembly, 20% were complete single-copies,

51% were complete duplicates and 15% were fragmented. The RR assembly showed a marked increase to 46% in the proportion of single-copy orthologs.

3.3.2 Assembly condensation and unigene reconstruction with BALLISTA

The RR assembly and *A. thaliana* TAIR10 proteome were used as inputs for BALLISTA, which ran to completion in ~2 hrs. The resultant BALLISTA reassembly (B-reassembly) (Table 4.1) was 114.19 Mbp in length, comprised of 106,105 contigs derived from reassembled clusters of transcripts matching an *A. thaliana* reference locus, and a further 64,537 contigs that did not match (a total of 170,642 contigs). Contigs lengths of the B-assembly ranged from 209 to 16,770 bp in length, with an N50 of 789 bp; the contig size distribution is presented in Figure 4.3.

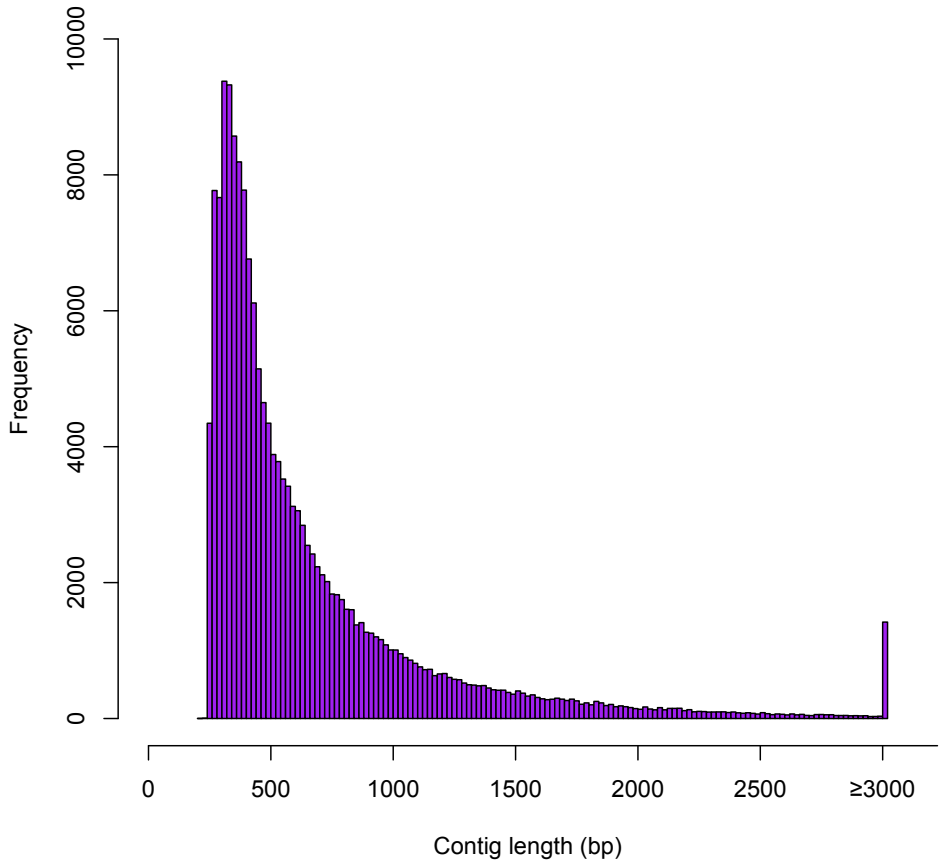


Figure 4.3 Distribution of contig lengths in the BALLISTA-assisted *Linum tenue* transcriptome assembly. Bars represent counts of contigs in each length category.

The B-reassembly yielded the best mapping results (Table 4.1), with an overall mapping rate of 82%, with 67% of reads mapping to unique transcriptomic locations and 15% showing multiple mapping. Of the mapped reads, the mean mapping quality score was 149.06, and the mean transcriptome-wide coverage was ~667X following a uniform distribution (Figure 4.4). Similarly, the BUSCO analysis of the B-reassembly (Table 4.1) demonstrated the best scores when compared to each individual initial assembly and the RR-assembly, with a total of 86% of single-copy orthologs found. Of these, 58% were complete single-copies, 12% were complete duplicates and 16% were fragmented.

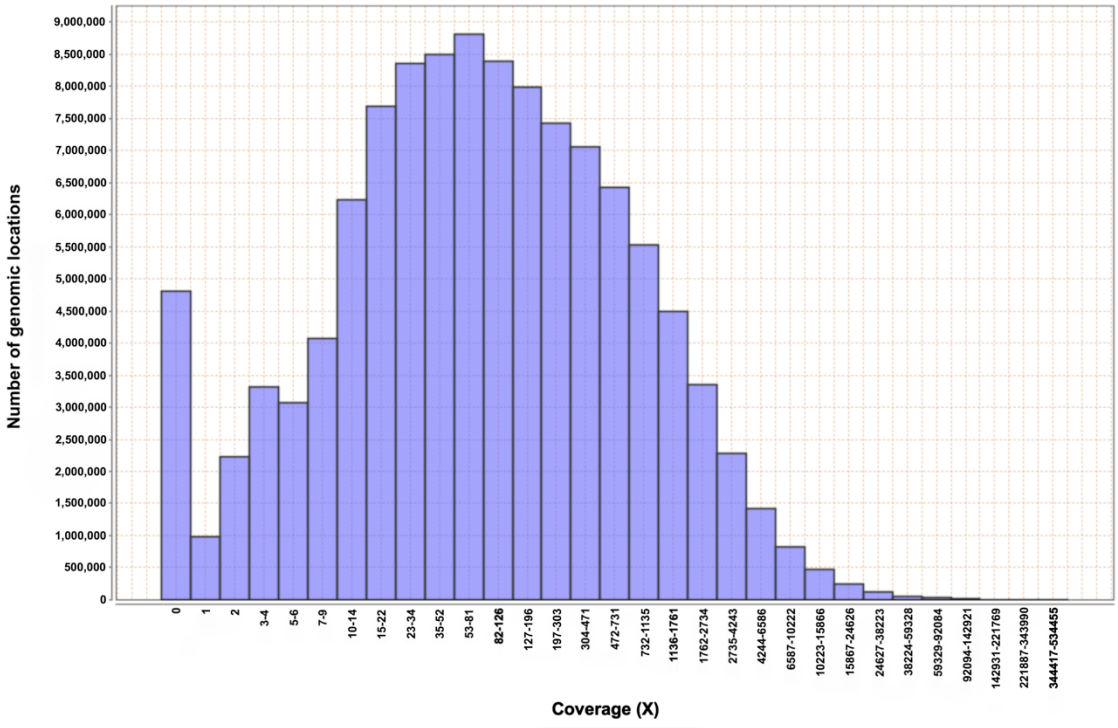


Figure 4.4 Distribution of read coverage across the *L. tenue* transcriptome.

Table 4.1 Summary statistics of the assembly, read mapping and BUSCO analyses.

	Oases k21	Oases k23	Oases k35	Oases k40	Trinity k25	Reduced Redundancy	BALLISTA Reassembly
Assembly Statistics							
Assembly length (Mbp)	88.67	87.95	68.92	64.64	250.03	157.80	114.19
No. contigs	183,954	178,271	139,975	131,686	313,014	249,320	170,642
Smallest contig length (bp)	197	220	248	250	251	209	209
Longest contig length (bp)	4,883	4,708	4,867	5,507	16,770	16,770	16,770
Average contig length (bp)	482.01	493.36	492.35	490.85	798.79	632.92	669.15
N50 (bp)	512	527	526	523	1,074	708	789
Read Mapping							
Uniquely mapped reads (%)	18	20	25	27	35	56	67
Multiply mapped reads (%)	25	26	22	21	55	27	15
Unmapped reads (%)	57	55	53	53	10	17	18
BUSCO Analysis							
Complete single-copy BUSCOs (%)	3	2	2	4	20	46	58
Complete duplicate BUSCOs (%)	2	2	3	3	51	24	12
Fragmented BUSCOs (%)	19	18	21	21	15	16	16
Missing BUSCOs (%)	77	77	75	73	14	14	14

4.3.3 Comparison of *A. thaliana* BALLISTA reassemblies

The initial *A. thaliana* assembly (*A. thaliana* IA) was comprised of 41,377 contigs, reaching 51.57 Mbp in length (Table 4.2). Mapping the *A. thaliana* preprocessed reads to the *A. thaliana* IA yielded an 96% overall mapping rate, with 54% mapping to unique and 42% mapping to multiple transcriptomic locations. A BUSCO analysis of the *A. thaliana* IA showed 8% of the total expected single-copy orthologs were missing; and of the single-copy orthologs that were present 66% were complete single-copies, 24% were complete duplicates and 3% were fragmented (Table 4.2).

The subsequent BALLISTA reassemblies of the *A. thaliana* IA showed similar results, with a uniform condensation of the number of contigs, ranging from 31,155-31,927, using each of the reference transcriptomes regardless of phylogenetic distance (Table 4.2). The total mapping rate remained at 96%, but displayed improvements in the multi-mapping rates, decreasing from 42% to 24-26% (Table 4.2). Similarly, this was reflected in the BUSCO analyses, where the number of complete duplicated single-copy orthologs decreased from 24% to 7-8% (Table 4.2).

To confirm the quality of the reassembled unigene sequences, contigs were compared to the *A. thaliana* genome using BLASTN (Table 4.2). All references post-processed with BALLISTA retained >99% match rate to the *A. thaliana* genome. When filtered for high-fidelity matches, where the pairwise alignment had a percentage identity >90% and covered >60% of the of the unigene contig sequence, the BALLISTA post-processed references performed just as well as the Trinity assembly, though with slightly higher rates of matches.

4.3.4 Functional annotation of the *L. tenue* transcriptome

The features as identified in the Trinotate pipeline were combined into a gene feature format (GFF3) annotation. Of the 170,642 contigs in the final *L. tenue*, 133,358 (78%) could be annotated for at least one feature. The 37,478 annotated contigs with associated GO terms were distributed among 35 subcategories across the three primary GO functional categories: biological processes (21,180, 57%), cellular components (8,197, 22%), and molecular function (33,204, 89%), as shown in Table 4.3 and Fig. 4.5. Within the biological processes category, the dominant groups of subcategories were metabolic process (43.2%) and cellular processes (31.3%). Within the molecular functions category, the dominant subcategories were binding (56.9%) and catalytic (45.1%), and within the cellular components category, they were cell (12.2%), cell part (12.2%) and membrane part (11.4%) (Fig. 4.5a).

Of the unannotated contigs, a total of 155 contigs were found to have bidirectional best-hit BLAST matches with three of the tested related species and with *A. thaliana* (TAIR10) (Fig 4.6). Thirty-three contigs had bidirectional best-hit BLAST matches with all query species, whilst unique bidirectional best hits were also found in *L. usitatissimum* (43), *P. trichocarpa* (12), *S. purpurea* (15) and TAIR10 (9).

Table 4.2 Summary statistics of the *Arabidopsis thaliana* reassemblies, read mapping and BUSCO analyses. The first data column (*A. thaliana* IA) presents data for the initial *A. thaliana* assembly as created using the single-*k* Trinity method. All subsequent columns present the data for separate BALLISTA reassemblies of the *A. thaliana* IA data using the listed species as proteomic references. For the TAIR10 genome (chromosomes), only the relevant statistics of read mapping are provided for comparison purposes.

	***A. Thaliana* IA**	***A. lyrata***	***B. oleracea***	***C. papaya***	***P. trichocarpa***	***A. coerulea***	***A. trichopoda***	**TAIR10_chromosomes**
Assembly Statistics								
Assembly length (Mbp)	57.51	42.46	43.36	43.36	43.41	43.13	43.21	-
No. contigs	41,377	31,155	31,982	31,927	32,083	31,804	31,660	-
Smallest contig length (bp)	251	251	251	251	251	251	251	-
Longest contig length (bp)	13,019	13,019	13,019	13,019	13,019	13,019	13,019	-
Average contig length (bp)	1,389.83	1,362.88	1,355.87	1,358.14	1,353.08	1,356.00	1,364.68	-
N50 (bp)	1,801	1,776	1,773	1,774	1,768	1,773	1,786	-
Read Mapping								
Uniquely mapped reads (%)	54	71	70	70	70	70	70	96
Multiply mapped reads (%)	42	24	25	26	25	25	25	1
Unmapped reads (%)	4	4	4	4	4	4	4	3
BUSCO Analysis								
Complete single-copy BUSCOs (%)	66	78	78	77	78	78	78	-
Complete duplicate BUSCOs (%)	24	8	8	8	8	7	8	-
Fragmented BUSCOs (%)	3	6	6	6	6	6	6	-
Missing BUSCOs (%)	8	8	8	8	8	8	8	-
BLASTN to TAIR10 genome								
No. contigs with BLAST matches	41,361	31,141	31,968	31,913	32,070	31,790	31,646	-
No. contigs with BLAST matches (%)	99.961	99.955	99.956	99.956	99.959	99.956	99.956	-
No. high-fidelity BLAST matches	16,520	12,997	13,409	13,413	13,516	13,398	13,360	-
No. high-fidelity BLAST matches (%)	39.93	41.72	41.93	42.01	42.13	42.13	42.20	-

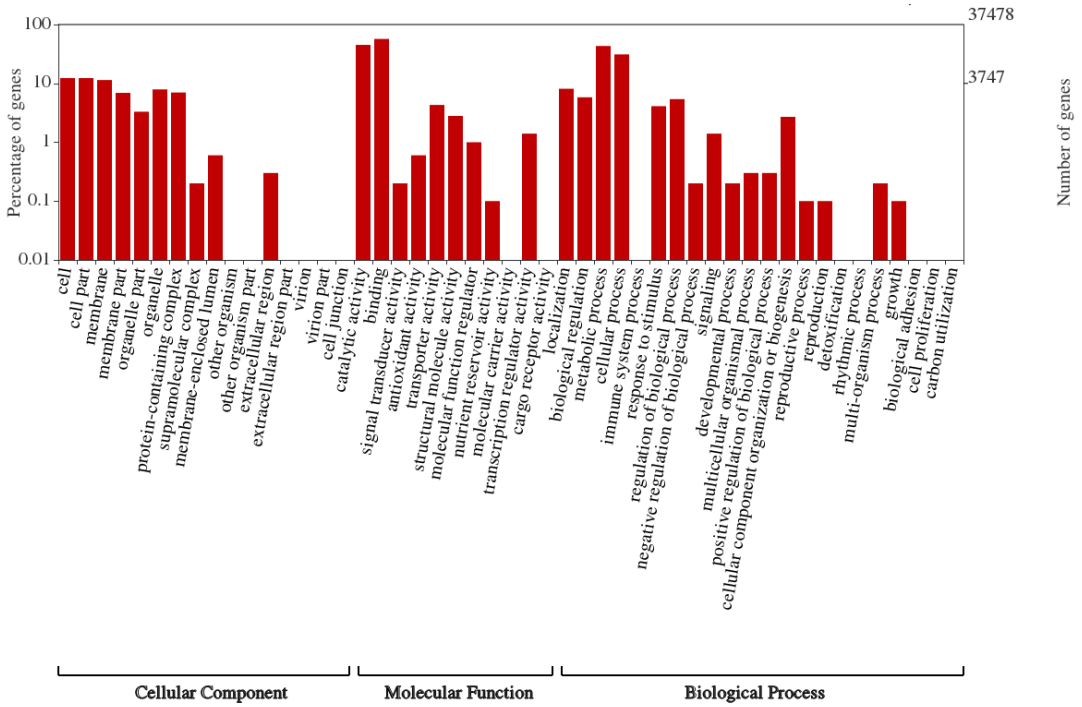


Figure 4.5 GO annotation of the *L. tenue* BALLISTA reassembly. A total of 37,478 GO terms were derived from the annotated contigs and classified into the three major categories of the GO hierarchy (cellular component, molecular function, and biological process) and 35 subcategories.

Table 4.3 Summary of analysis of GO terms using WEGO.

		L. tenue
Annotated Features		37,478
GO Terms	Biological	21,180
	Cellular	8,197
	Function	33,204
	Total	62,581

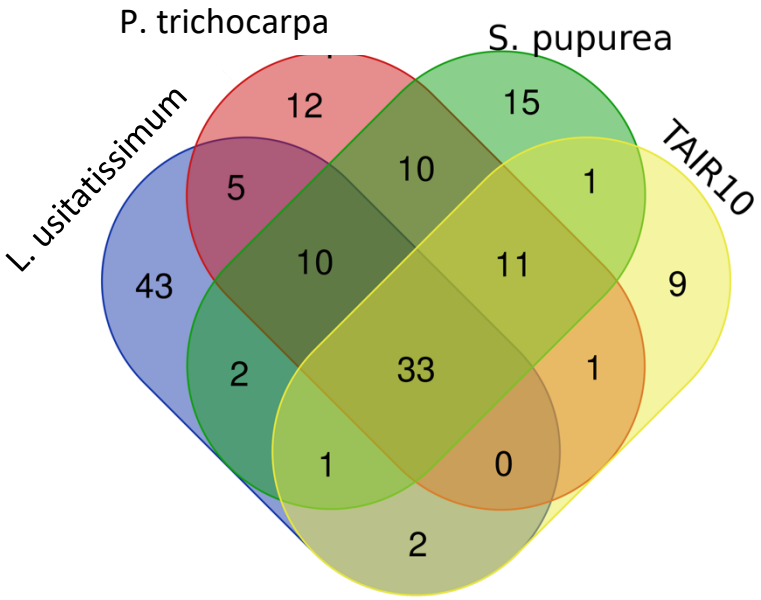


Figure 4.6 Venn diagram of the unannotated contigs and their bidirectional best-hit BLAST matches with related reference species (*Linum usitatissimum*, *Populus trichocarpa*, *Salix purpurea* and *Arabidopsis thaliana* (TAIR10).

4.4 Discussion

The availability and reduced cost of NGS technology has led to transcriptomic analyses in an increasing range of species (Todd, Black and Gemmell, 2016). However, with the advancement of technology there is often a lag in the development of the tools that facilitate its implementation to realising its full potential, and RNAseq studies in non-model organisms still rely heavily on the quality of the reference *de novo* transcriptome. Here we present the BALLISTA pipeline as an alternative post-assembly processing method to improve the quality of a *de novo* mRNA transcriptome assembly, even with distantly-related reference species. BALLISTA uses proteomic information from the proteome of a related reference species to reassemble clusters of unigenes and allelic variants into an accurate reference transcriptome, as demonstrated by our *de novo* transcriptome assemblies of *Linum tenue* and *Arabidopsis thaliana* from Illumina mRNAseq datasets.

4.4.1 Pre-processing of the FASTQ reads

The use of *k*-mers (sub-sequences of the sequence reads of length *k*) in the construction and extension of de Bruijn graphs in modern short-read assembly algorithms can lead to assemblies with high numbers of contigs, representing only fragments of the parent transcripts. Sequencing errors along with sequence attributes, such as polymorphisms, sequence repeats and poor coverage of lowly expressed transcripts, have been shown to contribute further to this fragmentation (Heo *et al.*, 2014; Song, Florea and Langmead, 2014; Li, 2015). Sequencing errors are often the leading cause of rare *k*-mers (Song and Florea, 2015), and it is thus good practice to correct or remove them. Rcorrecter is a fast, efficient, high-accuracy tool for sequence error correction designed to handle the variance in coverage and alternative isoforms present in RNAseq data. Erroneous reads that cannot be corrected are also flagged for removal, which can be done with simple scripts.

Examination of the sequence read quality, particularly for adapter contamination post-demultiplexing, which is a common occurrence, should also be an important consideration when aiming to produce a high quality contiguous, error-free, and complete transcriptome. However, it should be noted that aggressive trimming through the use of high quality thresholds can be suboptimal and adversely affect an assembly (MacManes, 2014). TrimGalore! is a trimming tool that has been shown to be more effective at removing adapter sequeces than comparable programs such as Trimmomatic both in the literature (Stubbs *et al.*, 2017) and in our dataset (personal observations).

The final step in the pre-processing pipeline should be to filter the *k*-corrected and trimmed reads against a custom ribosomal RNA database, created through a combination of large subunit, small subunit and 5S ribosomal RNA databases. This is particularly

recommended to researchers using only poly-A capture methods to remove abundant rRNA, as resultant libraries can still be highly composed of ribosomal RNA. Whilst the Illumina TruSeq Stranded mRNA kits utilise both ribo-depletion and ploy-A capture methods, we would recommend filtering residual ribosomal reads out of the dataset to ensure that the assembly contains only mRNA transcripts. The reads resulting from this process were subsequently assessed and were shown to be of demonstrably high quality (Phred quality score >30).

Pre-processing of the raw reads beyond 3' trimming and adapter removal is rarely suggested in the literature, and we urge researchers to consider this or similar pre-processing pipelines for RNAseq investigations.

4.4.2 Multiple *k*-mer approach for the *de novo L. tenue* transcriptome assembly

The creation of a merged RR assembly from multiple *k*-mer values and different assemblers was a useful transcriptome assembly method supported by our results. Software packages for *de novo* assembly offer a whole suite of parameters that can be adjusted to optimise the resultant assembly. The *k*-mer length is the main parameter that is often focused on as it has one of the largest effects on the structure of the de Bruijn graph. Choosing the right value of *k* can be a subjective endeavour due to trade-offs between low and high values: low values of *k* increase the number of reads that contribute to the graph and thus improve coverage, while high values of *k* are more specific and thus less sensitive to sequencing errors, producing more contiguous and error-free assemblies. Approaches combining assemblies of varying *k*-mer length (Surget-Groba and Montoya-Burgos, 2010) have thus been proposed to maximise both coverage and accuracy. However different assembly tools behave differently in the way they handle the variable

transcript coverage inherent in RNAseq data, and the choice of assembler has a large impact of the assembly. For example, a Velvet/Oases pipeline has been shown to be better at identifying novel isoforms, whilst Trinity was shown to be better at producing contiguous assemblies from low read coverage (Zhao *et al.*, 2011). A number of recent studies have thus started merging multiple assemblies from different assembly tools and varying *k*-mer lengths (Nakasugi *et al.*, 2014; Orsini *et al.*, 2016; Sales *et al.*, 2017).

The data for our initial assemblies (Table 4.1) strongly support this approach. Obtaining assembly statistics are useful for descriptive purposes but can be misleading for evaluation, as, in contrast to genomes, transcriptomes are by their nature more fragmented and composed of shorter sequences. Evaluation of transcriptome assembly quality should therefore focus more on how well it is supported by the original reads, and how complete the coverage is in terms of the presence of expected transcripts, as opposed to the use of statistics such as the N50 (which is a measure of contig size distribution). The use of BUSCO scores to evaluate transcriptome assembly completeness are particularly useful, as they indicate both the coverage of a transcriptome and the quality of the assembly of its protein-coding elements. Our *L. tenue* RR assembly demonstrated better read support and completeness scores overall than any of the initial assemblies individually, as indicated by improved rate of uniquely-mapped reads and the proportion of complete single-copy universal orthologs. The slight decrease in overall mapping rate in the RR assembly compared to the Trinity-*k*25 assembly (90 to 83%) can be explained as a consequence of the condensation of redundant transcripts that would require additional adjustment of the alignment parameters to take account of allelic mismatches and splice junctions.

Of interest in our analysis is the low coverage of the Velvet/Oases assemblies in terms of their BUSCO scores (ranging from 33-37% completeness). This result is largely explained by the normalisation steps performed on the sequence data using ORNA prior to

assembly. The Velvet assembler is notorious for its difficulties in resolving nodes of the de Bruijn graph with high coverage (Zhao *et al.*, 2011), commonly leading to computational failures in the assembly process (personal observations from our investigations and from collaborators). ORNA is thus designed to be overzealous in its normalisation of high-coverage *k*-mers, which can lead to a large reduction of the overall input dataset. Our results indicate that this can come at the cost of complete and contiguous assemblies. However, the slight increase in fragmented BUSCO scores between the Trinity-*k*25 and RR assembly (15-16%) indicate that the Velvet/Oases pipeline assembled some transcripts that Trinity was unable to. This supports the hypothesis that Velvet/Oases is more effective at assembling novel or lowly expressed isoforms than Trinity, although improvements to the normalisation process are warranted.

4.4.3 Assembly of consensus unigene sequences with BALLISTA

Further improvements to the quality of the *L. tenue* transcriptome assembly were made by applying the BALLSITA pipeline. In the absence of a genomic reference, the uncertainty produced through the multiple-mapping of reads to different locations on the transcriptome is a ubiquitous problematic issue inherent in short-read RNAseq. Small reads representing fragments of sequence may map to multiple locations if their parent transcripts share exons through alternate-splicing or conserved protein functional domains. Promising tools are now available that aim to reduce the issues associated with allocating multi-mapped reads to a transcriptomic reference through Bayesian quantification approaches (Li *et al.*, 2010; Leng *et al.*, 2013) or the stitching together of related sequences into 'super transcripts' (Davidson, Hawkins and Oshlack, 2017). However, in our experience, there can still be great difficulty in the analysis if a *de novo* transcriptome assembly has not been adequately post-processed to remove the

redundancy of allelic variants. Our results indicate that the BALLISTA pipeline as applied to the RR assembly is effective at reducing the issues associated with multiple mapping, as demonstrated by the 11% increase in uniquely aligned reads, and the 12% increase in the presence of complete single-copy BUSCOs (Table 4.1).

The use of translational information in the reconstruction and condensation of unigenes is another strength of the BALLSITA pipeline for use on mRNAseq data. Building on approaches proposed by Surget-Groba & Montoya-Burgos (2011), BALLISTA demonstrates improved effectiveness at increasing the contiguity and incorporating redundant contigs by removing the assumption that the transcripts being reassembled must resemble the reference proteome in organisation and structure. The reassembly of the *A. thaliana* transcriptome from reference proteomes of varying phylogenetic distance demonstrate that even distantly-related species can be used as an effective reference, with BALLISTA consistently improving unique mapping rates by 26-27%, and increasing the proportions of complete single-copy BUSCOs by 11-12% (Table 4.2). Through reassembling all transcripts that comprise a unigene, as opposed to only reassembling regions that partially overlap positionally, BALLISTA also successfully incorporates smaller contigs that are fully contained within larger ones. This is more effective for the condensation of allelic-level information and allows distantly-related species to be used as a reference if more closely-related proteomes are unavailable. The use of BLASTX to cluster transcripts by matches to protein reference loci is a biologically accurate method for unigene clustering as it matches transcripts to known proteins. This method is less arbitrary than, and preferable to, clustering transcripts based on their relationships with each other within the assembly; as was attempted during the *de novo* transcriptome assembly of *Orchis italica* (De Paolo *et al.*, 2014), using CD-HIT (a clustering tool) set to 85% sequence identity. Such approaches run the risk of merging loci from related gene families.

4.5 Concluding remarks

The concept of condensing all the transcripts comprising a unigene into a consensus transcriptomic sequences is not a new practise. Liang *et al.* (2000) and Pertea *et al.* (2003) developed methods which constructed clusters of unigenes from EST data, through pairwise sequence similarity, which were then individually reassembled using CAP3 or similar tools. The employment of similar methods have been used more recently to resolve redundancy issues in the *de novo* transcriptome assembly of the tetraploid *Nicotiana benthamiana* (Nakasugi *et al.*, 2014). The modifications to the CAP3 parameters in the our pipeline have been informed by these studies (Pertea *et al.*, 2003; Lee *et al.*, 2004), and BALLISTA revives the method for use in the NGS era.

- Chapter 5 -

General Discussion

The tenuous jungle. Image by Alireza Foroozani

5.0 General Discussion

5.1 Synthesis

This thesis has investigated the morphological nature and transcriptomics of distyly in *Linum tenue*, with the aim of furthering our understanding of the molecular genetic basis of heterostyly in the Linaceae. This work has illustrated the complexities of the distyly syndrome in *L. tenue*, highlighting the functional differences between pin and thrum traits, possible developmental mechanisms responsible for differences in tall and short style heights in pin and thrum morphs, and indicating stages of development in which flowers begin their morph-specific trajectories. This work underlines the significance of developing a high-quality, reliable *de novo* transcriptomic reference and presents a method by which this can be achieved. This method has been used to create a vegetative and floral transcriptome reference for *L. tenue*, which is an extremely useful resource for the *Linum* research community. Furthermore, this project has investigated patterns of differential expression between pin and thrum flowers to provide a list of putative candidate transcripts (Appendix III) for loci that comprise the *S* locus, are *S* locus-linked, or are controlled by the *S* locus of *L. tenue*.

The findings of this project have implications at a range of scales in different fields. In this section these will be discussed in light of key themes: *(i)* the value of constructing a high-quality reference transcriptome, *(ii)* new insights into the expression of distyly in *L. tenue*, *(iii)* evolution of pin and thrum morphs of *L. tenue*, and *(iv)* the *S* locus of *L. tenue*. In Section 5.2 I will present the gaps and questions raised by this research, suggesting avenues for improvements and further work; and finally, in Section 5.3 I will outline my thoughts and predictions for the directions that the field of bioinformatics will take in the future.

5.1.1 The value of constructing a high-quality reference transcriptome

In Chapter 4.0, we presented a method to reconstruct consensus unigene sequences from the mRNA sequences that are assembled with contemporary transcriptome assembly software tools. From combining numerous assemblies generated from different software tools at a range of different parameters, we found that a combined multiple *k*-mer and reduced redundancy approach resulted in a transcriptomic reference that was of higher quality than individual references assembled from a single *k*-mer or a single software tool alone. Further improvements to the quality of this reference was made through implementation of the BALLISTA pipeline, a user-friendly automated tool we have developed to reduce the allelic and alternate splice variants commonly identified as transcript isoforms in the raw outputs of assembly tools.

Deriving useful biological information from a transcriptome without a genomic reference is challenging. Researchers need to create a reference transcriptome as a catalogue of the transcripts present in the study samples, which can be done using a genome-guided approach (with the use of a related reference genome) or through *de novo* assembly of the sequenced cDNA reads. Given that using a hetero-specific reference as a template focuses on reassembling reads that map to known transcript models, and declines in performance with increasing sequence divergence (Vijay *et al.*, 2013), we elected to use a *de novo* assembly approach. However, the resultant fragmented and high-complexity (i.e. containing alternate-splice and allelic variants) references that are commonly assembled can be of poor quality (Chang, Wang and Li, 2014), which naturally impacts the quality of downstream expression analyses.

Through the use of translational information, BALLISTA uses the proteome of a reference species to reduce the complexity and redundancy of the *de novo* transcripts by reconstructing consensus unigene sequences without relying on assumptions of synteny with the reference. Only reassembling the transcripts that match to a single locus on the reference, the BALLISTA pipeline also reduces the risk of falsely merging transcripts from closely-related loci. Post-processing of assemblies with the BALLISTA pipeline produced references that were of demonstrably higher quality, with higher rates of unique alignments of the original sequence reads, and increased assemblies of universal single-copy orthologs. The efficacy of the BALLISTA pipeline to function consistently even with distantly-related references was also demonstrated through the *de novo* assembly of the *Arabidopsis thaliana* transcriptome from mRNAseq sequence data, similarly improving the rates of unique read alignment and universal single-copy orthologs whilst retaining high-fidelity matches of the unigene sequences to the *A. thaliana* genome.

Retaining allelic and alternate-splice information may be a point of interest for some RNAseq studies, and this information is retrievable from the BALLISTA process. For each consensus unigene sequence the alignment of its constituent input contigs is known. However, in our study focused on finding candidate loci, isoform-level information was not a priority. Furthermore we found the reduced levels of multimapping during the alignment stage, and reduction of information at the differential expression-level, was more helpful for our analyses.

We would also like to highlight to researchers in the field the importance of pre-processing the sequence reads prior to assembly and downstream analyses. Bias and errors in the library preparation and sequencing processes will have negative impacts on reference generation and transcript quantification. Implementation of such steps are not commonly

reported in the literature, and we urge researchers to follow the pre-processing steps we have taken, or similar approaches, when using short-read Illumina data.

5.1.2 New insights into the expression of distyly in *L. tenue*

This work has contributed significantly to our knowledge of the ecology and the heterostyly syndrome in *L. tenue*. We made use of the new adaptive inaccuracy measures of heterostylous reciprocity, proposed by Armbruster *et al.* (2017), to reveal interesting patterns in the functional expression of distyly at the population level (Table 2.3). This method uses equal positioning of the reciprocal stigmas and anthers as the optimum phenotype, and extent of phenotypic deviations in these organs from the optimum (inaccuracy) that can be measured in terms of the mean departure of floral organ position from the optimum (bias) and intrapopulation variance in each organ height (imprecision). We found a greater bias for short reproductive organs, generally as a result of shorter thrum pistils than pin stamens, and greater imprecision for tall reproductive organs. The developmental data from our study indicates that the greater imprecision in tall organs result from greater variance in pin pistil height in the later stages of flower development.

That there are differences and differing developmental mechanisms between the different morphs and organs is a significant finding, and provides strong evidence that heterostyly in *L. tenue* is a pleiotropic trait with different determinants for the male and female sexual organs, as is consistent with the expectations of an *S* locus-controlled regulation of heterostyly (Lewis and Jones, 1992). The reduced imprecision in the length of thrum pistils suggests that development of the organ is under tighter genetic control. Our findings thus suggest that reduced cell elongation in the thrum style tissues leads to increased robustness of the trait. Our results also suggest different levels of selection pressure for tall and short

organs. It is commonly reported in the literature that the positioning of taller organs at the opening of the flower increase the likelihood of contact with pollinators than the shorter organs, suggesting there is less pressure for precise positioning. This could have implications for evolutionary transitions to selfing or dioecy if the thrum flower begins to suffer from reduced female fitness.

5.1.3 Evolution of pin and thrum morphs of *L. tenue*

Analyses of differential gene expression between pin and thrum morphs of *Linum tenue* flowers in Chapter 3.0 revealed a list of features that are representative of morph-specific transcriptional activity. This list is an important first step towards finding candidate loci for the *L. tenue S* locus. Analysis of these differentially expressed features uncovered interesting expression patterns between the morphotypes, which, if representative of the underlying biology, have very interesting implications for our understanding of the molecular genetics and evolution of heterostyly in *Linum*.

Research in the mating system of *Linum* species is currently very active (Ushijima *et al.*, 2015; Kappel, Huu and Lenhard, 2017), with a number of groups working to identify the location and architecture of the *S* locus. The current evidence suggests that heterostyly in *Linum grandiflorum* exhibits hemizygous determination of the thrum morph (Ushijima *et al.*, 2012), similar to the genetic control observed in *Primula* (Nowak *et al.*, 2015). However, without characterisation of *S* locus allelic variants, or an understanding of the dominance or epistatic interactions of *S* locus alleles, little is known regarding the behaviour of loci comprising or controlled by the *Linum S* locus. Our findings have implications for the molecular genetics of distyly in *L. tenue*, and *speculatively* provide some support for a model of the evolutionary development of heterostyly in the species.

Our results for differences in cell lengths for style and stamen filament organs between pin and thrum morphs (Chapter 2.0) is parsimonious with the model of heterostyly evolution that distyly developed from approach herkogamous flowers (pin morph) (Lloyd and Webb, 1992a, 1992b). Here, an evolutionary event, possibly a gene duplication as proposed in *Primula* (Li *et al.*, 2016), leads to extension of the stamen filaments to the height of the style in a long homostyle morph. Over time, as more genes are recruited to the *S* locus, the length of the style in this new long homostyle decreases and results in the thrum morphs we see in populations today. Our finding that the short style of the thrum is largely developmentally driven by reduced cell elongation in the tissues of the style, whilst the long stamen filaments putatively develop from an increase in cell number in the tissues of the filament (as cells of the pin style, pin filament and thrum filament are of relatively comparable length) (Figure 2.3), is consistent with this evolutionary model. Our cell length data suggest at least two different mechanisms for the development of long and short sexual organs in the two morphs, consistent with two separate evolutionary events: first the creation of long stamen filaments, then a short style.

The morph-specific differences in observed gene expression patterns in our analyses also fit with this model of distyly evolution. Differentially expressed features show strong grouping by morphotype; but the pin group shows strong structural subgrouping by individual, whereas this pattern is much less apparent in the thrum group (Figure 3.7, Figure 3.8). This suggests a higher variance in expression *between* pin individuals and a higher variance in expression *within* thrum individuals. If this difference is a result of greater allelic diversity in pin *S* locus-linked loci than in thrum *S* locus-linked loci, a younger evolutionary origin of the thrum morph (with less time to accumulate allelic variants in functional genes or genic regions) provides a reasonable explanation for this subgroup patterning.

5.1.4 The *S* locus of *L. tenue*

Until recent years, our understanding of the genetic basis and functional genetic basis of heterostyly remained elusive. Modern sequencing technologies have made allowed models to be made from non-model organisms, and *Primula* has become the flagship system for understanding heterostyly. However, independent evolutionary emergences of heterostyly have occurred across the angiosperm phylogeny; though the *S* locus may be under constraints that can dictate which genes are recruited, the genetic determinants and mechanisms between taxa are unlikely to be completely equivalent. That of the five *L. grandiflorum* thrum-specific loci found by Ushijima *et al.* (2012) we found only *TSS1* to be differentially expressed is demonstrative of this. Even though *L. grandiflorum* is a con-generic, according to certain phylogenies (McDill *et al.*, 2009) it may represent an evolutionary acquisition of heterostyly within the Linaceae independent of *L. tenue*. Though some elements of the *S* locus may be shared, we might expect the genomic architecture and constitutive determinants to be different. We were also unable to find a homolog for *TSS1* in the newly published genome of *Primula vulgaris* (Cocker *et al.*, 2018), though publication of its predicted protein sequences may change this. Nonetheless, the genetics of heterosty in *Linum* remains very much an open field. Though with the Ushijima group in Okayama chasing the functional genetics in *L. grandiflorum*, and the Slotte Lab in Stockholm sequencing the genomes of three heterostylous *Linum* species, this may soon change.

The discovery of Contig_141165 as a candidate gene for the *G* locus is a significant contribution to the field. We not only demonstrate a late-development thrum-specific signature of expression, we present evidence that Contig_141165 is divergent from other

putative paralogs within the genome. Furthermore, we show that Contig_141165 is likely a homolog of *VUP1* in *A. thaliana*, a gene of unknown function that has dramatic pleiotropic events on multiple pathways (Grienenberger and Douglas, 2014), and functions as a dominant inhibitory element. The cell length microscopy we have conducted revealing reduced cell elongation in the tissues of the thrum style further support our synthesis that Contig_141165 may have a similar effect in *L. tenue*. To our knowledge, this thesis represents the only work to recognise the work done by Grienenberger and Douglas (2014) realising the downstream effects of *VUP1*, if not its direct function, and reporting a homolog in *L. tenue* with similar (putative) activity in style tissue.

Further investigations into Contig_141165 must be done in order to define it as a *S* locus constituent. Knowledge of the genomic sequence will greatly aid subsequent functional genetics work to confirm its activity and effects of expression. Characterising the protein, inferring and discovering its function will be a significant task, as Contig_141165 and its homologs display very low levels of similarity and no functional or structural domains have been identified so far. This demonstrates the limitations of bioinformatics for various tasks, as functional genetic studies remain the strongest avenue for the foreseeable future to characterise Contig_141165 and its activity. This is a challenge the authors welcome to receive.

5.2 Further work and improvements

5.2.1 The BALLISTA pipeline

Whether used on a reference produced from a single *de novo* assembly or a redundancy-reduction tool such as EvidentialGene (Section 4.2.2), the BALLISTA pipeline combines

alternate-splice variants into consensus unigene sequences. Our tool could be further developed to produce ancillary files that detail the location of these 'features' within the unigene sequences. This information is present in the temporary files that BALLISTA creates, but a consolidated report detailing which of the original contigs constitute each of the unigene sequences could be useful to users. Similarly, BALLISTA has been created to be accessible, and BED files that detail the precise location of the original features on the new unigene reference sequences could also be created to aid downstream analyses.

5.2.2 The *L. tenue* reference transcriptome

The current build of the *L. tenue* reference transcriptome (Section 4.2.2) is the highest quality reference built to date. However, the multiple *k*-mer assemblies generated with Velvet/Oases (Zerbino and Birney, 2008; Schulz *et al.*, 2012) demonstrated lower rates of overall alignment of the original sequence reads, and higher rates of missing universal single-copy orthologs, than the single-*k* Trinity (Garber *et al.*, 2011) assembly. This is due to aggressive *in silico* normalisation of the reads with ORNA (Durai and Schulz, 2017), which was a necessary step and recommended tool to prepare reads for assembly with Velvet/Oases; Trinity makes use of a bundled normalisation tool, Jellyfish (Marçais and Kingsford, 2011). Future builds of the *L. tenue* reference transcriptome should either include further optimisation of ORNA parameters for input to Velvet/Oases or use Jellyfish independently to create a universal set of normalised reads for input to both Velvet/Oases and Trinity tools. Additional assembly tools, such as Trans-ABySS (Robertson *et al.*, 2010) and SOAPdenovo-Trans (Xie *et al.*, 2014), could also be utilised to capture as broad a complement of the sample transcript populations as possible.

5.2.3 Improving the power of the *L.tenue* floral RNAseq experiment

Further sequencing of additional biological replicates would improve the power of the expression analyses by *(i)* increasing the sample size of pin and thrum individuals and *(ii)* ameliorating the levels of dispersion in library size. Through the inclusion of more individuals in the analyses, a wider range of allelic variation at the *S* locus could be captured, providing greater confidence for the subgroup structures of pin- and thrum-specific patterns of expression. The variance in per-treatment feature counts tends to decrease with the inclusion of biological replicates, which can greatly improve the replicability of an experiment. This in turn increases the power of the experiment to avoid type II false negative errors. The increased sampling would also allow pairwise comparisons of pin and thrum samples to be carried out by growth stage, as the lower levels of count data dispersion in treatments by single growth stages would provide enough power to offset the decrease in the power that pooling all growth stages into a larger treatment would provide. This would allow a more detailed analysis of the phases of flower development and the discovery of more differentially expressed features exclusively present in specific developmental stages.

5.2.4 Refining the list of putative candidate loci

Streamlining of the list of differentially expressed features and improvements in the biological interpretation of the dataset would be achieved through Gene Ontology (GO) enrichment and pathway analyses. Such analyses would reveal which GO terms and molecular pathways are over- or under-represented in the gene set, providing valuable and important insights into the molecular functions of the putative candidate loci. Thrum-specific repression of the brassinosteroid, gibberellic acid, and auxin response pathways

in particular would be tell-tale signals for the influence of Contig_141165. Analyses of coexpression are also great avenues for further work. Particular focus on genes exhibiting expression patterns correlated with the activity of Contig_141165 may uncover loci downstream of its regulation network, or even further *S* locus candidates.

The results of the genomic clustering showed some promise, particularly with the high levels of co-localising features on Chromosomes 3 and 4 of *A. thaliana*. A more detailed analysis of the *L. tenue* contigs mapping to these regions could provide candidate regions in *L. tenue* through synteny mapping.

5.2.5 Developmental mechanisms behind tall and short floral organs of *L. tenue*

Our study demonstrated that the mechanism driving the shorter height of the thrum style relative to the pin style appears to be largely a difference in cell elongation. Though the comparable cell lengths of pin and thrum stamen filaments suggest the height difference in male organs is driven by cell number, this was not directly measured in our study. Future investigations would benefit from further microscopy work to count all the epidermal cells on a single plane from the base to the tip of the stamen filaments in order to gain insights into the mechanism driving differences in male organ heights.

Developmental genetic studies could also greatly complement this work, as known pathways controlling cell length and cell division can be used to observe the behaviour of genes of interest in the expression datasets. Further developmental genetic investigations would strongly benefit from further differentiating samples by morph-specific organs, as opposed to the whole flower, to observe how gene expression affects the male and female organs of each morph.

5.2.6 Gene flow dynamics within and between *L. tenue* populations

Population genetic studies of *L. tenue* and other *Linum* species would be of immense value to the future of this project. First, with increasing sample sizes and the addition of more populations to the study, we could gain a deeper understanding of morph frequency bias in *L. tenue* populations. We found there was no significant difference in morph frequencies in the five wild populations we observed in the field (Table 2.3), though disequilibrium of morph ratios in heterostylous populations is not uncommon and can affect population-level measures of genetic diversity (Meeus *et al.*, 2012). It has also been noted through personal observations (AF) that some populations of *L. tenue* in the field appeared to show slightly skewed morph ratios, though this was not investigated in our study. The inclusion of a more rigorous sampling of populations would provide further insights on this issue.

Next-generation sequencing (NGS) approaches allow population genetics to be conducted on a genomic scale, through the use of techniques such as restriction site-associated DNA sequencing (RADseq). Primarily, this could be used to conduct association mapping and identify broad-scale candidate markers for the *S* locus. This could allow us to identify candidate genomic regions of interest, and could open up avenues of investigation for levels of allelic diversity in *S* locus loci determining the different male and female organ traits. Population genetic studies would also allow us to investigate the evolutionary implications of disyly in *Linum*, particularly with studies conducted across species with varied mating systems. In this way, the evolutionary consequences of heterostyly, self-incompatibility and self-compatibility could be examined.

5.3 The future of bioinformatics perspectives

We are currently in a transitional period between two generations of sequencing technologies. Where the previous decade has been dominated by second-generation short-read Illumina sequencing, the next ten years may well see a shift to the widespread employment of single-molecule sequencing. This so-called 'third-generation' sequencing, currently led by Pacific Biosciences (PacBio) and Oxford Nanopore, is still in its infancy, yet if it develops to its full potential is set to paradigmatically change the molecular biology and bioinformatics research scenes.

In 2011, PacBio commercialised what has become known as single-molecule real-time sequencing; a sequencing-by-synthesis method that eavesdrops on the natural activity of DNA polymerase, which can replicate 3000 nucleotides per minute. Through the use of rolling circle repetitive sequencing, where the cDNA is circularised, the base-calling accuracy rates can rival that of current day Illumina technologies. In 2011, PacBio claimed that this technology could soon be used to sequence whole human genomes in minutes for under $100, though the current cost-per-base-sequencing is relatively high.

In contrast, Oxford Nanopore technology uses a charged membrane that is punctured with pores <1 nM in diameter, that can either be composed of either protein channels (biological membranes) or solid materials (solid-state nanopores), through which nucleic acids can be pulled by electrophoresis. As the molecules pass through the membranes, the charge signature of each base is recorded. The major benefits of this technology are that, as the sequencing is not dependant on DNA synthesis, RNA and even amino acid molecules can also be sequenced. This can have a resounding impact on the nature of genomics in two ways. Firstly, the ability to directly sequence RNA molecules removes the need to prepare cDNA libraries. This can not only be laboriously intensive and time-

consuming (as can be attested through personal experience!), but removes the reverse-transcription and PCR bias of library preparation and sequencing. Such biases can change the composition of the molecule pool and lead to composition bias between samples, or the over- and under-representation of certain RNA species or transcripts from particular loci. This also allows transcript counts to be acquired in real time without the need for aligning short reads back to a reference genome or transcriptome, which removes many uncertainties that can arise during expression analyses with Illumina data. Secondly, the ability to easily and cost-effectively conduct direct quantitative proteomics has the potential to change the way research is conducted. With the exception of certain RNA species, proteins are largely the main functional elements of the genome: RNAseq is usually used as a proxy for understanding the proteome, which means the technique has a limited lifespan in the future. There is evidence to suggest that interactions between transcriptome and proteome activity do not always correlate (Feussner and Polle, 2015; Ishitsuka, Akutsu and Nacher, 2016). Proteins are also the elements of the genome that selection directly acts on, and thus proteomics will afford us a greater understanding of how organisms interact with their environment.

A rise in approaches that use longer reads could have impacts for the tools used in bioinformatics. Longer reads could cause a shift in the algorithms that assembly tools use, and we could see a reversion to the use of Hamiltonian pathways in de Bruin graph-based software. In Hamiltonian cycles, the reads are assigned to the nodes of the de Bruijn graph and the overlaps to the edges, meaning an algorithm needs to function to visit each edge once to assemble the sequence. This was the technique used during the era of Sanger sequencing and was used to assemble the human genome. However, with the advent of Illumina data, the increased size and complexity of the de Bruijn graphs became very difficult to solve. de Bruijn-based algorithms thus transitioned to use Eulerian cycles, whereby the reads are now assigned to the edges of the graph and overlaps to the nodes,

requiring each node to be visited once for resolution. This is much less intensive on computational resources and is easier to solve. Such approaches were in fact designed for the early days of the microarray, which was originally designed to sequence the human genome through sequencing-by-hybridisation. Many of the modern NGS assembly tools available today have adapted Eulerian cycle-based algorithms that were originally designed for microarray data. This fluctuation in algorithms used by tools therefore follows popularity trends in the nature of data that is collected, and it is thus a near-certainty that we will see a recycling of approaches in the future.

Applications of NGS which make use of Illumina data that are unlikely to make the transition to single-molecule sequencing are micro RNA sequencing and RADseq. Single-molecule sequencing is less accurate for shorter molecules, and short-reads are likely to be lost at the alignment stage of such pipelines. Micro RNAs and their diagnostic precursor transcripts require sequencing of molecules between ~22-70 nucleotides in length and thus rely on short-read sequencing. Current alignment tools also tend to use algorithms that are best-suited to either short- or long-read mapping, making the application of single-molecule sequencing of limited value here. The use of Illumina is similarly useful to population genetics as the ability to multiplex large sample sizes is highly appealing. Population genomic approaches that use whole genome sequences are very useful and can reveal interesting patterns of genomic evolution, but the benefit of RADseq to population genetics is that it is a reduced-genome representation approach; and researchers may always prefer to balance the benefits of wider sampling over more genetic information. These applications may allow the research community to eke out the use of Illumina technology into the future.

Depending on the type of sequencing and the platform used, error rates with Nanopore can range from 1-16%. In conjunction with high cost-per-base sequencing and low

awareness of the technology in the research community currently leads to limited use of Nanopore sequencing; however, as the technologies improve and costs of sequencing fall, these techniques could become more widely adopted in time. Given the overhaul in approaches that NGS has brought over the last decade, it is not too far a stretch to imagine the possibilities that may come in the not too distant future.

References

Aii, Jotaro, Mio Nagano, A. Greg Penner, G. Clayton Campbell, and Taiji Adachi. 1998. "Identification of RAPD Markers Linked to the Homostylar (Ho) Gene in Buckwheat." *Ikushugaku Zasshi* 48 (1): 59–62. https://doi.org/10.1270/jsbbs1951.48.59.

Albert, V. A., W. B. Barbazuk, C. W. dePamphilis, J. P. Der, J. Leebens-Mack, H. Ma, J. D. Palmer, et al. 2013. "The Amborella Genome and the Evolution of Flowering Plants." *Science* 342 (6165): 1241089–1241089. https://doi.org/10.1126/science.1241089.

Allen, Alexandra M., Christopher J. Thorogood, Matthew J. Hegarty, Christian Lexer, and Simon J. Hiscock. 2011. "Pollenpistil Interactions and Self-Incompatibility in the Asteraceae: New Insights from Studies of Senecio Squalidus (Oxford Ragwort)." *Annals of Botany*. https://doi.org/10.1093/aob/mcr147.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Anderson, M. A., E. C. Cornish, S. L. Mau, E. G. Williams, R. Hoggart, A. Atkinson, I. Bonig, et al. 1986. "Cloning of CDNA for a Stylar Glycoprotein Associated with Expression of Self-Incompatibility in Nicotiana Alata." *Nature* 321 (6065): 38–44. https://doi.org/10.1038/321038a0.

Andrews, Simon. 2010. "A Quality Control Tool for High Throughput Sequence Data."

Appel, Ron D., Amos Bairoch, and Denis F. Hochstrasser. 1994. "A New Generation of Information Retrieval Tools for Biologists: The Example of the ExPASy WWW Server." *Trends in Biochemical Sciences* 19 (6): 258–60. https://doi.org/10.1016/0968-0004(94)90153-8.

Armbruster, W. Scott, Geir H. Bolstad, Thomas F. Hansen, Barbara Keller, Elena Conti, and Christophe Pélabon. 2017. "The Measure and Mismeasure of Reciprocity in Heterostylous Flowers." *New Phytologist* 215 (2): 906–17. https://doi.org/10.1111/nph.14604.

Armbruster, W. Scott, Rocío Pérez-Barrales, Juan Arroyo, Mary E. Edwards, and Pablo Vargas. 2006. "Three-Dimensional Reciprocity of Floral Morphs in Wild Flax (Linum Suffruticosum): A New Twist on Heterostyly." *New Phytologist* 171 (3): 581–90. https://doi.org/10.1111/j.1469-8137.2006.01749.x.

Artimo, P., M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. de Castro, S. Duvaud, et al. 2012. "ExPASy: SIB Bioinformatics Resource Portal." *Nucleic Acids Research* 40 (W1): W597–603. https://doi.org/10.1093/nar/gks400.

Arunkumar, Ramesh, Wei Wang, Stephen I. Wright, and Spencer C. H. Barrett. 2017. "The Genetic Architecture of Tristyly and Its Breakdown to Self-Fertilization." *Molecular Ecology* 26 (3): 752–65. https://doi.org/10.1111/mec.13946.

Athanasiou, A, and J S Shore. 1997. "Morph-Specific Proteins in Pollen and Styles of Distylous Turnera (Turneraceae)." *Genetics* 146 (2): 669–79. http://www.ncbi.nlm.nih.gov/pubmed/9178015.

Athanasiou, A., D. Khosravi, F. Tamari, and J. S. Shore. 2003. "Characterization and Localization of Short-Specific Polygalacturonase in Distylous Turnera Subulata (Turneraceae)." *American Journal of Botany* 90 (5): 675–82. https://doi.org/10.3732/ajb.90.5.675.

Barrett, S C H. 2002. "Sexual Interference of the Floral Kind." *Heredity* 88 (2): 154–59. https://doi.org/10.1038/sj.hdy.6800020.

Barrett, S. C. H. 1992. "Heterostylous Genetic Polymorphisms: Model Systems for Evolutionary Analysis." In , 1–29. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-86656-2_1.

Barrett, S. C. H., and J. S. Shore. 2008. "New Insights on Heterostyly: Comparative Biology, Ecology and Genetics." In *Self-Incompatibility in Flowering Plants*, 3–32. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68486-2_1.

Barrett, Spencer C H. 2002. "The Evolution of Plant Sexual Diversity." *Nature Reviews Genetics* 3 (April): 274. http://dx.doi.org/10.1038/nrg776.

Barrett, Spencer C. H., Martin T. Morgan, and Brian C. Husband. 1989. "The Dissolution of a Complex Genetic Polymorphism: The Evolution of Self-Fertilisation in Tristylous *Eichhornia Paniculata* (Pontederiaceae)." *Evolution* 43 (7): 1398–1416. https://doi.org/10.1111/j.1558-5646.1989.tb02591.x.

Barrett, Spencer C.H. 2013. "The Evolution of Plant Reproductive Systems: How Often Are Transitions Irreversible?" *Proceedings of the Royal Society B: Biological Sciences*. Royal Society. https://doi.org/10.1098/rspb.2013.0913.

Barrett, Spencer C.H., and Josh Hough. 2013. "Sexual Dimorphism in Flowering Plants." *Journal of Experimental Botany*. https://doi.org/10.1093/jxb/ers308.

Barrett, Spencer C.H., Linley K. Jesson, and Angela M. Baker. 2000. "The Evolution and Function of Stylar Polymorphisms in Flowering Plants." *Annals of Botany* 85 (March): 253–65. https://doi.org/10.1006/ANBO.1999.1067.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2012. "NCBI GEO: Archive for Functional Genomics Data Sets—Update." *Nucleic Acids Research* 41 (D1): D991–95. https://doi.org/10.1093/nar/gks1193.

Bartoń, K. 2009. "MuMIn : Multi-Model Inference, R Package Version 0.12.0." *Http://R-Forge.r-Project.Org/Projects/Mumin/*.

Bateman, Alex, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, et al. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1): D158–69. https://doi.org/10.1093/nar/gkw1099.

Bateson, W., and R. P. Gregory. 1905. "On the Inheritance of Heterostylism in Primula." *Proceedings of the Royal Society B: Biological Sciences* 76 (513): 581–86. https://doi.org/10.1098/rspb.1905.0049.

Bell, Charles D., Douglas E. Soltis, and Pamela S. Soltis. 2010. "The Age and Diversification of the Angiosperms Re-Revisited." *American Journal of Botany* 97 (8): 1296–1303. https://doi.org/10.3732/ajb.0900346.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. https://doi.org/10.2307/2346101.

Bryant, Donald M, Kimberly Johnson, Tia DiTommaso, Timothy Tickle, Matthew Brian Couger, Duygu Payzin-Dogru, Tae J Lee, et al. 2017. "A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors." *Cell Reports* 18 (3): 762–76. https://doi.org/10.1016/j.celrep.2016.12.063.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Buermans, H.P.J., and J.T. den Dunnen. 2014. "Next Generation Sequencing Technology: Advances and Applications." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842 (10): 1932–41. https://doi.org/10.1016/J.BBADIS.2014.06.015.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in MRNA-Seq Experiments." *BMC Bioinformatics* 11 (February): 94. https://doi.org/10.1186/1471-2105-11-94.

Cao, Dongni, Hui Cheng, Wei Wu, Hui Meng Soo, and Jinrong Peng. 2006. "Gibberellin Mobilizes Distinct DELLA-Dependent Transcriptomes to Regulate Seed Germination and Floral Development in Arabidopsis." *Plant Physiology* 142 (2): 509–25. https://doi.org/10.1104/pp.106.082289.

Chang, Zheng, Zhenjia Wang, and Guojun Li. 2014. "The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study." Edited by F. Nina Papavasiliou. *PLoS ONE* 9 (4): e94825. https://doi.org/10.1371/journal.pone.0094825.

Charlesworth, D, and B Charlesworth. 1987. "Inbreeding Depression and Its Evolutionary Consequences." *Annual Review of Ecology and Systematics* 18 (1): 237–68. https://doi.org/10.1146/annurev.es.18.110187.001321.

Charlesworth, D., and B. Charlesworth. 1979. "A Model for the Evolution of Distyly." *The American Naturalist*. The University of Chicago PressThe American Society of Naturalists. https://doi.org/10.1086/283496.

Charlesworth, Deborah. 2006. "Evolution of Plant Breeding Systems." *Current Biology*. https://doi.org/10.1016/j.cub.2006.07.068.

Chase, M. W., M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, D. J. Mabberley, A. N. Sennikov, P. S. Soltis, and P. F. Stevens. 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): 1–20. https://doi.org/10.1111/boj.12385.

Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton. 2004. "The Jalview Java Alignment Editor." *Bioinformatics* 20 (3): 426–27. https://doi.org/10.1093/bioinformatics/btg430.

Cocker, Jonathan M., Jonathan Wright, Jinhong Li, David Swarbreck, Sarah Dyer, Mario Caccamo, and Philip M. Gilmartin. 2018. "Primula Vulgaris (Primrose) Genome Assembly, Annotation and Gene Expression, with Comparative Genomics on the Heterostyly Supergene." *Scientific Reports* 8 (1): 17942. https://doi.org/10.1038/s41598-018-36304-4.

Cohen, James I. 2011. "A Phylogenetic Analysis of Morphological and Molecular Characters of Lithospermum L. (Boraginaceae) and Related Taxa: Evolutionary Relationships and Character Evolution." *Cladistics* 27 (6): 559–80. https://doi.org/10.1111/j.1096-0031.2011.00352.x.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (1): 13. https://doi.org/10.1186/s13059-016-0881-8.

Corner, E. J. H. (Edred John Henry). 1964. *The Life of Plants*. Cleveland, Ohio: World Publishing Co.

Crane, P. R. 1993. "Plant Evolution: Time for the Angiosperms." *Nature*. https://doi.org/10.1038/366631a0.

Cunnane, Stephen C. 2003. "The Contribution of α-Linolenic Acid in Flaxseed to Human Health." In *Flax: The Genus Linum*, 150–80. CRC Press.

Darwin, Charles. 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 of London. Botany* 6 (22): 77–96. https://doi.org/10.1111/j.1095-8312.1862.tb01218.x.

Darwin, Charles. 1877. "The Different Forms of Flowers on Plants of the Same Species." *Murray*. https://books.google.co.uk/books?hl=en&lr=&id=FhOP8L8GAbIC&oi=fnd&pg=PA1&dq=Darwin+C.+1877.+The+different+forms+of+flowers+on+plants+of

+the+same+species.+London,+UK:+Murray.&ots=bUDMMmIndp&sig=ucSO2hMnBWhdQ5qM0aRz9L57e_o#v=onepage&q&f=false.

Darwin, Charles. 1876. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom - Charles Darwin - Google Books*. John Murray, London. https://books.google.co.uk/books?hl=en&lr=&id=4nlYKu2SorAC&oi=fnd&pg=PR1&dq=effects+of+cross+and+self-fertilisation+in+veg+kingdom&ots=lmrIw2EHLR&sig=n__FndpuC9Cg5Y_onCfyoKo5_EE#v=onepage&q=effects of cross and self-fertilisation in veg king.

Davidson, Nadia M., Anthony D. K. Hawkins, and Alicia Oshlack. 2017. "SuperTranscripts: A Data Driven Reference for Analysis and Visualisation of Transcriptomes." *Genome Biology* 18 (1): 148. https://doi.org/10.1186/s13059-017-1284-1.

Paolo, Sofia De, Marco Salvemini, Luciano Gaudio, and Serena Aceto. 2014. "De Novo Transcriptome Assembly from Inflorescence of Orchis Italica: Analysis of Coding and Non-Coding Transcripts." *PloS One* 9 (7): e102155. https://doi.org/10.1371/journal.pone.0102155.

Vos, Jurriaan M de, Colin E Hughes, Gerald M Schneeweiss, Brian R Moore, and Elena Conti. 2014. "Heterostyly Accelerates Diversification via Reduced Extinction in Primroses." *Proceedings. Biological Sciences* 281 (1784): 20140075. https://doi.org/10.1098/rspb.2014.0075.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, et al. 2013. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83. https://doi.org/10.1093/bib/bbs046.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dohm, Juliane C., Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. 2008. "Substantial Biases in Ultra-Short Read Data Sets from High-Throughput DNA Sequencing." *Nucleic Acids Research* 36 (16). https://doi.org/10.1093/nar/gkn425.

Donoghue, M, and J Doyle. 1989. *Phylogenetic Analysis of Angiosperms and the Relationships of Hamamelidae. In P. R. Crane and S. Blackmore [Eds.], Evolution, Systematics, and Fossil History of the Hamamelidae, Vol. 1, Introduction and "Lower" Hamamelidae*. Oxford, UK: Clarendon Press.

Dulberger, Rivka. 1992. "Modifications in the Interapertural Exine Following Pollinations of Linum Grandiflorum." In *Angiosperm Pollen and Ovules*, 253–58. Springer New York. https://doi.org/10.1007/978-1-4612-2958-2_41.

Eckert, Christopher G., and Spencer C.H. Barrett. 1994. "Tristyly, Self-compatibility and Floral Variation in Decodon Verticillatus (Lythraceae)." *Biological Journal of the Linnean Society* 53 (1): 1–30. https://doi.org/10.1111/j.1095-8312.1994.tb01000.x.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." Edited by William R. Pearson. *PLoS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, R. C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Ekblom, R, and J Galindo. 2011. "Applications of next Generation Sequencing in Molecular Ecology of Non-Model Organisms." *Heredity* 107 (1): 1–15. https://doi.org/10.1038/hdy.2010.152.

Engström, Pär G, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, et al. 2013. "Systematic Evaluation of Spliced Alignment Programs for RNA-Seq Data." *Nature Methods* 10 (12): 1185–91. https://doi.org/10.1038/nmeth.2722.

Ernst, A. 1928. "60. A. Ernst: Zur Vererbung Der Morphologischen Heterostyliemerkmale." *Berichte Der Deutschen Botanischen Gesellschaft* 46 (8): 573–88. https://doi.org/10.1111/J.1438-8677.1928.TB00355.X.

Ernst, Alfred. 1936. "Heterostylie-Forschung." *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre* 71 (1): 156–230. https://doi.org/10.1007/BF01848862.

Ernst, Alfred. 1955. "Self-Fertility in Monomorphic Primulas." *Genetica* 27 (1): 391–448. https://doi.org/10.1007/BF01664170.

Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. 2018. "Selecting Between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions." *Briefings in Bioinformatics* 19 (5): 776–92. https://doi.org/10.1093/bib/bbx008.

Faivre, A E. 2000. "Ontogenetic Differences in Heterostylous Plants and Implications for Development from a Herkogamous Ancestor." *Evolution; International Journal of Organic Evolution* 54 (3): 847–58. http://www.ncbi.nlm.nih.gov/pubmed/10937258.

Faivre, Amy E., and Lucinda A. Mcdade. 2001. "Population-Level Variation in the Expression of Heterostyly in Three Species of Rubiaceae: Does Reciprocal Placement of Anthers and Stigmas Characterize Heterostyly?" *American Journal of Botany* 88 (5): 841–53. https://doi.org/10.2307/2657036.

Ferrero, V., I. Chapela, J. Arroyo, and L. Navarro. 2011. "Reciprocal Style Polymorphisms Are Not Easily Categorised: The Case of Heterostyly in Lithodora and Glandora (Boraginaceae)." *Plant Biology* 13 (SUPPL. 1): 7–18. https://doi.org/10.1111/j.1438-8677.2009.00307.x.

Feussner, Ivo, and Andrea Polle. 2015. "What the Transcriptome Does Not Tell — Proteomics and Metabolomics Are Closer to the Plants' Patho-Phenotype." *Current Opinion in Plant Biology* 26 (August): 26–31. https://doi.org/10.1016/J.PBI.2015.05.023.

Finotello, F., and B. Di Camillo. 2015. "Measuring Differential Gene Expression with RNA-Seq: Challenges and Strategies for Data Analysis." *Briefings in Functional Genomics* 14 (2): 130–42. https://doi.org/10.1093/bfgp/elu035.

Foote, H. C., J. P. Ride, V. E. Franklin-Tong, E. A. Walker, M. J. Lawrence, and F. C. Franklin. 1994. "Cloning and Expression of a Distinctive Class of Self-Incompatibility (S) Gene from Papaver Rhoeas L." *Proceedings of the National Academy of Sciences* 91 (6): 2265–69. https://doi.org/10.1073/pnas.91.6.2265.

Friedman, William E. 2006. "Embryological Evidence for Developmental Lability during Early Angiosperm Evolution." *Nature* 441 (7091): 337–40. https://doi.org/10.1038/nature04690.

Friis, Else Marie, Kaj Raunsgaard Pedersen, and Peter R Crane. 2005. "When Earth Started Blooming: Insights from the Fossil Record." *Current Opinion in Plant Biology* 8 (1): 5–12. https://doi.org/10.1016/j.pbi.2004.11.006.

Frohlich, Michael W., and David S. Parker. 2000. "The Mostly Male Theory of Flower Evolutionary Origins: From Genes to Fossils." *Systematic Botany* 25 (2): 155. https://doi.org/10.2307/2666635.

Ganders, Fred R. 1979. "The Biology of Heterostyly." *New Zealand Journal of Botany* 17: 607–35. https://www.tandfonline.com/doi/pdf/10.1080/0028825X.1979.10432574?needAccess=true.

Ganders, Fred R. 1974. " Disassortative Pollination in the Distylous Plant Jepsonia Heterandra ." *Canadian Journal of Botany* 52 (11): 2401–6. https://doi.org/10.1139/b74-311.

Gao, Jian, Ying Zhang, Chunling Zhang, Feiyan Qi, Xueping Li, Shaohua Mu, and Zhenhua Peng. 2014. "Characterization of the Floral Transcriptome of Moso Bamboo (Phyllostachys Edulis) at Different Flowering Developmental Stages by Transcriptome Sequencing and RNA-Seq Analysis." Edited by Emmanuel Gaquerel. *PLoS ONE* 9 (6): e98910. https://doi.org/10.1371/journal.pone.0098910.

Garber, R J & Quisenberry, K. S. 1927. "THE INHERITANCE OF LENGTH OF STYLE IN BUCKWHEAT ^." *Journal of Agricultural Research* 34 (2): 181–83. https://naldc.nal.usda.gov/download/IND43967384/PDF.

García-Robledo, Carlos. 2008. "Asymmetry in Pollen Flow Promotes Gender Specialization in Morphs of the Distylous Neotropical Herb Arcytophyllum Lavarum (Rubiaceae)." *Evolutionary Ecology* 22 (6): 743–55. https://doi.org/10.1007/s10682-007-9198-0.

Gasteiger, E., Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D. Appel, and Amos Bairoch. 2003. "ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis." *Nucleic Acids Research* 31 (13): 3784–88. https://doi.org/10.1093/nar/gkg563.

Glover, Beverley J. (Beverley Jane). 2007. *Understanding Flowers and Flowering: An Integrated Approach*. First Edit. Oxford University Press.

Goda, Hideki, Eriko Sasaki, Kenji Akiyama, Akiko Maruyama-Nakashita, Kazumi Nakabayashi, Weiqiang Li, Mikihiro Ogawa, et al. 2008. "The AtGenExpress Hormone and Chemical Treatment Data Set: Experimental Design, Data Evaluation, Model Data Analysis and Data Access." *The Plant Journal* 55 (3): 526–42. https://doi.org/10.1111/j.1365-313X.2008.03510.x.

Goda, Hideki, Shinichiro Sawa, Tadao Asami, Shozo Fujioka, Yukihisa Shimada, and Shigeo Yoshida. 2004. "Comprehensive Comparison of Auxin-Regulated and Brassinosteroid-Regulated Genes in Arabidopsis [W]." https://doi.org/10.1104/pp.103.034736.

Goldberg, Emma E., Joshua R. Kohn, Russell Lande, Kelly A. Robertson, Stephen A. Smith, and Boris Igić. 2010. "Species Selection Maintains Self-Incompatibility." *Science* 330 (6003): 493–95. https://doi.org/10.1126/science.1194513.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Gray, J. E., B. A. McClure, I. Bonig, M. A. Anderson, and A. E. Clarke. 1991. "Action of the Style Product of the Self-Incompatibility Gene of Nicotiana Alata (S-RNase) on in Vitro-Grown Pollen Tubes." *The Plant Cell*, March, 271–83. https://doi.org/10.1105/tpc.3.3.271.

Gregory, R. P., D. de Winton, and W. Bateson. 1923. "Genetics of Primula Sinensis." *Journal of Genetics* 13 (2): 219–53. https://doi.org/10.1007/BF02983056.

Grienenberger, Etienne, and Carl J Douglas. 2014. "Arabidopsis VASCULAR-RELATED UNKNOWN PROTEIN1 Regulates Xylem Development and Growth by a Conserved Mechanism That Modulates Hormone Signaling 1[W][OPEN]." https://doi.org/10.1104/pp.114.236406.

Guo, Yan, Shilin Zhao, Chung I. Li, Quanhu Sheng, and Y. Shyr. 2014. "RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment." *Cancer Informatics* 13 (October): 1–5. https://doi.org/10.4137/CIN.S17688.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8 (8): 1494–1512. https://doi.org/10.1038/nprot.2013.084.

Hamston, Tracey J., Robert J. Wilson, Natasha de Vere, Tim C. G. Rich, Jamie R. Stevens, and James E. Cresswell. 2017. "Breeding System and Spatial Isolation from Congeners Strongly Constrain Seed Set in an Insect-Pollinated Apomictic Tree: Sorbus Subcuneata (Rosaceae)." *Scientific Reports* 7 (March): 45122. https://doi.org/10.1038/srep45122.

Hansen, Kasper D., Steven E. Brenner, and Sandrine Dudoit. 2010. "Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming." *Nucleic Acids Research* 38 (12). https://doi.org/10.1093/nar/gkq224.

Harder, Lawrence D., and Spencer Charles Hilton. Barrett. 2006. *Ecology and Evolution of Flowers*. Oxford University Press. https://books.google.co.uk/books?hl=en&lr=&id=3c8TDAAAQBAJ&oi=fnd&pg=PR9&dq=harder+and+barrett+2006&ots=hxUw4OMQsy&sig=4dqjSgE2j2y6s269ws6oDGDbO_k#v=onepage&q=harder and barrett 2006&f=false.

Hegarty, Matthew J., Gary L. Barker, Ian D. Wilson, Richard J. Abbott, Keith J. Edwards, and Simon J. Hiscock. 2006. "Transcriptome Shock after Interspecific Hybridization in Senecio Is Ameliorated by Genome Duplication." *Current Biology* 16 (16): 1652–59. https://doi.org/10.1016/J.CUB.2006.06.071.

Heo, Yun, Xiao-Long Wu, Deming Chen, Jian Ma, and Wen-Mei Hwu. 2014. "BLESS: Bloom Filter-Based Error Correction Solution for High-Throughput Sequencing Reads." *Bioinformatics* 30 (10): 1354–62. https://doi.org/10.1093/bioinformatics/btu030.

Heywood, V. H. (Vernon Hilton). 1978. *Flowering Plants of the World*. New York, USA: Mayflower Books.

Hicks, Stephanie C., and Rafael A. Irizarry. 2014. "When to Use Quantile Normalization?" *BioRxiv*, December, 012203. https://doi.org/10.1101/012203.

Hiscock, Simon J., and Stephanie M. McInnis. 2003. "Pollen Recognition and Rejection during the Sporophytic Self-Incompatibility Response: Brassica and Beyond." *Trends in Plant Science*. Elsevier Ltd. https://doi.org/10.1016/j.tplants.2003.10.007.

Hiscock, Simon J., Stephanie M. McInnis, David A. Tabah, Catherine A. Henderson, and Adrian C. Brennan. 2003. "Sporophytic Self-Incompatibility in Senecio Squalidus L. (Asteraceae) - The Search for S." In *Journal of Experimental Botany*, 54:169–74. Oxford University Press. https://doi.org/10.1093/jxb/erg005.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular Biology and Evolution* 35 (2): 518–22. https://doi.org/10.1093/molbev/msx281.

Hrdlickova, Radmila, Masoud Toloue, and Bin Tian. 2017. "RNA-Seq Methods for Transcriptome Analysis." *Wiley Interdisciplinary Reviews: RNA* 8 (1): e1364. https://doi.org/10.1002/wrna.1364.

Huang, Xiaoqiu, and Anup Madan. 1999. "CAP3: A DNA Sequence Assembly Program." *Genome Research* 9 (9): 868–77. https://doi.org/10.1101/gr.9.9.868.

Hughes, Norman F. (Norman Francis). 1994. *The Enigma of Angiosperm Origins*. Cambridge University Press.

Huu, Cuong Nguyen, Christian Kappel, Barbara Keller, Adrien Sicard, Yumiko Takebayashi, Holger Breuninger, Michael D Nowak, et al. 2016. "Presence versus Absence of CYP734A50 Underlies the Style-Length Dimorphism in Primroses." *ELife* 5 (September): e17956. https://doi.org/10.7554/eLife.17956.

Ishitsuka, Masayuki, Tatsuya Akutsu, and Jose C. Nacher. 2016. "Critical Controllability in Proteome-Wide Protein Interaction Network Integrating Transcriptome." *Scientific Reports* 6 (1): 23541. https://doi.org/10.1038/srep23541.

Jiang, Xian Feng, Xing Fu Zhu, Ling Ling Chen, and Qing Jun Li. 2018. "What Ecological Factors Favor the Shift from Distyly to Homostyly? A Study from the Perspective of Reproductive Assurance." *Journal of Plant Ecology* 11 (4): 645–55. https://doi.org/10.1093/jpe/rtx036.

Jin, Jinpu, Feng Tian, De-Chang Yang, Yu-Qi Meng, Lei Kong, Jingchu Luo, and Ge Gao. 2017. "PlantTFDB 4.0: Toward a Central Hub for Transcription Factors and Regulatory Interactions in Plants." *Nucleic Acids Research* 45 (D1): D1040–45. https://doi.org/10.1093/nar/gkw982.

Johnson, Steven D., and Kim E. Steiner. 2003. "Specialized Pollination Systems in Southern Africa." *South African Journal of Science*, no. 99: 59–66.

Johnson, and Steiner. 2000. "Generalization versus Specialization in Plant Pollination Systems." *Trends in Ecology & Evolution* 15 (4): 140–43. http://www.ncbi.nlm.nih.gov/pubmed/10717682.

Johnston, Mark O, Emmanuelle Porcher, Pierre-Olivier Cheptou, Christopher G Eckert, Elizabeth Elle, Monica A Geber, Susan Kalisz, et al. 2009. "Correlations among Fertility Components Can Maintain Mixed Mating in Plants." *The American Naturalist* 173 (1): 1–11. https://doi.org/10.1086/593705.

Kadam, B.S., and S.M. Patel. 1938. "Anthesis in Flax." *Journal of the American Society of Agronomy* 30 (11): 932–40.

Kalisz, Susan, April Randle, David Chaiffetz, Melisa Faigeles, Aileen Butera, and Craig Beight. 2012. "Dichogamy Correlates with Outcrossing Rate and Defines the Selfing Syndrome in the Mixed-Mating Genus Collinsia." *Annals of Botany* 109 (3): 571–82. https://doi.org/10.1093/aob/mcr237.

Kálmán, K., A. Medvegy, Z. Pénzes, and E. Mihalik. 2007. "Morph-Specific Variation of Floral Traits Associated with Reciprocal Herkogamy in Natural Populations of Primula Vulgaris and Primula Veris." *Plant Systematics and Evolution* 268 (1–4): 15–27. https://doi.org/10.1007/s00606-007-0575-5.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." *Nature Methods* 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.

Kandasamy, M. K., D. J. Paolillo, C. D. Faraday, J. B. Nasrallah, and M. E. Nasrallah. 1989. "The S-Locus Specific Glycoproteins of Brassica Accumulate in the Cell

Wall of Developing Stigma Papillae." *Developmental Biology* 134 (2): 462–72. https://doi.org/10.1016/0012-1606(89)90119-X.

Kappel, Christian, Cuong Nguyen Huu, and Michael Lenhard. 2017. "A Short Story Gets Longer: Recent Insights into the Molecular Basis of Heterostyly." *Journal of Experimental Botany* 68 (21–22): 5719–30. https://doi.org/10.1093/jxb/erx387.

Kearns, Carol Ann, and David W. Inouye. 1994. "Fly Pollination of Linum Lewish (Linaceae)." *American Journal of Botany* 81 (9): 1091–95. https://doi.org/10.1002/j.1537-2197.1994.tb15602.x.

Keller, Barbara, Jurriaan M. De Vos, and Elena Conti. 2012. "Decrease of Sexual Organ Reciprocity between Heterostylous Primrose Species, with Possible Functional and Evolutionary Implications." *Annals of Botany* 110 (6): 1233–44. https://doi.org/10.1093/aob/mcs199.

Keller, Barbara, James D. Thomson, and Elena Conti. 2014. "Heterostyly Promotes Disassortative Pollination and Reduces Sexual Interference in Darwin's Primroses: Evidence from Experimental Studies." Edited by Gaku Kudo. *Functional Ecology* 28 (6): 1413–25. https://doi.org/10.1111/1365-2435.12274.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4): R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kissling, Jonathan, and Spencer C H Barrett. 2013. "Variation and Evolution of Herkogamy in Exochaenium (Gentianaceae): Implications for the Evolution of Distyly." *Annals of Botany* 112 (1): 95–102. https://doi.org/10.1093/aob/mct097.

Kubo, Ken Ichi, Tetsuyuki Entani, Akie Takara, Ning Wang, Allison M. Fields, Zhihua Hua, Mamiko Toyoda, et al. 2010. "Collaborative Non-Self Recognition System in S-RNase-Based Self-Incompatibility." *Science* 330 (6005): 796–99. https://doi.org/10.1126/science.1195243.

Kurian, Valsa, and A J Richards. 1997. "A New Recombinant in the Heteromorphy 'S' Supergene in Prim u/A." *Heredity* 78: 383–90. https://www.nature.com/articles/hdy199761.pdf.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "LmerTest Package: Tests in Linear Mixed Effects Models ." *Journal of Statistical Software* 82 (13). https://doi.org/10.18637/jss.v082.i13.

Labonne, Jdj, F Tamari, Js Shore, and Heredity. 2010. "Characterization of X-Ray-Generated Floral Mutants Carrying Deletions at the S-Locus of Distylous Turnera Subulata." *Heredity* 10539 (10510): 235–43. https://doi.org/10.1038/hdy.2010.83.

Labonne, Jonathan J. D., Alina Goultiaeva, and Joel S. Shore. 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 (6): 673–85. https://doi.org/10.1007/s00438-009-0439-5.

Lai, Zhao, Wenshi Ma, Bin Han, Lizhi Liang, Yansheng Zhang, Guofan Hong, and Yongbiao Xue. 2002. "An F-Box Gene Linked to the Self-Incompatibility (S) Locus of Antirrhinum Is Expressed Specifically in Pollen and Tapetum." *Plant Molecular Biology* 50 (1): 29–42. https://doi.org/10.1023/a:1016050018779.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lau, Pablo, and Carlos Bosque. 2003. "Pollen Flow in the Distylous Palicourea Fendleri (Rubiaceae): An Experimental Test of the Disassortative Pollen Flow Hypothesis." *Oecologia* 135 (4): 593–600. https://doi.org/10.1007/s00442-003-1216-5.

Lee, Y., J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung, and J. Quackenbush. 2004. "The TIGR Gene Indices: Clustering and Assembling EST and Known Genes and Integration with Eukaryotic Genomes." *Nucleic Acids Research* 33 (Database issue): D71–74. https://doi.org/10.1093/nar/gki064.

Leng, Ning, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. 2013. "EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments." *Bioinformatics* 29 (8): 1035–43. https://doi.org/10.1093/bioinformatics/btt087.

Lewis, D. 1954. "Comparative Incompatibility in Angiosperms and Fungi." *Advances in Genetics* 6 (C): 235–85. https://doi.org/10.1016/S0065-2660(08)60131-5.

Lewis, D., and D. A. Jones. 1992. "The Genetics of Heterostyly." In , 129–50. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-86656-2_5.

Li, Bo, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. 2010. "RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty." *Bioinformatics* 26 (4): 493–500. https://doi.org/10.1093/bioinformatics/btp692.

Li, J., B. Dudas, M. A. Webster, H. E. Cook, B. H. Davies, and P. M. Gilmartin. 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 (12): 5664–68. https://doi.org/10.1073/pnas.0910955107.

Li, Jinhong, Jonathan M. Cocker, Jonathan Wright, Margaret A. Webster, Mark McMullan, Sarah Dyer, David Swarbreck, Mario Caccamo, Cock van Oosterhout, and Philip M. Gilmartin. 2016. "Genetic Architecture and Evolution of the S Locus Supergene in Primula Vulgaris." *Nature Plants* 2 (12): 16188. https://doi.org/10.1038/nplants.2016.188.

Li, Jinhong, Margaret A. Webster, Jonathan Wright, Jonathan M. Cocker, Matthew C. Smith, Farah Badakshi, Pat Heslop-Harrison, and Philip M. Gilmartin. 2015. "Integration of Genetic and Physical Maps of the *Primula Vulgaris S* Locus and Localization by Chromosome *in Situ* Hybridization." *New Phytologist* 208 (1): 137–48. https://doi.org/10.1111/nph.13373.

Li, Jinhong, Margaret Webster, Brigitta Dudas, Holly Cook, Iain Manfield, Brendan Davies, and Philip M. Gilmartin. 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." *The Plant Journal* 56 (1): 1–12. https://doi.org/10.1111/j.1365-313X.2008.03584.x.

Li, Jinhong, Margaret Webster, Masaki Furuya, and Philip M. Gilmartin. 2007. "Identification and Characterization of Pin and Thrum Alleles of Two Genes That Co-Segregate with the Primula S Locus." *The Plant Journal* 51 (1): 18–31. https://doi.org/10.1111/j.1365-313X.2007.03125.x.

Liang, F, I Holt, G Pertea, S Karamycheva, S L Salzberg, and J Quackenbush. 2000. "An Optimized Protocol for Analysis of EST Sequences." *Nucleic Acids Research* 28 (18): 3657–65. http://www.ncbi.nlm.nih.gov/pubmed/10982889.

Lin, Brenda B. 2011. "Resilience in Agriculture through Crop Diversification: Adaptive Management for Environmental Change." *BioScience* 61 (3): 183–93. https://doi.org/10.1525/bio.2011.61.3.4.

Liu, Mingming, Zach N. Adelman, Kevin M. Myles, and Liqing Zhang. 2014. "A Transcriptome Post-Scaffolding Method for Assembling High Quality Contigs." *Computational Biology Journal* 2014: 1–4. https://doi.org/10.1155/2014/961823.

Liu, Shengyi, Yumei Liu, Xinhua Yang, Chaobo Tong, David Edwards, Isobel A.P. Parkin, Meixia Zhao, et al. 2014. "The Brassica Oleracea Genome Reveals the Asymmetrical Evolution of Polyploid Genomes." *Nature Communications* 5 (1): 3930. https://doi.org/10.1038/ncomms4930.

Lloyd, D. G., and C. J. Webb. 1992a. "The Evolution of Heterostyly." In *Evolution and Function of Heterostyly*, edited by S. C. H. Barrett, 151–78. Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-86656-2_6.

Lloyd, D. G., and C. J. Webb. 1992b. "The Selection of Heterostyly." In *Evolution and Function of Heterostyly*, edited by S. C. H. Barrett, 179–207. Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-86656-2_7.

Lloyd, David G., and Jocelyn M. A. Yates. 1982. "Intrasexual Selection and the Segregation of Pollen and Stigmas in Hermaphrodite Plants, Exemplified by Wahlenbergia Albomarginata (Campanulaceae)." *Evolution* 36 (5): 903. https://doi.org/10.2307/2408071.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

MacManes, Matthew D. 2014. "On the Optimal Trimming of High-Throughput MRNA Sequence Data." *Frontiers in Genetics* 5 (January): 13. https://doi.org/10.3389/fgene.2014.00013.

Manfield, I. W., Vassily K Pavlov, Jinhong Li, Holly E Cook, Florian Hummel, and Philip M Gilmartin. 2005. "Molecular Characterization of DNA Sequences from

the Primula Vulgaris S-Locus." *Journal of Experimental Botany* 56 (414): 1177–88. https://doi.org/10.1093/jxb/eri110.

Marçais, Guillaume, and Carl Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers." *Bioinformatics* 27 (6): 764–70. https://doi.org/10.1093/bioinformatics/btr011.

Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, et al. 2015. "CDD: NCBI's Conserved Domain Database." *Nucleic Acids Research* 43 (D1): D222–26. https://doi.org/10.1093/nar/gku1221.

Martin, Jeffrey A., and Zhong Wang. 2011. "Next-Generation Transcriptome Assembly." *Nature Reviews Genetics* 12 (10): 671. https://doi.org/10.1038/nrg3068.

Massinga, Paulo H., Steven D. Johnson, and Lawrence D. Harder. 2005. "Heteromorphic Incompatibility and Efficiency of Pollination in Two Distylous Pentanisia Species (Rubiaceae)." In *Annals of Botany*, 95:389–99. https://doi.org/10.1093/aob/mci040.

Matsui, K., T. Tetsuka, T. Nishio, and T. Hara. 2003. "Heteromorphic Incompatibility Retained in Self-Compatible Plants Produced by a Cross between Common and Wild Buckwheat." *New Phytologist* 159 (3): 701–8. https://doi.org/10.1046/j.1469-8137.2003.00840.x.

McClure, Bruce. 2016. "Reproduction: The Genetic Basis of Heterostyly." *Nature Plants*. Palgrave Macmillan Ltd. https://doi.org/10.1038/nplants.2016.184.

McClure, Bruce A., and Vernonica Franklin-Tong. 2006. "Gametophytic Self-Incompatibility: Understanding the Cellular Mechanisms Involved in 'Self' Pollen Tube Inhibition." *Planta*. https://doi.org/10.1007/s00425-006-0284-2.

McCubbin, Andrew G., Christina Lee, and Amy Hetrick. 2006. "Identification of Genes Showing Differential Expression between Morphs in Developing Flowers of Primula Vulgaris." *Sexual Plant Reproduction* 19 (2): 63–72. https://doi.org/10.1007/s00497-006-0022-8.

McDill, Joshua, Miriam Repplinger, Beryl B. Simpson, and Joachim W. Kadereit. 2009. "The Phylogeny of <I>Linum</I> and Linaceae Subfamily Linoideae, with Implications for Their Systematics, Biogeography, and Evolution of Heterostyly." *Systematic Botany* 34 (2): 386–405. https://doi.org/10.1600/036364409788606244.

Meeus, S., O. Honnay, R. Brys, and H. Jacquemyn. 2012. "Biased Morph Ratios and Skewed Mating Success Contribute to Loss of Genetic Diversity in the Distylous Pulmonaria Officinalis." *Annals of Botany* 109 (1): 227–35. https://doi.org/10.1093/aob/mcr272.

Miljuš-Đukić, J., S. Ninković, S. Radović, V. Maksimović, J. Brkljačić, and M. Nešković. 2004. "Detection of Proteins Possibly Involved in Self-Incompatibility Response in Distylous Buckwheat." *Biologia Plantarum* 48 (2): 293–96. https://doi.org/10.1023/B:BIOP.0000033459.48057.8b.

Minh, B. Q., M. A. T. Nguyen, and A. von Haeseler. 2013. "Ultrafast Approximation for Phylogenetic Bootstrap." *Molecular Biology and Evolution* 30 (5): 1188–95. https://doi.org/10.1093/molbev/mst024.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. https://doi.org/10.1038/nmeth.1226.

Muir, Alister D., and Neil D. Westcott. 2003. *Flax : The Genus Linum*. Routledge.

Murray, Brian G. 1986. "Floral Biology and Self-Incompatibility in Linum." *Botanical Gazette* 147 (3): 327–33. https://doi.org/10.1086/337599.

Nakasugi, Kenlee, Ross Crowhurst, Julia Bally, and Peter Waterhouse. 2014. "Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant Nicotiana Benthamiana." Edited by Omprakash Mittapalli. *PLoS ONE* 9 (3): e91776. https://doi.org/10.1371/journal.pone.0091776.

Nasrallah, J. B., T. H. Kao, M. L. Goldberg, and M. E. Nasrallah. 1985. "A CDNA Clone Encoding an S-Locus-Specific Glycoprotein from Brassica Oleracea." *Nature* 318 (6043): 263–67. https://doi.org/10.1038/318263a0.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Nicholls, Marc S. 1986. "POPULATION COMPOSITION, GENDER SPECIALIZATION, AND THE ADAPTIVE SIGNIFICANCE OF DISTYLY IN LINUM PERENNE (LINACEAE)." *New Phytologist* 102 (1): 209–17. https://doi.org/10.1111/j.1469-8137.1986.tb00811.x.

Nicholls, Marc S. 1985. "The Evolutionary Breakdown of Distyly in Linum Tenuifolium (Linaceae)." *Plant Systematics and Evolution*. Springer. https://doi.org/10.2307/23673690.

Nickrent, D.L., M. Costea, J.F. Barcelona, P.B. Pelser, and K. Nixon. 2006 onwards. PhytoImages.

Niklas, KJ. 1997. *The Evolutionary Biology of Plants*. Chicago: The University of Chicogo Press. https://books.google.co.uk/books?hl=en&lr=&id=2aSrw70rOKgC&oi=fnd&pg=PR9&dq=Niklas,+K.+J.+(1997)+The+evolutionary+biology+of+plants.+&ots=ClbyvKDwNN&sig=U0tq8n23SA800zqHaAWI2NAmZ2Q.

Nowak, Michael D, Giancarlo Russo, Ralph Schlapbach, Cuong Huu, Michael Lenhard, and Elena Conti. 2015. "The Draft Genome of Primula Veris Yields Insights into the Molecular Basis of Heterostyly." *Genome Biology* 16 (1): 12. https://doi.org/10.1186/s13059-014-0567-z.

Ornduff, R. 1992. "Historical Perspectives on Heterostyly." In , 31–39. https://doi.org/10.1007/978-3-642-86656-2_2.

Orsini, Luisa, Donald Gilbert, Ram Podicheti, Mieke Jansen, James B. Brown, Omid Shams Solari, Katina I. Spanier, et al. 2016. "Daphnia Magna Transcriptome by RNA-Seq across 12 Environmental Stressors." *Scientific Data* 3 (May): 160030. https://doi.org/10.1038/sdata.2016.30.

Oshlack, Alicia, and Matthew J. Wakefield. 2009. "Transcript Length Bias in RNA-Seq Data Confounds Systems Biology." *Biology Direct* 4 (April). https://doi.org/10.1186/1745-6150-4-14.

Pamela, V, and J Dowrick. 1956. "Heterostyly and Homostyly in Primula Obconica." *Heredity* 10 (2): 219–36. https://doi.org/10.1038/hdy.1956.19.

Pannell, John R., Marcel E. Dorken, and Sarah M. Eppley. 2005. "'Haldane's Sieve' in a Metapopulation: Sifting through Plant Reproductive Polymorphisms." *Trends in Ecology and Evolution* 20 (7): 374–79. https://doi.org/10.1016/j.tree.2005.05.004.

Parker, G. A., R. R. Baker, and V. G.F. Smith. 1972. "The Origin and Evolution of Gamete Dimorphism and the Male-Female Phenomenon." *Journal of Theoretical Biology* 36 (3): 529–53. https://doi.org/10.1016/0022-5193(72)90007-0.

Partek Inc. 2018. "Partek® Flow®." Partek Inc.

Paszkiewicz, K., and D. J. Studholme. 2010. "De Novo Assembly of Short Sequence Reads." *Briefings in Bioinformatics* 11 (5): 457–72. https://doi.org/10.1093/bib/bbq020.

Pérez-Barrales, Rocío, and J Arroyo. 2010. "Pollinator Shifts and the Loss of Style Polymorphism in Narcissus Papyraceus (Amaryllidaceae)." *Journal of Evolutionary Biology* 23 (6): 1117–28. https://doi.org/10.1111/j.1420-9101.2010.01988.x.

Pérez-Barrales, Rocío, Violeta I. Simón-Porcar, Rocío Santos-Gally, and Juan Arroyo. 2014. "Phenotypic Integration in Style Dimorphic Daffodils (Narcissus, Amaryllidaceae) with Different Pollinators." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1649). https://doi.org/10.1098/rstb.2013.0258.

Pérez, Rocío, Pablo Vargas, and Juan Arroyo. 2004. "Convergent Evolution of Flower Polymorphism in Narcissus (Amaryllidaceae)." *New Phytologist* 161 (1): 235–52. https://doi.org/10.1046/j.1469-8137.2003.00955.x.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, et al. 2003. "TIGR Gene Indices Clustering Tools (TGICL): A Software System for Fast Clustering of Large EST Datasets." *Bioinformatics* 19 (5): 651–52. https://doi.org/10.1093/bioinformatics/btg034.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. "SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions." *Nature Methods* 8 (10): 785–86. https://doi.org/10.1038/nmeth.1701.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner. 2007. "SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB." *Nucleic Acids Research* 35 (21): 7188–96. https://doi.org/10.1093/nar/gkm864.

Qiao, Hong, Fei Wang, Lan Zhao, Junli Zhou, Zhao Lai, Yansheng Zhang, Timothy P. Robbins, and Yongbiao Xue. 2004. "The F-Box Protein AhSLF-S 2 Controls the Pollen Function of S-RNase-Based Self-Incompatibility." *Plant Cell* 16 (9): 2307–22. https://doi.org/10.1105/tpc.104.024919.

Qiu, Yin Long, Jungho Lee, Fabiana Bernasconi-Quadroni, Douglas E. Soltis, Pamela S. Soltis, Michael Zanis, Elizabeth A. Zimmer, Zhiduan Chen, Vincent Savolainen, and Mark W. Chase. 1999. "The Earliest Angiosperms: Evidence from Mitochondrial, Plastid and Nuclear Genomes." *Nature* 402 (6760): 404–7. https://doi.org/10.1038/46536.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Rawat, Vimal, Ahmed Abdelsamad, Björn Pietzenuk, Danelle K. Seymour, Daniel Koenig, Detlef Weigel, Ales Pecinka, and Korbinian Schneeberger. 2015. "Improving the Annotation of Arabidopsis Lyrata Using RNA-Seq Data." Edited by Nicholas James Provart. *PLoS ONE* 10 (9): e0137391. https://doi.org/10.1371/journal.pone.0137391.

Renner, Susanne S. 2014. "The Relative and Absolute Frequencies of Angiosperm Sexual Systems: Dioecy, Monoecy, Gynodioecy, and an Updated Online Database." *American Journal of Botany* 101 (10): 1588–96. https://doi.org/10.3732/ajb.1400196.

Richards, A. J. 1997. *Plant Breeding Systems*. Chapman & Hall.

Richards, Jennifer H., and Suzanne Koptur. 1993. "Floral Variation and Distyly in Guettarda Scabra (Rubiaceae)." *American Journal of Botany* 80 (1): 31–40. https://doi.org/10.1002/j.1537-2197.1993.tb13764.x.

Richardson, James E., Frans M. Weitz, Michael F. Fay, Quentin C.B. Cronk, H. Peter Linder, G. Reeves, and Mark W. Chase. 2001. "Rapid and Recent Origin of Species Richness in the Cape Flora of South Africa." *Nature* 412 (6843): 181–83. https://doi.org/10.1038/35084067.

Roberts, Adam, Harold Pimentel, Cole Trapnell, and Lior Pachter. 2011. "Identification of Novel Transcripts in Annotated Genomes Using RNA-Seq." *Bioinformatics* 27 (17): 2325–29. https://doi.org/10.1093/bioinformatics/btr355.

Robertson, Ashley, Timothy C G Rich, Alexandra M Allen, Libby Houston, Cat Roberts, Jon R Bridle, Stephen A Harris, and Simon J Hiscock. 2010. "Hybridization and Polyploidy as Drivers of Continuing Evolution and Speciation in Sorbus." *Molecular Ecology* 19 (8): 1675–90. https://doi.org/10.1111/j.1365-294X.2010.04585.x.

Robertson, Gordon, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, et al. 2010. "De Novo Assembly and Analysis of RNA-Seq Data." *Nature Methods* 7 (11): 909–12. https://doi.org/10.1038/nmeth.1517.

Robinson, Mark D, and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rogers, C. M. 1979. "Distyly and Pollen Dimorphism in Linum Suffruticosum (Linaceae)." *Plant Systematics and Evolution* 131 (1–2): 127–32. https://doi.org/10.1007/BF00984126.

Rogers, C. M., R. Mildner, and B. D. Harris. 1972. "Some Additional Chromosome Numbers in the Linaceae." *Brittonia* 24 (3): 313–16. https://doi.org/10.2307/2805668.

Ruiz-Martín, J., R. Santos-Gally, M. Escudero, J. J. Midgley, R. Pérez-Barrales, and J. Arroyo. 2018. "Style Polymorphism in Linum (Linaceae): A Case of Mediterranean Parallel Evolution?" *Plant Biology* 20 (January): 100–111. https://doi.org/10.1111/plb.12670.

Sá, Túlio, Marco T. Furtado, Victoria Ferrero, Rocio Pérez-Barrales, Ebenézer B. Rodrigues, Isabela G. dos Santos, and Hélder Consolaro. 2016. "Floral Biology, Reciprocal Herkogamy and Breeding System in Four Psychotria Species (Rubiaceae) in Brazil." *Botanical Journal of the Linnean Society* 182 (3): 689–707. https://doi.org/10.1111/boj.12476.

Sales, Gabriele, Bruce E. Deagle, Enrica Calura, Paolo Martini, Alberto Biscontin, Cristiano De Pittà, So Kawaguchi, et al. 2017. "KrillDB: A de Novo Transcriptome Database for the Antarctic Krill (Euphausia Superba)." Edited by Cristiano Bertolucci. *PLOS ONE* 12 (2): e0171908. https://doi.org/10.1371/journal.pone.0171908.

Sánchez, J M, V Ferrero, J Arroyo, and L Navarro. 2010. "Patterns of Style Polymorphism in Five Species of the South African Genus Nivenia (Iridaceae)." *Annals of Botany* 106 (2): 321–31. https://doi.org/10.1093/aob/mcq111.

Sanchez, Jose M., Victoria Ferrero, and Luis Navarro. 2008. "A New Approach to the Quantification of Degree of Reciprocity in Distylous (Sensu Lato) Plant Populations." *Annals of Botany* 102 (3): 463–72. https://doi.org/10.1093/aob/mcn111.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri. 2012. "NIH Image to ImageJ: 25 Years of Image Analysis." *Nature Methods*. https://doi.org/10.1038/nmeth.2089.

Schopfer, C R, M E Nasrallah, and J B Nasrallah. 1999. "The Male Determinant of Self-Incompatibility in Brassica." *Science (New York, N.Y.)* 286 (5445): 1697–1700. https://doi.org/10.1126/science.286.5445.1697.

Schulz, Marcel H, Daniel R Zerbino, Martin Vingron, and Ewan Birney. 2012. "Oases: Robust de Novo RNA-Seq Assembly across the Dynamic Range of Expression Levels." *Bioinformatics (Oxford, England)* 28 (8): 1086–92. https://doi.org/10.1093/bioinformatics/bts094.

Seetharam, A. 1972. “Interspecific Hybridization in Linum.” *Euphytica* 21 (3): 489–95. https://doi.org/10.1007/BF00039344.

Seetharam, A., and D. Srinivasachar. 1972. “Cytomorphological Studies in the Genus Linum.” *CYTOLOGIA* 37 (4): 661–71. https://doi.org/10.1508/cytologia.37.661.

Shore, Joel S, and Spencer C H Barrett. 1985. “The Genetics of Distyly and Homostyly in Turnera Ulmifolia L. (Turneraceae).” *Heredity* 55: 167–74. https://www.nature.com/articles/hdy198588.pdf.

Sijacic, Paja, Xi Wang, Andrea L Skirpan, Yan Wang, Peter E Dowd, Andrew G McCubbin, Shihshieh Huang, and Teh-Hui Kao. 2004. “Identification of the Pollen Determinant of S-RNase-Mediated Self-Incompatibility.” *Nature* 429 (6989): 302–5. https://doi.org/10.1038/nature02523.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Simón-Porcar, Violeta I. 2018. “Late-Acting Self-Incompatibility and a Narrow Floral Tube as Selective Forces for Stylar Dimorphism in Narcissus (Amaryllidaceae).” *Iee* 11: 64–73. https://doi.org/10.4033/iee.2018.11.8.n.

Simón-Porcar, Violeta I., Rocío Santos-Gally, and Juan Arroyo. 2014. “Long-Tongued Insects Promote Disassortative Pollen Transfer in Style-Dimorphic Narcissus Papyraceus (Amaryllidaceae).” *Journal of Ecology* 102 (1): 116–25. https://doi.org/10.1111/1365-2745.12179.

Society, G Hillman - Proceedings of the Prehistoric, and undefined 1975. n.d. “The Plant Remains from Tell Abu Hureyra: A Preliminary Report.”

Soltis, P S, D E Soltis, and M W Chase. 1999. “Angiosperm Phylogeny Inferred from Multiple Genes as a Tool for Comparative Biology.” *Nature* 402 (6760): 402–4. https://doi.org/10.1038/46528.

Song, Li, and Liliana Florea. 2015. “Rcorrector: Efficient and Accurate Error Correction for Illumina RNA-Seq Reads.” *GigaScience* 4 (1): 48. https://doi.org/10.1186/s13742-015-0089-y.

Song, Li, Liliana Florea, and Ben Langmead. 2014. “Lighter: Fast and Memory-Efficient Sequencing Error Correction without Counting.” *Genome Biology* 15 (11): 509. https://doi.org/10.1186/s13059-014-0509-9.

Stebbins, G. Ledyard. 1957. “Self Fertilization and Population Variability in the Higher Plants.” *The American Naturalist* 91 (861): 337–54. https://doi.org/10.1086/281999.

Stein, J C, R Dixit, M E Nasrallah, and J B Nasrallah. 1996. “SRK, the Stigma-Specific S Locus Receptor Kinase of Brassica, Is Targeted to the Plasma Membrane in Transgenic Tobacco.” *The Plant Cell* 8 (3): 429–45. https://doi.org/10.1105/tpc.8.3.429.

Stein, J. C., B. Howlett, D. C. Boyes, M. E. Nasrallah, and J. B. Nasrallah. 1991. "Molecular Cloning of a Putative Receptor Protein Kinase Gene Encoded at the Self-Incompatibility Locus of Brassica Oleracea." *Proceedings of the National Academy of Sciences of the United States of America* 88 (19): 8816–20. https://doi.org/10.1073/pnas.88.19.8816.

Stone, Sophia L., Erin M. Anderson, Robert T. Mullen, and Daphne R. Goring. 2003. "ARC1 Is an E3 Ubiquitin Ligase and Promotes the Ubiquitination of Proteins during the Rejection of Self-Incompatible Brassica Pollen." *Plant Cell* 15 (4): 885–98. https://doi.org/10.1105/tpc.009845.

Stubbs, Thomas M., Marc Jan Bonder, Anne-Katrien Stark, Felix Krueger, Ferdinand von Meyenn, Oliver Stegle, and Wolf Reik. 2017. "Multi-Tissue DNA Methylation Age Predictor in Mouse." *Genome Biology* 18 (1): 68. https://doi.org/10.1186/s13059-017-1203-5.

Sun, Ge, Qiang Ji, David L Dilcher, Shaolin Zheng, Kevin C Nixon, and Xinfu Wang. 2002. "Archaefructaceae, a New Basal Angiosperm Family." *Science (New York, N.Y.)* 296 (5569): 899–904. https://doi.org/10.1126/science.1069439.

Sun, Yu, Xi-Ying Fan, Dong-Mei Cao, Wenqiang Tang, Kun He, Jia-Ying Zhu, Jun-Xian He, et al. 2010. "Integration of Brassinosteroid Signal Transduction with the Transcription Network for Plant Growth Regulation in Arabidopsis." *Developmental Cell* 19 (5): 765–77. https://doi.org/10.1016/j.devcel.2010.10.010.

Surget-Groba, Yann, and Juan I Montoya-Burgos. 2010. "Optimization of de Novo Transcriptome Assembly from Next-Generation Sequencing Data." *Genome Research* 20 (10): 1432–40. https://doi.org/10.1101/gr.103846.109.

Szövényi, Péter, Nicolas Devos, David J. Weston, Xiaohan Yang, Zsófia Hock, Jonathan A. Shaw, Kentaro K. Shimizu, Stuart F. McDaniel, and Andreas Wagner. 2014. "Efficient Purging of Deleteriousmutations in Plants with Haploid Selfing." *Genome Biology and Evolution* 6 (5): 1238–52. https://doi.org/10.1093/gbe/evu099.

Szymanski, Maciej, Andrzej Zielezinski, Jan Barciszewski, Volker A. Erdmann, and Wojciech M. Karlowski. 2016. "5SRNAdb: An Information Resource for 5S Ribosomal RNAs." *Nucleic Acids Research* 44 (D1): D180–83. https://doi.org/10.1093/nar/gkv1081.

Takasaki, Takeshi, Katsunori Hatakeyama, Go Suzuki, Masao Watanabe, Akira Isogai, and Kokichi Hinata. 2000. "The S Receptor Kinase Determines Self-Incompatibility in Brassica Stigma." *Nature* 403 (6772): 913–16. https://doi.org/10.1038/35002628.

Takayama, S, H Shimosato, H Shiba, M Funato, F S Che, M Watanabe, M Iwano, and A Isogai. 2001. "Direct Ligand-Receptor Complex Interaction Controls Brassica Self-Incompatibility." *Nature* 413 (6855): 534–38. https://doi.org/10.1038/35097104.

Takayama, Seiji, Hiroshi Shiba, Megumi Iwano, Hiroko Shimosato, Fang Sik Che, Naoko Kai, Masao Watanabe, Go Suzuki, Kokichi Hinata, and Akira Isogai. 2000.

"The Pollen Determinant of Self-Incompatibility in Brassica Campestris." *Proceedings of the National Academy of Sciences of the United States of America* 97 (4): 1920–25. https://doi.org/10.1073/pnas.040556397.

Todd, Erica V., Michael A. Black, and Neil J. Gemmell. 2016. "The Power and Promise of RNA-Seq in Ecology and Evolution." *Molecular Ecology* 25 (6): 1224–41. https://doi.org/10.1111/mec.13526.

Tsuchimatsu, Takashi, Keita Suwabe, Rie Shimizu-Inatsugi, Sachiyo Isokawa, Pavlos Pavlidis, Thomas Städler, Go Suzuki, Seiji Takayama, Masao Watanabe, and Kentaro K Shimizu. 2010. "Evolution of Self-Compatibility in Arabidopsis by a Mutation in the Male Specificity Gene." *Nature* 464 (7293): 1342–46. https://doi.org/10.1038/nature08927.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, et al. 2006. "The Genome of Black Cottonwood, Populus Trichocarpa (Torr. & Gray)." *Science* 313 (5793): 1596–1604. https://doi.org/10.1126/science.1128691.

Ushijima, Koichiro, Kazuo Ikeda, Ryohei Nakano, Miyo Matsubara, Yuri Tsuda, and Yasutaka Kubo. 2015. "Genetic Control of Floral Morph and Petal Pigmentation in Linum Grandiflorum Desf., a Heterostylous Flax." *The Horticulture Journal* 84 (3): 261–68. https://doi.org/10.2503/hortj.MI-045.

Ushijima, Koichiro, Ryohei Nakano, Mayu Bando, Yukari Shigezane, Kazuo Ikeda, Yuko Namba, Saori Kume, Toshiyuki Kitabata, Hitoshi Mori, and Yasutaka Kubo. 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." *The Plant Journal* 69 (2): 317–31. https://doi.org/10.1111/j.1365-313X.2011.04792.x.

Uyenoyama, Marcy K. 2004. "Evolution under Tight Linkage to Mating Type." *New Phytologist* 165 (1): 63–70. https://doi.org/10.1111/j.1469-8137.2004.01246.x.

Vaisey-Genser, M, and D. H. Morris. 2003. "Introduction: History of the Cultivation and Uses of Flaxseed." In *Flax: The Genus Linum*, 13–33. CRC Press. https://doi.org/10.1201/9780203437506-5.

Valdés-Castrillón, B., S. Talavera-Lozano, and F.E. Fernández-Galiano. 1987. *Flora Vascular de Andalucía Occidental, 3 Volúmenes*. Barcelona, España: Ketres Editora SA.

Maaten, Laurens Van Der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research*. Vol. 9. http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Vandenbussche, Michiel, Jan Zethof, Stefan Royaert, Koen Weterings, and Tom Gerats. 2004. "The Duplicated B-Class Heterodimer Model: Whorl-Specific Effects and Complex Genetic Interactions in Petunia Hybrida Flower Development." *Plant Cell* 16 (3): 741–54. https://doi.org/10.1105/tpc.019166.

Vijay, Nagarjun, Jelmer W. Poelstra, Axel Künstner, and Jochen B. W. Wolf. 2013. "Challenges and Strategies in Transcriptome Assembly and Differential Gene

Expression Quantification. A Comprehensive *in Silico* Assessment of RNA-Seq Experiments." *Molecular Ecology* 22 (3): 620–34. https://doi.org/10.1111/mec.12014.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. 2012. "Measurement of MRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent among Samples." *Theory in Biosciences* 131 (4): 281–85. https://doi.org/10.1007/s12064-012-0162-3.

Wajid, Bilal, and Erchin Serpedin. 2012. "Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers." *Genomics, Proteomics & Bioinformatics* 10 (2): 58–73. https://doi.org/10.1016/J.GPB.2012.05.006.

Wan, Yizhen, Kai Tang, Dayong Zhang, Shaojun Xie, Xiaohong Zhu, Zegang Wang, and Zhaobo Lang. 2015. "Transcriptome-Wide High-Throughput Deep M6A-Seq Reveals Unique Differential M6A Methylation Patterns between Three Organs in Arabidopsis Thaliana." *Genome Biology* 16 (1): 272. https://doi.org/10.1186/s13059-015-0839-2.

Warden, Charles D, Yate-Ching Yuan, and Xiwei Wu. 2013. "Optimal Calculation of RNA-Seq Fold-Change Values." *International Journal of Computational Bioinformatics and In Silico Modeling*. Vol. 2. http://www.bioconductor.org/.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. 2009. "Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench." *Bioinformatics* 25 (9): 1189–91. https://doi.org/10.1093/bioinformatics/btp033.

Wheeler, Michael J, Barend H J de Graaf, Natalie Hadjiosif, Ruth M Perry, Natalie S Poulter, Kim Osman, Sabina Vatovec, Andrea Harper, F Christopher H Franklin, and Vernonica E Franklin-Tong. 2009. "Identification of the Pollen Self-Incompatibility Determinant in Papaver Rhoeas." *Nature* 459 (7249): 992–95. https://doi.org/10.1038/nature08027.

Williams, Joseph H., and William E. Friedman. 2002. "Identification of Diploid Endosperm in an Early Angiosperm Lineage." *Nature* 415 (6871): 522–26. https://doi.org/10.1038/415522a.

Wolfe, L. M. 2001. "Associations among Multiple Floral Polymorphisms in Linum Pubescens (Linaceae), a Heterostylous Plant." *International Journal of Plant Sciences* 162 (2): 335–42. https://doi.org/10.1086/319578.

Wos, G., and Y. Willi. 2018. "Thermal Acclimation in Arabidopsis Lyrata: Genotypic Costs and Transcriptional Changes." *Journal of Evolutionary Biology* 31 (1): 123–35. https://doi.org/10.1111/jeb.13208.

Xie, Y., G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, et al. 2014. "SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics* 30 (12): 1660–66. https://doi.org/10.1093/bioinformatics/btu077.

Yasui, Yasuo, Hideki Hirakawa, Mariko Ueno, Katsuhiro Matsui, Tomoyuki Katsube-Tanaka, Soo Jung Yang, Jotaro Aii, Shingo Sato, and Masashi Mori. 2016. "Assembly of the Draft Genome of Buckwheat and Its Applications in Identifying Agronomically Useful Genes." *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes* 23 (3): 215–24. https://doi.org/10.1093/dnares/dsw012.

Yasui, Yasuo, Masashi Mori, Jotaro Aii, Tomoko Abe, Daiki Matsumoto, Shingo Sato, Yoriko Hayashi, Ohmi Ohnishi, and Tatsuya Ota. 2012. "S-LOCUS EARLY FLOWERING 3 Is Exclusively Present in the Genomes of Short-Styled Buckwheat Plants That Exhibit Heteromorphic Self-Incompatibility." Edited by Edward Newbigin. *PLoS ONE* 7 (2): e31264. https://doi.org/10.1371/journal.pone.0031264.

Yasui, Yasuo, Yingjie Wang, Ohmi Ohnishi, and Clayton G Campbell. 2004. "Amplified Fragment Length Polymorphism Linkage Analysis of Common Buckwheat (*Fagopyrum Esculentum*) and Its Wild Self-Pollinated Relative *Fagopyrum Homotropicum*." *Genome* 47 (2): 345–51. https://doi.org/10.1139/g03-126.

Ye, Jia, Lin Fang, Hongkun Zheng, Yong Zhang, Jie Chen, Zengjin Zhang, Jing Wang, et al. 2006. "WEGO: A Web Tool for Plotting GO Annotations." *Nucleic Acids Research* 34 (WEB. SERV. ISS.). https://doi.org/10.1093/nar/gkl031.

Ye, Jia, Yong Zhang, Huihai Cui, Jiawei Liu, Yuqing Wu, Yun Cheng, Huixing Xu, et al. 2018. "WEGO 2.0: A Web Tool for Analyzing and Plotting GO Annotations, 2018 Update." *Nucleic Acids Research* 46: 71–75. https://doi.org/10.1093/nar/gky400.

Yuan, Shuai, Spencer C H Barrett, Tingting Duan, Xin Qian, Miaomiao Shi, and Dianxiang Zhang. 2017. "Ecological Correlates and Genetic Consequences of Evolutionary Transitions from Distyly to Homostyly." *Annals of Botany* 120 (5): 775–89. https://doi.org/10.1093/aob/mcx098.

Zerbino, D. R., and E. Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhao, Qiong-Yi, Yi Wang, Yi-Meng Kong, Da Luo, Xuan Li, and Pei Hao. 2011. "Optimizing de Novo Transcriptome Assembly from Short-Read RNA-Seq Data: A Comparative Study." *BMC Bioinformatics* 12 (Suppl 14): S2. https://doi.org/10.1186/1471-2105-12-S14-S2.

Zhou, Wei, Spencer C. H. Barrett, Hong Wang, and De-Zhu Li. 2015. "Reciprocal Herkogamy Promotes Disassortative Mating in a Distylous Species with Intramorph Compatibility." *New Phytologist* 206 (4): 1503–12. https://doi.org/10.1111/nph.13326.

Zhu, Xing-Fu, Xian-Feng Jiang, Li Li, Zhi-Qiang Zhang, and Qing-Jun Li. 2015. "Asymmetrical Disassortative Pollination in a Distylous Primrose: The Complementary Roles of Bumblebee Nectar Robbers and Syrphid Flies." *Scientific Reports* 5 (1): 7721. https://doi.org/10.1038/srep07721.

Zohary, D., and M. Hopf. 2000. "Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley." *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley.*, no. Ed.3.

Zyprych-Walczak, J, A Szabelska, L Handschuh, K Górczak, K Klamecka, M Figlerowicz, and I Siatkowski. 2015. "The Impact of Normalization Methods on RNA-Seq Data Analysis." *BioMed Research International* 2015 (June): 621690. https://doi.org/10.1155/2015/621690.

Appendix I

Table S1 Location data for Linum tenue sample populations. Collectors were Dr Adrian Brennan (ACB), Mr Alireza Foroozani (AF), and Dr Rocío Pérez-Barrales (RPB). Experiments were field (F) or glasshouse (G) measures.

Location	Name	Year	Collector	Latitude	Longitude	Experiment
Manga Villaluenga	mva	2015	RPB	36.6886	-5.4095	F
Puerto de las Palomas	pdp	2015	RPB	36.7881	-5.3764	F
Ronda 10 km	r10	2015	RPB	36.7814	-5.0946	F
Ronda 17 km	r17	2015	RPB	36.7933	-4.9913	F
Sierra de las Nieves	sdn	2015	RPB	36.6619	-5.7725	F
Alhaurin de al Torre	alt	2013	ACB	36.6687	-4.5543	G
Aracena	ara	2014	AF	37.9414	-6.5249	G
Burguillos	bur	2014	AF	37.5950	-5.9753	G
Cazalla de la Sierra	caz	2014	AF	37.9374	-5.7612	G
Cabra	cbt	2013	ACB	37.4671	-4.4229	G
El Bosque	ebo	2014	AF	36.7774	-5.5180	G
El Burgo	elb	2014	AF	36.6275	-4.9901	G
Grazalema	grt	2013	ACB	36.8189	-5.3463	G
Hinojales	hin	2014	AF	38.0129	-6.5872	G
La Zubia	laz	2014	AF	37.1142	-3.5747	G
La Umbria	lum	2014	AF	37.8623	-6.4809	G
Mairena del Aljarafe	mda	2014	AF	37.3414	-6.0476	G
Monachil	mon	2014	AF	37.1368	-3.5077	G
Pinos Genil	pig	2014	AF	37.1609	-3.5123	G
Sierra Nevada	snv	2014	AF	37.1389	-3.4609	G
Sevilla	svt	2013	ACB	37.3553	-5.9909	G

Appendix II

Table S2 Mixed model analysis results for *all* Linum tenue open flower measurements. Mixed models were performed on non-transformed data using the lmer REML fit function of the R lmerTest. The *p* values of mixed effects were evaluated using t-tests with Satterthwaite degrees of freedom approximations, while the *p* values of random effects were evaluated by sequentially dropping random effects from the model and comparing the prior model using the anova function with likelihood ratio tests. R^2 values were calculated using the r.squaredGLMM function of the R MuMIn package and are either conditional for the full mixed model or marginal for fixed effects only. Shaded rows indicate trait results presented in Table 1.

Response	Random effects	No. obs.	Variance	SD	*P* value	R^2 cond.	Fixed effects	Estimate	SE	*p* value	R^2 marg.
Pistil length	individual x population	150	0.117	0.343	**9.66^{e-15}**	0.955	intercept	4.822	0.241	**$<2.00^{e-16}$**	0.916
							morph	-2.009	0.392	**4.94^{e-07}**	
	population	16	0.001	0.024	**1.50^{e-02}**		petal length	0.191	0.017	**$<2.00^{e-16}$**	
	residual		0.140	0.374			morph*petal length	-0.099	0.027	**3.52^{e-04}**	
Stamen length	individual x population	115	0.106	0.325	**2.65^{e-07}**	0.905	intercept	3.418	0.334	**$<2.00^{e-16}$**	0.841
							morph	1.868	0.483	**1.35^{e-04}**	
	population	16	0.004	0.061	**2.92^{e-02}**		petal length	0.142	0.024	**1.19^{e-08}**	
	residual		0.159	0.399			morph*petal length	0.024	0.034	4.76^{e-01}	
Herkogamy (pistil – stamen)	individual x population	115	0.141	0.375	**$<2.00^{e-16}$**	0.986	intercept	1.658	0.307	**1.39^{e-07}**	0.968
							morph	-4.143	0.443	**$<2.00^{e-16}$**	
	population	16	<0.001	<0.001	7.71^{e-02}		petal length	0.030	0.022	**1.85^{e-01}**	
	residual		0.119	0.345			morph*petal length	-0.105	0.031	**9.44^{e-04}**	
Ovary+Style	individual x population	131	0.067	0.259	**5.95^{e-07}**	0.949	intercept	4.291	0.231	**$<2.00^{e-16}$**	0.918
		16	0.009	0.097	**8.11^{e-04}**		morph	-2.136	0.372	**2.40^{e-08}**	
	population		0.123	0.350			petal length	0.160	0.017	**$<2.00^{e-16}$**	
	residual						morph*petal length	-0.065	0.026	**1.30^{e-02}**	

Filament	individual x population population population residual	131 16	0.090 0.016 0.145	0.300 0.127 0.381	**5.78^{e-08}** **3.65^{e-04}**	0.922	intercept morph petal length morph*petal length	2.812 1.859 0.118 0.036	0.256 0.411 0.018 0.029	**<2.00^{e-16}** **9.06^{e-06}** **5.92^{e-10}** 2.07^{e-01}	0.865
Herkogamy (ovary+style – filament)	individual x population population population residual	131 16	0.084 0.030 0.119	0.289 0.175 0.346	**2.10^{e-08}** **4.26^{e-06}**	0.984	intercept morph petal length morph*petal length	1.541 -4.086 -0.036 -0.094	0.242 0.382 0.017 0.027	**7.57^{e-10}** **<2.00^{e-16}** **3.45^{e-02}** **4.97^{e-04}**	0.969
Stigma length	individual x population population population residual	131 16	0.012 0.002 0.018	0.111 0.045 0.133	**3.85^{e-09}** **7.89^{e-04}**	0.784	intercept morph petal length morph*petal length	0.651 -0.156 0.026 -0.021	0.091 0.146 0.007 0.010	**8.61^{e-12}** 2.87^{e-01} **1.01^{e-04}** **3.95^{e-02}**	0.608
Anther length	individual x population population population residual	131 16	0.006 0.002 0.011	0.075 0.044 0.106	**1.03^{e-07}** **1.16^{e-05}**	0.440	intercept morph petal length morph*petal length	1.057 -0.090 0.012 0.002	0.069 0.111 0.005 0.008	**<2.00^{e-16}** 4.21^{e-01} **1.48^{e-02}** 7.91^{e-01}	0.066
Petal width	individual x population population population residual	150 16	0.289 <0.001 0.512	0.538 <0.001 0.715	**4.88^{e-10}** 4.29^{e-01}	0.613	intercept morph petal length morph*petal length	1.692 0.097 0.367 -0.015	0.433 0.704 0.031 0.049	**1.11^{e-04}** 8.91^{e-01} **<2.00^{e-16}** 7.59^{e-01}	0.394
Sepal length	individual x population population population residual	129 16	0.094 0.035 0.240	0.306 0.186 0.490	**9.12^{e-04}** **2.65^{e-03}**	0.457	intercept morph petal length morph*petal length	3.007 0.014 0.134 -0.020	0.321 0.510 0.023 0.036	**<2.00^{e-16}** 9.78^{e-01} **1.59^{e-08}** 5.85^{e-01}	0.166

Appendix III

List of features found to be differentially expressed in the full dataset, filtered for a minimum fold change of +/- 2 and and FDR threshold of 0.05.

Contig_129342	Contig_145822	Contig_020380	Contig_159123	Contig_162660	Contig_132262
Contig_125780	Contig_028132	Contig_071319	Contig_015074	Contig_054620	Contig_007980
Contig_095635	Contig_080751	Contig_071368	Contig_112734	Contig_083136	Contig_005292
Contig_168818	Contig_086326	Contig_042165	Contig_123146	Contig_124597	Contig_070336
Contig_151932	Contig_107300	Contig_078474	Contig_082492	Contig_117373	Contig_142068
Contig_070269	Contig_128263	Contig_007465	Contig_063789	Contig_127015	Contig_018454
Contig_117644	Contig_127921	Contig_062262	Contig_007994	Contig_069675	Contig_133359
Contig_167741	Contig_001428	Contig_090758	Contig_046535	Contig_119682	Contig_079997
Contig_054975	Contig_122937	Contig_167359	Contig_106753	Contig_036859	Contig_097512
Contig_129990	Contig_108580	Contig_060016	Contig_146722	Contig_109163	Contig_121521
Contig_097696	Contig_167732	Contig_051786	Contig_091415	Contig_097699	Contig_126342
Contig_026462	Contig_020391	Contig_125868	Contig_028320	Contig_030985	Contig_060754
Contig_008558	Contig_004732	Contig_143046	Contig_170321	Contig_066254	Contig_008310
Contig_164093	Contig_094272	Contig_062164	Contig_061177	Contig_043195	Contig_127147
Contig_078102	Contig_043888	Contig_145326	Contig_088122	Contig_044970	Contig_127185
Contig_127925	Contig_159228	Contig_005294	Contig_035675	Contig_056928	Contig_089689
Contig_060228	Contig_039640	Contig_125027	Contig_105745	Contig_117739	Contig_002368
Contig_143758	Contig_079510	Contig_125317	Contig_015125	Contig_098552	Contig_056649
Contig_089334	Contig_047709	Contig_108576	Contig_126446	Contig_164619	Contig_128751
Contig_019094	Contig_050956	Contig_095114	Contig_010314	Contig_118266	Contig_110734
Contig_096616	Contig_113249	Contig_048993	Contig_016548	Contig_106723	Contig_002877
Contig_119744	Contig_040308	Contig_165581	Contig_159291	Contig_063071	Contig_136084
Contig_064925	Contig_070598	Contig_007940	Contig_147374	Contig_035537	Contig_161110
Contig_032223	Contig_156142	Contig_073540	Contig_138051	Contig_133178	Contig_080750
Contig_133387	Contig_110007	Contig_094689	Contig_046520	Contig_055302	Contig_082150
Contig_083134	Contig_013610	Contig_157342	Contig_134252	Contig_005491	Contig_130560
Contig_109724	Contig_125708	Contig_098553	Contig_074745	Contig_130465	Contig_168264
Contig_111750	Contig_100473	Contig_110428	Contig_006411	Contig_170224	Contig_139887
Contig_157631	Contig_068833	Contig_120246	Contig_083024	Contig_077819	Contig_069541
Contig_013606	Contig_119500	Contig_055251	Contig_007413	Contig_159973	Contig_164084
Contig_007469	Contig_090982	Contig_165440	Contig_073149	Contig_118351	Contig_032661
Contig_168293	Contig_159112	Contig_129297	Contig_041721	Contig_100938	Contig_054205
Contig_146603	Contig_144826	Contig_130200	Contig_007287	Contig_037924	Contig_149880
Contig_161189	Contig_099636	Contig_152598	Contig_158273	Contig_123204	Contig_113244
Contig_128032	Contig_035016	Contig_035536	Contig_163216	Contig_009754	Contig_093445
Contig_039635	Contig_129913	Contig_066053	Contig_087698	Contig_149416	Contig_112094
Contig_138449	Contig_155937	Contig_156638	Contig_110712	Contig_084397	Contig_130344
Contig_154292	Contig_055576	Contig_124571	Contig_089333	Contig_155738	Contig_018946
Contig_033030	Contig_126742	Contig_088121	Contig_065112	Contig_106649	Contig_036206
Contig_081191	Contig_121053	Contig_155880	Contig_004911	Contig_022634	Contig_151385
Contig_157894	Contig_145024	Contig_048185	Contig_087702	Contig_041836	Contig_091825
Contig_142554	Contig_111529	Contig_040489	Contig_054654	Contig_166208	Contig_147459
Contig_148818	Contig_123175	Contig_043783	Contig_024453	Contig_138622	Contig_000845
Contig_049345	Contig_063076	Contig_163208	Contig_135046	Contig_015126	Contig_141643
Contig_094968	Contig_020299	Contig_128342	Contig_102066	Contig_160685	Contig_046527
Contig_109337	Contig_117889	Contig_148911	Contig_131607	Contig_011810	Contig_052617
Contig_032672	Contig_088790	Contig_151288	Contig_141869	Contig_141795	Contig_088662
Contig_005542	Contig_034595	Contig_154849	Contig_108758	Contig_125380	Contig_075481
Contig_033726	Contig_147164	Contig_157562	Contig_114740	Contig_083141	Contig_139098
Contig_047708	Contig_156090	Contig_135277	Contig_046757	Contig_086787	Contig_035676
Contig_151821	Contig_102000	Contig_131551	Contig_067934	Contig_143028	Contig_013777
Contig_111660	Contig_040535	Contig_049185	Contig_034894	Contig_029286	Contig_087691
Contig_133650	Contig_109942	Contig_015300	Contig_033579	Contig_167758	Contig_109239
Contig_121005	Contig_052552	Contig_071637	Contig_021778	Contig_162365	Contig_079699
Contig_000974	Contig_152729	Contig_083814	Contig_167020	Contig_133605	Contig_113608
Contig_022504	Contig_161921	Contig_036483	Contig_117067	Contig_135847	Contig_014490
Contig_153510	Contig_087869	Contig_043971	Contig_120081	Contig_151804	Contig_062218
Contig_041186	Contig_060830	Contig_020379	Contig_072173	Contig_138290	Contig_030123
Contig_062948	Contig_096613	Contig_147478	Contig_046068	Contig_019060	Contig_091846
Contig_123443	Contig_084066	Contig_009484	Contig_082353	Contig_014600	Contig_136309
Contig_073543	Contig_163189	Contig_018474	Contig_027388	Contig_060749	Contig_089335
Contig_043973	Contig_149150	Contig_038173	Contig_073948	Contig_115748	Contig_042156

Contig_117933	Contig_033629	Contig_032302	Contig_100573	Contig_116680	Contig_032666
Contig_148814	Contig_085335	Contig_130327	Contig_163898	Contig_142551	Contig_019194
Contig_046533	Contig_073126	Contig_059286	Contig_158109	Contig_042507	Contig_059564
Contig_145893	Contig_069589	Contig_161341	Contig_060609	Contig_138464	Contig_024296
Contig_122961	Contig_064306	Contig_164899	Contig_017678	Contig_046600	Contig_124789
Contig_163983	Contig_127646	Contig_043885	Contig_021686	Contig_137380	Contig_136257
Contig_098780	Contig_155413	Contig_004977	Contig_131853	Contig_115720	Contig_078762
Contig_096947	Contig_041237	Contig_075704	Contig_069702	Contig_168366	Contig_165481
Contig_075744	Contig_145322	Contig_002871	Contig_117048	Contig_120032	Contig_056890
Contig_169649	Contig_056144	Contig_102410	Contig_056679	Contig_003030	Contig_019522
Contig_062854	Contig_129291	Contig_076480	Contig_131818	Contig_085804	Contig_124574
Contig_060363	Contig_081662	Contig_057000	Contig_050198	Contig_122851	Contig_049376
Contig_031878	Contig_156596	Contig_167047	Contig_069036	Contig_052906	Contig_055706
Contig_066919	Contig_136421	Contig_092693	Contig_159856	Contig_059595	Contig_119722
Contig_164114	Contig_051600	Contig_039271	Contig_162321	Contig_005524	Contig_086899
Contig_025047	Contig_130353	Contig_085895	Contig_131783	Contig_026410	Contig_160450
Contig_095002	Contig_115086	Contig_139970	Contig_155584	Contig_059516	Contig_039483
Contig_028586	Contig_100512	Contig_025914	Contig_054880	Contig_017648	Contig_102254
Contig_088653	Contig_105244	Contig_089306	Contig_022684	Contig_019833	Contig_090545
Contig_015850	Contig_081105	Contig_028515	Contig_054738	Contig_166463	Contig_010489
Contig_029939	Contig_049081	Contig_011148	Contig_101519	Contig_033468	Contig_053596
Contig_080421	Contig_168325	Contig_069345	Contig_077737	Contig_164734	Contig_059952
Contig_153351	Contig_024750	Contig_111768	Contig_121336	Contig_004614	Contig_169260
Contig_039634	Contig_079780	Contig_044315	Contig_091821	Contig_045464	Contig_012516
Contig_122347	Contig_165740	Contig_029686	Contig_125672	Contig_029001	Contig_003878
Contig_109504	Contig_020764	Contig_099734	Contig_155974	Contig_067491	Contig_056122
Contig_016578	Contig_030299	Contig_161065	Contig_115196	Contig_164836	Contig_152913
Contig_035319	Contig_154813	Contig_089888	Contig_128008	Contig_108180	Contig_053745
Contig_093025	Contig_105216	Contig_102327	Contig_118004	Contig_166682	Contig_107465
Contig_005186	Contig_130429	Contig_107417	Contig_090154	Contig_044141	Contig_151290
Contig_100155	Contig_144132	Contig_082975	Contig_149751	Contig_088658	Contig_132053
Contig_007557	Contig_050646	Contig_158442	Contig_004970	Contig_138778	Contig_034521
Contig_056895	Contig_012510	Contig_124030	Contig_156420	Contig_071606	Contig_077843
Contig_040216	Contig_062955	Contig_088774	Contig_026142	Contig_151678	Contig_046378
Contig_065713	Contig_022651	Contig_112826	Contig_065452	Contig_161533	Contig_009406
Contig_069506	Contig_081523	Contig_015979	Contig_156346	Contig_138975	Contig_027822
Contig_053624	Contig_141279	Contig_142976	Contig_079988	Contig_093997	Contig_133203
Contig_045747	Contig_085368	Contig_113995	Contig_087788	Contig_093046	Contig_038587
Contig_054826	Contig_059338	Contig_128220	Contig_110222	Contig_050039	Contig_160217
Contig_078313	Contig_061503	Contig_133535	Contig_165379	Contig_137538	Contig_148031
Contig_113766	Contig_058891	Contig_000708	Contig_135520	Contig_020373	Contig_026371
Contig_048663	Contig_040331	Contig_114322	Contig_082156	Contig_013481	Contig_165820
Contig_066035	Contig_150122	Contig_118837	Contig_050276	Contig_127576	Contig_071886
Contig_151202	Contig_020483	Contig_136543	Contig_106421	Contig_157199	Contig_147050
Contig_137214	Contig_162742	Contig_066433	Contig_119550	Contig_141203	Contig_135405
Contig_091705	Contig_038849	Contig_132075	Contig_023461	Contig_028036	Contig_115383
Contig_039648	Contig_041247	Contig_120132	Contig_013474	Contig_101938	Contig_019321
Contig_002283	Contig_071635	Contig_017522	Contig_163924	Contig_076839	Contig_086557
Contig_066510	Contig_020016	Contig_135930	Contig_139370	Contig_128790	Contig_032074
Contig_104634	Contig_095500	Contig_155942	Contig_014025	Contig_015947	Contig_033580
Contig_156263	Contig_158714	Contig_021079	Contig_165135	Contig_141239	Contig_016665
Contig_050275	Contig_157618	Contig_036710	Contig_026662	Contig_028939	Contig_118267
Contig_111648	Contig_023729	Contig_073645	Contig_141903	Contig_076575	Contig_040309
Contig_116856	Contig_028569	Contig_007975	Contig_074322	Contig_035307	Contig_125669
Contig_125503	Contig_005746	Contig_118668	Contig_021704	Contig_012259	Contig_023541
Contig_039925	Contig_138423	Contig_003397	Contig_058347	Contig_095020	Contig_010563
Contig_037414	Contig_099633	Contig_107911	Contig_002872	Contig_026696	Contig_031709
Contig_001506	Contig_114946	Contig_044330	Contig_062088	Contig_147340	Contig_127772
Contig_091409	Contig_163598	Contig_008102	Contig_075703	Contig_082568	Contig_148844
Contig_126598	Contig_156823	Contig_022736	Contig_129430	Contig_086636	Contig_117638
Contig_041527	Contig_124992	Contig_078857	Contig_156564	Contig_161396	Contig_157599
Contig_106764	Contig_016233	Contig_019973	Contig_137229	Contig_018411	Contig_045574
Contig_062453	Contig_139244	Contig_118341	Contig_100127	Contig_057529	Contig_002112
Contig_023372	Contig_139246	Contig_043062	Contig_141932	Contig_063810	Contig_036628
Contig_132371	Contig_078535	Contig_129274	Contig_057128	Contig_152793	Contig_091513
Contig_143365	Contig_088661	Contig_128258	Contig_067809	Contig_116367	Contig_079013
Contig_068628	Contig_145205	Contig_147627	Contig_158166	Contig_036277	Contig_034840
Contig_066828	Contig_031150	Contig_075702	Contig_077886	Contig_019192	Contig_144859
Contig_114682	Contig_064187	Contig_022267	Contig_105284	Contig_073810	Contig_018434
Contig_166310	Contig_049651	Contig_078548	Contig_153417	Contig_082587	Contig_074714

Contig_095342	Contig_005269	Contig_102785	Contig_166010	Contig_062182	Contig_024325
Contig_137559	Contig_019642	Contig_091903	Contig_110235	Contig_056741	Contig_100577
Contig_033676	Contig_057493	Contig_076068	Contig_064952	Contig_115758	Contig_047793
Contig_003316	Contig_015036	Contig_081577	Contig_108150	Contig_098117	Contig_049526
Contig_065244	Contig_063572	Contig_046752	Contig_160592	Contig_068679	Contig_050866
Contig_028120	Contig_000907	Contig_163054	Contig_162912	Contig_049060	Contig_012653
Contig_092119	Contig_117799	Contig_112302	Contig_132375	Contig_053528	Contig_139979
Contig_035538	Contig_155192	Contig_113028	Contig_055704	Contig_157620	Contig_043337
Contig_124577	Contig_145877	Contig_107825	Contig_092702	Contig_169561	Contig_047387
Contig_041771	Contig_050072	Contig_118215	Contig_032024	Contig_019065	Contig_016441
Contig_157475	Contig_140273	Contig_088990	Contig_040732	Contig_031306	Contig_019239
Contig_136402	Contig_085781	Contig_127022	Contig_096158	Contig_011098	Contig_021324
Contig_130806	Contig_131475	Contig_090754	Contig_058761	Contig_049484	Contig_163287
Contig_047901	Contig_030274	Contig_097824	Contig_137912	Contig_041833	Contig_054674
Contig_048062	Contig_061626	Contig_115460	Contig_055252	Contig_113742	Contig_076447
Contig_062656	Contig_029293	Contig_146092	Contig_059289	Contig_078547	Contig_078885
Contig_148040	Contig_063398	Contig_078481	Contig_093615	Contig_159889	Contig_074890
Contig_075883	Contig_143118	Contig_115732	Contig_122335	Contig_003844	Contig_001315
Contig_128449	Contig_103087	Contig_129057	Contig_015062	Contig_008750	Contig_108870
Contig_116517	Contig_031872	Contig_043356	Contig_045966	Contig_052786	Contig_069875
Contig_090261	Contig_089693	Contig_156007	Contig_129212	Contig_145050	Contig_149652
Contig_027377	Contig_124687	Contig_061412	Contig_048094	Contig_052686	Contig_158572
Contig_002798	Contig_119819	Contig_004759	Contig_059371	Contig_006610	Contig_117162
Contig_165677	Contig_046820	Contig_041990	Contig_021445	Contig_054026	Contig_060440
Contig_154062	Contig_085397	Contig_001873	Contig_100868	Contig_135125	Contig_028987
Contig_113143	Contig_124399	Contig_043066	Contig_047677	Contig_101790	Contig_161499
Contig_103568	Contig_117656	Contig_099848	Contig_121603	Contig_120637	Contig_023735
Contig_057961	Contig_110229	Contig_155319	Contig_032759	Contig_051741	Contig_052860
Contig_149718	Contig_005335	Contig_041613	Contig_144197	Contig_154358	Contig_112675
Contig_141560	Contig_089032	Contig_142376	Contig_098699	Contig_165449	Contig_114878
Contig_027391	Contig_101937	Contig_090716	Contig_115009	Contig_119321	Contig_113597
Contig_013462	Contig_042226	Contig_061082	Contig_098615	Contig_063438	Contig_045203
Contig_135612	Contig_052278	Contig_000724	Contig_087699	Contig_148165	Contig_019350
Contig_101505	Contig_003407	Contig_068063	Contig_045877	Contig_074269	Contig_085322
Contig_154157	Contig_035039	Contig_143581	Contig_035093	Contig_132184	Contig_061693
Contig_044739	Contig_054676	Contig_009955	Contig_020758	Contig_026881	Contig_136699
Contig_064735	Contig_112391	Contig_151688	Contig_095380	Contig_005777	Contig_056897
Contig_021312	Contig_048638	Contig_143362	Contig_039384	Contig_147747	Contig_057267
Contig_077870	Contig_040969	Contig_033181	Contig_117095	Contig_153641	Contig_050859
Contig_109470	Contig_049497	Contig_026361	Contig_050230	Contig_155022	Contig_002507
Contig_135116	Contig_157309	Contig_100684	Contig_132895	Contig_021837	Contig_125598
Contig_067204	Contig_019681	Contig_149117	Contig_167593	Contig_094089	Contig_170226
Contig_025913	Contig_103522	Contig_050123	Contig_058427	Contig_136894	Contig_146280
Contig_018626	Contig_099720	Contig_147396	Contig_098747	Contig_146290	Contig_035426
Contig_110438	Contig_063824	Contig_024056	Contig_044319	Contig_013461	Contig_074823
Contig_092745	Contig_066822	Contig_145451	Contig_147705	Contig_161968	Contig_074501
Contig_154133	Contig_108005	Contig_020651	Contig_169809	Contig_155838	Contig_128997
Contig_106043	Contig_164761	Contig_131221	Contig_101824	Contig_100497	Contig_018471
Contig_129703	Contig_170142	Contig_142271	Contig_133693	Contig_043966	Contig_135448
Contig_036147	Contig_102250	Contig_124814	Contig_115685	Contig_158093	Contig_018690
Contig_164307	Contig_046529	Contig_010384	Contig_168604	Contig_083118	Contig_065469
Contig_026230	Contig_002867	Contig_162400	Contig_144745	Contig_005274	Contig_052337
Contig_049961	Contig_015949	Contig_119670	Contig_008469	Contig_158334	Contig_135796
Contig_055275	Contig_114933	Contig_073600	Contig_045156	Contig_155972	Contig_055339
Contig_010025	Contig_123832	Contig_086763	Contig_094342	Contig_076546	Contig_165945
Contig_092279	Contig_004467	Contig_135732	Contig_095370	Contig_079541	Contig_082976
Contig_019518	Contig_037944	Contig_009734	Contig_000833	Contig_087414	Contig_146142
Contig_072586	Contig_052631	Contig_159761	Contig_087132	Contig_148309	Contig_088789
Contig_055340	Contig_078092	Contig_064649	Contig_053064	Contig_163296	Contig_129409
Contig_127852	Contig_155199	Contig_042521	Contig_114392	Contig_128050	Contig_048230
Contig_048908	Contig_125877	Contig_086500	Contig_158886	Contig_101690	Contig_127196
Contig_016130	Contig_061680	Contig_161945	Contig_058527	Contig_071246	Contig_022277
Contig_165280	Contig_166873	Contig_150063	Contig_007592	Contig_078475	Contig_097521
Contig_153338	Contig_023634	Contig_133242	Contig_042983	Contig_121427	Contig_074866
Contig_158693	Contig_002286	Contig_096433	Contig_011345	Contig_019440	Contig_004961
Contig_040330	Contig_048495	Contig_118204	Contig_165626	Contig_125499	Contig_130873
Contig_096880	Contig_023570	Contig_054632	Contig_042895	Contig_062190	Contig_043869
Contig_114245	Contig_126084	Contig_134725	Contig_101627	Contig_105379	Contig_164005
Contig_005293	Contig_055081	Contig_161328	Contig_060081	Contig_081824	Contig_108364
Contig_019845	Contig_007724	Contig_146114	Contig_084683	Contig_064109	Contig_062174

Contig_002741	Contig_107230	Contig_140707	Contig_032078	Contig_074490	Contig_149964
Contig_095250	Contig_095539	Contig_040484	Contig_053804	Contig_007184	Contig_087696
Contig_070507	Contig_139704	Contig_033423	Contig_049804	Contig_161757	Contig_142809
Contig_146350	Contig_010838	Contig_037608	Contig_156103	Contig_052191	Contig_062177
Contig_164521	Contig_129311	Contig_003884	Contig_147573	Contig_119907	Contig_060112
Contig_074431	Contig_073594	Contig_091353	Contig_106922	Contig_140076	Contig_022017
Contig_007468	Contig_055331	Contig_060971	Contig_132255	Contig_055664	Contig_125330
Contig_079154	Contig_055887	Contig_010912	Contig_049579	Contig_114268	Contig_117852
Contig_050133	Contig_116890	Contig_083924	Contig_039650	Contig_168046	Contig_025450
Contig_057368	Contig_075435	Contig_098359	Contig_140551	Contig_002088	Contig_002279
Contig_129319	Contig_140384	Contig_114794	Contig_052346	Contig_007169	Contig_101157
Contig_135179	Contig_085888	Contig_005792	Contig_150211	Contig_049936	Contig_099144
Contig_086558	Contig_004973	Contig_059717	Contig_150707	Contig_157280	Contig_025538
Contig_106877	Contig_170411	Contig_124575	Contig_098722	Contig_120316	Contig_076661
Contig_004044	Contig_091762	Contig_090815	Contig_085377	Contig_044269	Contig_129776
Contig_148897	Contig_073729	Contig_016555	Contig_163943	Contig_056796	Contig_053854
Contig_087961	Contig_092007	Contig_085621	Contig_081525	Contig_046485	Contig_020200
Contig_119794	Contig_116122	Contig_014264	Contig_123329	Contig_029502	Contig_159873
Contig_054073	Contig_036344	Contig_066265	Contig_068186	Contig_154568	Contig_115664
Contig_165890	Contig_038688	Contig_094090	Contig_070307	Contig_016025	Contig_146751
Contig_152998	Contig_086183	Contig_079805	Contig_154136	Contig_150728	Contig_127738
Contig_102858	Contig_161404	Contig_151406	Contig_008714	Contig_031154	Contig_065743
Contig_053795	Contig_135671	Contig_153851	Contig_077489	Contig_130328	Contig_028789
Contig_061015	Contig_020762	Contig_161747	Contig_028720	Contig_053129	Contig_080100
Contig_049223	Contig_027514	Contig_099935	Contig_099142	Contig_148615	Contig_113119
Contig_106252	Contig_075699	Contig_016599	Contig_151785	Contig_064645	Contig_006957
Contig_065627	Contig_078732	Contig_154965	Contig_082626	Contig_146291	Contig_032835
Contig_150907	Contig_135254	Contig_040642	Contig_079279	Contig_029670	Contig_058423
Contig_082344	Contig_056233	Contig_030845	Contig_089311	Contig_169953	Contig_087832
Contig_039307	Contig_127452	Contig_024696	Contig_159035	Contig_136002	Contig_060157
Contig_011377	Contig_058994	Contig_162320	Contig_021478	Contig_081866	Contig_031037
Contig_159845	Contig_035615	Contig_086396	Contig_102062	Contig_067876	Contig_099332
Contig_142235	Contig_115604	Contig_049251	Contig_126333	Contig_150950	Contig_137915
Contig_133825	Contig_013936	Contig_123169	Contig_001295	Contig_126310	Contig_084292
Contig_041139	Contig_121168	Contig_140646	Contig_168160	Contig_086909	Contig_104408
Contig_051728	Contig_122337	Contig_065619	Contig_147577	Contig_028223	Contig_023204
Contig_028037	Contig_058628	Contig_090326	Contig_153027	Contig_079550	Contig_093926
Contig_137000	Contig_056417	Contig_093827	Contig_156317	Contig_121561	Contig_073592
Contig_136075	Contig_167851	Contig_127304	Contig_150368	Contig_129576	Contig_123149
Contig_100967	Contig_100123	Contig_037374	Contig_020983	Contig_039185	Contig_033108
Contig_023720	Contig_127044	Contig_068236	Contig_149301	Contig_082010	Contig_147281
Contig_152316	Contig_091412	Contig_128843	Contig_050898	Contig_113084	Contig_129600
Contig_073683	Contig_122839	Contig_003875	Contig_152273	Contig_108583	Contig_014594
Contig_030461	Contig_095982	Contig_156677	Contig_137960	Contig_040010	Contig_161999
Contig_059130	Contig_079548	Contig_010022	Contig_058301	Contig_137499	Contig_084384
Contig_093879	Contig_076298	Contig_012493	Contig_020376	Contig_159205	Contig_088262
Contig_087787	Contig_031983	Contig_072346	Contig_154771	Contig_070226	Contig_045121
Contig_144373	Contig_039626	Contig_000624	Contig_082452	Contig_110139	Contig_060158
Contig_073963	Contig_102409	Contig_155837	Contig_127574	Contig_167632	Contig_167222
Contig_069690	Contig_070115	Contig_044313	Contig_020985	Contig_006524	Contig_135892
Contig_158942	Contig_152070	Contig_166676	Contig_167874	Contig_018476	Contig_090138
Contig_155637	Contig_065246	Contig_044142	Contig_018040	Contig_071861	Contig_142234
Contig_164506	Contig_147555	Contig_054577	Contig_097697	Contig_115282	Contig_089536
Contig_041191	Contig_076132	Contig_016531	Contig_012464	Contig_005916	Contig_049759
Contig_040780	Contig_166677	Contig_009065	Contig_064393	Contig_098880	Contig_108508
Contig_050389	Contig_160114	Contig_158896	Contig_067869	Contig_146626	Contig_154693
Contig_132319	Contig_149411	Contig_139986	Contig_109912	Contig_010889	Contig_155415
Contig_018575	Contig_126777	Contig_102577	Contig_122962	Contig_023425	Contig_034390
Contig_046530	Contig_052689	Contig_060095	Contig_086181	Contig_151986	Contig_141429
Contig_108894	Contig_003776	Contig_052566	Contig_072798	Contig_042516	Contig_134106
Contig_008712	Contig_077717	Contig_131434	Contig_019703	Contig_098073	Contig_064017
Contig_139674	Contig_150158	Contig_130897	Contig_042124	Contig_002401	Contig_136647
Contig_107797	Contig_114984	Contig_165244	Contig_131499	Contig_059572	Contig_129348
Contig_136914	Contig_145214	Contig_044697	Contig_000975	Contig_111936	Contig_145642
Contig_008078	Contig_033386	Contig_097582	Contig_013590	Contig_110614	Contig_054162
Contig_089349	Contig_076714	Contig_002092	Contig_165590	Contig_003100	Contig_085879
Contig_047876	Contig_029785	Contig_009448	Contig_101439	Contig_005060	Contig_096539
Contig_111447	Contig_156751	Contig_055685	Contig_124976	Contig_145760	Contig_159189
Contig_161938	Contig_108026	Contig_158021	Contig_040075	Contig_043090	Contig_021767
Contig_106531	Contig_011145	Contig_152872	Contig_043353	Contig_149931	Contig_077116

Contig_075470	Contig_160061	Contig_110471	Contig_147279	Contig_049711	Contig_029298
Contig_019542	Contig_130289	Contig_002353	Contig_112070	Contig_054637	Contig_064396
Contig_017968	Contig_043819	Contig_144747	Contig_023632	Contig_093606	Contig_099934
Contig_086570	Contig_045535	Contig_144435	Contig_108504	Contig_148869	Contig_018541
Contig_048634	Contig_091015	Contig_036482	Contig_008858	Contig_156584	Contig_136136
Contig_006838	Contig_031653	Contig_048832	Contig_123522	Contig_054643	Contig_005276
Contig_107849	Contig_002548	Contig_015443	Contig_063982	Contig_123111	Contig_024622
Contig_007162	Contig_144194	Contig_124486	Contig_128629	Contig_027830	Contig_097972
Contig_000893	Contig_156784	Contig_040310	Contig_166652	Contig_041702	Contig_048138
Contig_071140	Contig_066988	Contig_078934	Contig_129565	Contig_128087	Contig_139867
Contig_124270	Contig_058158	Contig_141357	Contig_036043	Contig_002981	Contig_014718
Contig_074681	Contig_119838	Contig_152394	Contig_044494	Contig_056239	Contig_032665
Contig_137289	Contig_091047	Contig_078053	Contig_146327	Contig_136898	Contig_090013
Contig_090483	Contig_120949	Contig_097427	Contig_041474	Contig_062160	Contig_082922
Contig_053837	Contig_056342	Contig_146455	Contig_119414	Contig_055249	Contig_003359
Contig_040380	Contig_135813	Contig_042644	Contig_014038	Contig_007141	Contig_080639
Contig_049355	Contig_036834	Contig_069573	Contig_143169	Contig_106700	Contig_162007
Contig_046596	Contig_020858	Contig_015678	Contig_095796	Contig_126867	Contig_002190
Contig_024446	Contig_064289	Contig_044321	Contig_047386	Contig_118248	Contig_084992
Contig_085692	Contig_114039	Contig_105603	Contig_093016	Contig_017398	Contig_112374
Contig_155966	Contig_026091	Contig_000066	Contig_091583	Contig_062784	Contig_052773
Contig_106750	Contig_056680	Contig_049363	Contig_153444	Contig_004569	Contig_057789
Contig_039922	Contig_040608	Contig_135617	Contig_007848	Contig_011163	Contig_002968
Contig_062837	Contig_078455	Contig_054925	Contig_161455	Contig_082078	Contig_059966
Contig_139961	Contig_059499	Contig_066316	Contig_143760	Contig_068583	Contig_090123
Contig_155772	Contig_019538	Contig_003507	Contig_123200	Contig_105896	Contig_130204
Contig_008612	Contig_097756	Contig_145176	Contig_010713	Contig_142371	Contig_032959
Contig_107934	Contig_131765	Contig_069340	Contig_078675	Contig_096292	Contig_157372
Contig_041691	Contig_001864	Contig_014945	Contig_124851	Contig_167938	Contig_112557
Contig_118172	Contig_035665	Contig_077063	Contig_041810	Contig_001722	Contig_105774
Contig_091019	Contig_050263	Contig_115228	Contig_120708	Contig_122666	Contig_045042
Contig_044766	Contig_161235	Contig_013768	Contig_138968	Contig_051793	Contig_025394
Contig_153986	Contig_050450	Contig_098190	Contig_158032	Contig_134373	Contig_143174
Contig_135960	Contig_053687	Contig_072041	Contig_138364	Contig_093625	Contig_046605
Contig_032659	Contig_043263	Contig_155941	Contig_053244	Contig_090392	Contig_158424
Contig_135604	Contig_127761	Contig_023426	Contig_002203	Contig_005513	Contig_005163
Contig_118916	Contig_146634	Contig_130390	Contig_043939	Contig_075234	Contig_144315
Contig_138311	Contig_027229	Contig_069845	Contig_169754	Contig_142830	Contig_021491
Contig_168527	Contig_049086	Contig_061909	Contig_013604	Contig_095678	Contig_051735
Contig_050041	Contig_063719	Contig_107265	Contig_064211	Contig_098474	Contig_109509
Contig_068783	Contig_146018	Contig_142930	Contig_079492	Contig_161900	Contig_089172
Contig_020229	Contig_112681	Contig_011339	Contig_164543	Contig_165705	Contig_018129
Contig_020055	Contig_018309	Contig_085194	Contig_045211	Contig_003374	Contig_042023
Contig_051705	Contig_165261	Contig_139522	Contig_065644	Contig_070355	Contig_064069
Contig_141257	Contig_077893	Contig_033810	Contig_110989	Contig_135693	Contig_164406
Contig_148713	Contig_074843	Contig_088270	Contig_107485	Contig_013298	Contig_083626
Contig_022645	Contig_037227	Contig_031304	Contig_151092	Contig_078507	Contig_012179
Contig_020838	Contig_157401	Contig_163639	Contig_138071	Contig_093533	Contig_009981
Contig_112750	Contig_056662	Contig_048772	Contig_006804	Contig_022118	Contig_158932
Contig_148833	Contig_065156	Contig_056195	Contig_096135	Contig_020375	Contig_071136
Contig_099758	Contig_079281	Contig_003104	Contig_078956	Contig_092766	Contig_095508
Contig_103432	Contig_065418	Contig_077321	Contig_103403	Contig_115425	Contig_167867
Contig_103009	Contig_046448	Contig_066321	Contig_135685	Contig_146889	Contig_071970
Contig_145147	Contig_031447	Contig_066824	Contig_022888	Contig_160402	Contig_144773
Contig_070138	Contig_054762	Contig_001033	Contig_151429	Contig_116354	Contig_152822
Contig_135190	Contig_166696	Contig_166148	Contig_111993	Contig_008116	Contig_144088
Contig_046606	Contig_040846	Contig_028319	Contig_118854	Contig_050814	Contig_106216
Contig_013873	Contig_085983	Contig_049237	Contig_126372	Contig_037410	Contig_123322
Contig_142905	Contig_113925	Contig_027824	Contig_143272	Contig_045762	Contig_125838
Contig_163442	Contig_093601	Contig_106684	Contig_041893	Contig_077368	Contig_162060
Contig_082628	Contig_140308	Contig_036196	Contig_002428	Contig_116244	Contig_081005
Contig_111991	Contig_161046	Contig_052594	Contig_080947	Contig_123650	Contig_082908
Contig_034646	Contig_080205	Contig_143264	Contig_170023	Contig_018338	Contig_055186
Contig_158042	Contig_033882	Contig_031084	Contig_036888	Contig_142762	Contig_032794
Contig_092991	Contig_101497	Contig_166776	Contig_043424	Contig_084895	Contig_014442
Contig_060746	Contig_163476	Contig_065819	Contig_088282	Contig_073649	Contig_162726
Contig_025046	Contig_012683	Contig_048297	Contig_097698	Contig_062755	Contig_155876
Contig_133959	Contig_117338	Contig_047697	Contig_137733	Contig_145421	Contig_148449
Contig_020504	Contig_076306	Contig_138666	Contig_160193	Contig_112329	Contig_154872
Contig_013814	Contig_054911	Contig_028123	Contig_073038	Contig_051784	Contig_131839

Contig_085784	Contig_020812	Contig_094914	Contig_073431	Contig_076289	Contig_082208
Contig_022006	Contig_162107	Contig_025903	Contig_123164	Contig_155569	Contig_010734
Contig_089275	Contig_087654	Contig_015793	Contig_158497	Contig_115567	Contig_019943
Contig_154484	Contig_115066	Contig_064153	Contig_003657	Contig_083085	Contig_002460
Contig_146893	Contig_112540	Contig_145014	Contig_112056	Contig_039918	Contig_058986
Contig_121928	Contig_117330	Contig_003194	Contig_126112	Contig_021076	Contig_054090
Contig_133565	Contig_090860	Contig_158383	Contig_137932	Contig_153580	Contig_128546
Contig_046508	Contig_109237	Contig_123030	Contig_012245	Contig_078152	Contig_144066
Contig_044773	Contig_102491	Contig_064832	Contig_000761	Contig_160710	Contig_140950
Contig_101297	Contig_153227	Contig_061922	Contig_152414	Contig_143877	Contig_019193
Contig_062762	Contig_146261	Contig_025305	Contig_077481	Contig_066003	Contig_020494
Contig_116973	Contig_154305	Contig_039314	Contig_072923	Contig_071001	Contig_162076
Contig_137636	Contig_113687	Contig_138114	Contig_119860	Contig_032649	Contig_003491
Contig_054083	Contig_121845	Contig_053264	Contig_006040	Contig_148116	Contig_150316
Contig_040490	Contig_142090	Contig_076799	Contig_085451	Contig_165862	Contig_107177
Contig_046848	Contig_157921	Contig_054652	Contig_051894	Contig_007870	Contig_003918
Contig_050086	Contig_112813	Contig_042171	Contig_083091	Contig_055746	Contig_104565
Contig_038723	Contig_019385	Contig_114883	Contig_008480	Contig_161866	Contig_022425
Contig_052896	Contig_066566	Contig_158005	Contig_033668	Contig_038927	Contig_018028
Contig_135864	Contig_125629	Contig_008307	Contig_055271	Contig_088378	Contig_020739
Contig_004744	Contig_059500	Contig_013556	Contig_032932	Contig_092905	Contig_028830
Contig_050900	Contig_068666	Contig_034138	Contig_070687	Contig_045483	Contig_138358
Contig_088845	Contig_006869	Contig_012335	Contig_159458	Contig_058305	Contig_003530
Contig_152178	Contig_017552	Contig_085028	Contig_094535	Contig_020383	Contig_137121
Contig_094031	Contig_128845	Contig_097505	Contig_097822	Contig_085137	Contig_036114
Contig_015004	Contig_053125	Contig_073591	Contig_163821	Contig_106648	Contig_090894
Contig_107267	Contig_095544	Contig_053077	Contig_118875	Contig_148952	Contig_137330
Contig_016092	Contig_091429	Contig_155200	Contig_008152	Contig_169423	Contig_101584
Contig_008359	Contig_086901	Contig_122506	Contig_169931	Contig_129312	Contig_060261
Contig_117502	Contig_100866	Contig_015946	Contig_108671	Contig_120284	Contig_166757
Contig_002449	Contig_117773	Contig_040276	Contig_164032	Contig_003032	Contig_165824
Contig_146580	Contig_070356	Contig_083131	Contig_150569	Contig_044761	Contig_078428
Contig_062213	Contig_066678	Contig_089033	Contig_155701	Contig_005798	Contig_095139
Contig_060386	Contig_035164	Contig_061526	Contig_149092	Contig_065461	Contig_086852
Contig_048779	Contig_062810	Contig_087710	Contig_118972	Contig_133660	Contig_041146
Contig_072145	Contig_019640	Contig_066992	Contig_114687	Contig_073881	Contig_013526
Contig_069601	Contig_148908	Contig_069575	Contig_107348	Contig_046258	Contig_040335
Contig_145912	Contig_083696	Contig_170530	Contig_024042	Contig_051775	Contig_153939
Contig_083754	Contig_055393	Contig_056187	Contig_127317	Contig_049288	Contig_095050
Contig_132905	Contig_145586	Contig_070025	Contig_155288	Contig_134282	Contig_040757
Contig_143789	Contig_071352	Contig_091905	Contig_160595	Contig_036168	Contig_107026
Contig_143993	Contig_082545	Contig_154691	Contig_098813	Contig_046954	Contig_036464
Contig_161130	Contig_004704	Contig_034043	Contig_063426	Contig_113761	Contig_047249
Contig_097518	Contig_116690	Contig_140942	Contig_030637	Contig_159240	Contig_044343
Contig_020422	Contig_118385	Contig_129636	Contig_059037	Contig_074055	Contig_074974
Contig_055197	Contig_011538	Contig_162629	Contig_162648	Contig_057928	Contig_011853
Contig_062326	Contig_088025	Contig_051711	Contig_060405	Contig_056902	Contig_022668
Contig_061977	Contig_144085	Contig_037363	Contig_158267	Contig_115691	Contig_118307
Contig_158514	Contig_128343	Contig_004109	Contig_074447	Contig_044535	Contig_141165
Contig_119493	Contig_031098	Contig_041252	Contig_145825	Contig_080218	Contig_041673
Contig_134219	Contig_105864	Contig_089105	Contig_061481	Contig_160272	
Contig_153655	Contig_158614	Contig_162895	Contig_127515	Contig_148034	

Appendix IV

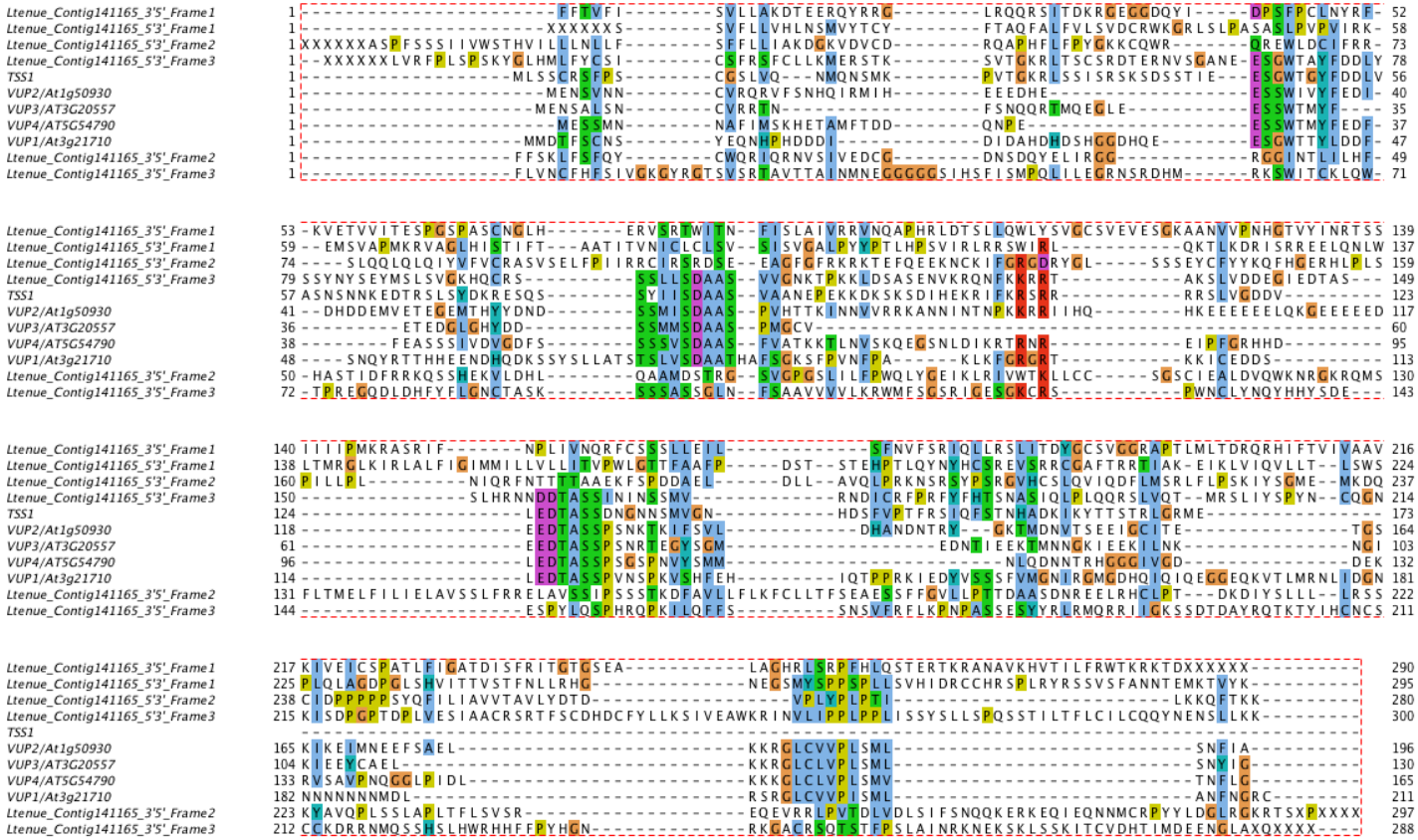


Figure S1 Alignment of full-length *Linum tenue* contigs Ltenue_Contig141165_5'3'_Frame1-3, Ltenue_Contig141165_3'5'_Frame1-3, TSS1 and VUP1-4. The sequnces are sorted by their position on the calculated tree (Fig S2) Alignments were made using MUSCLE v3.8.31, and amino acid residues coloured according to their physicochemical properties (Clustalx colouring system). M1-4 indicate conserved sequence motifs, as defined by Grienenberger & Douglas (2014), among VUP1 homologs.

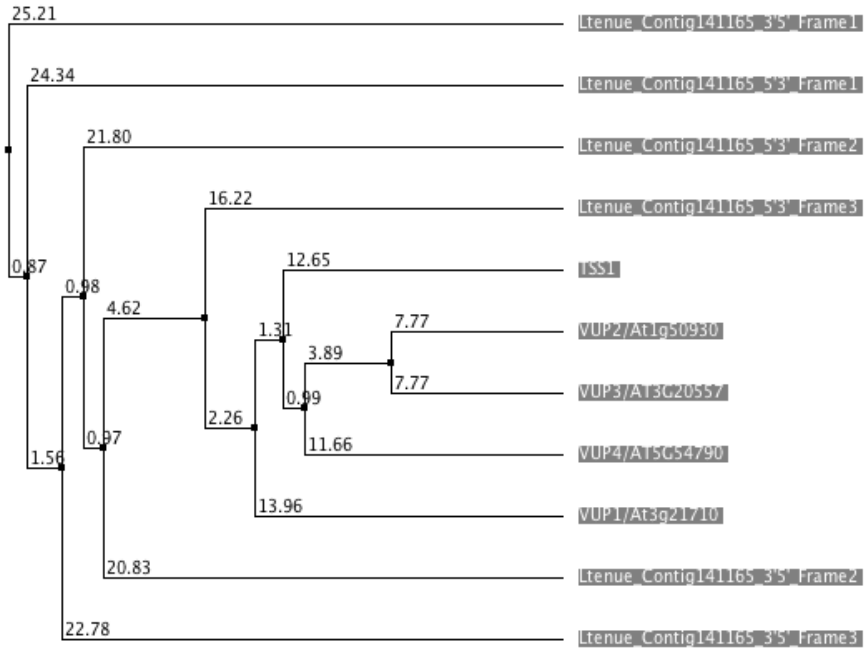


Figure S2 UPGMA tree of average distance for calculated using percentage sequence similarity with Jalview v2.9. Full-length *Linum tenue* contigs Ltenue_Contig141165_5'3'_Frame1-3, Ltenue_Contig141165_3'5'_Frame1-3, *TSS1 and VUP1-4*.

Appendix VI

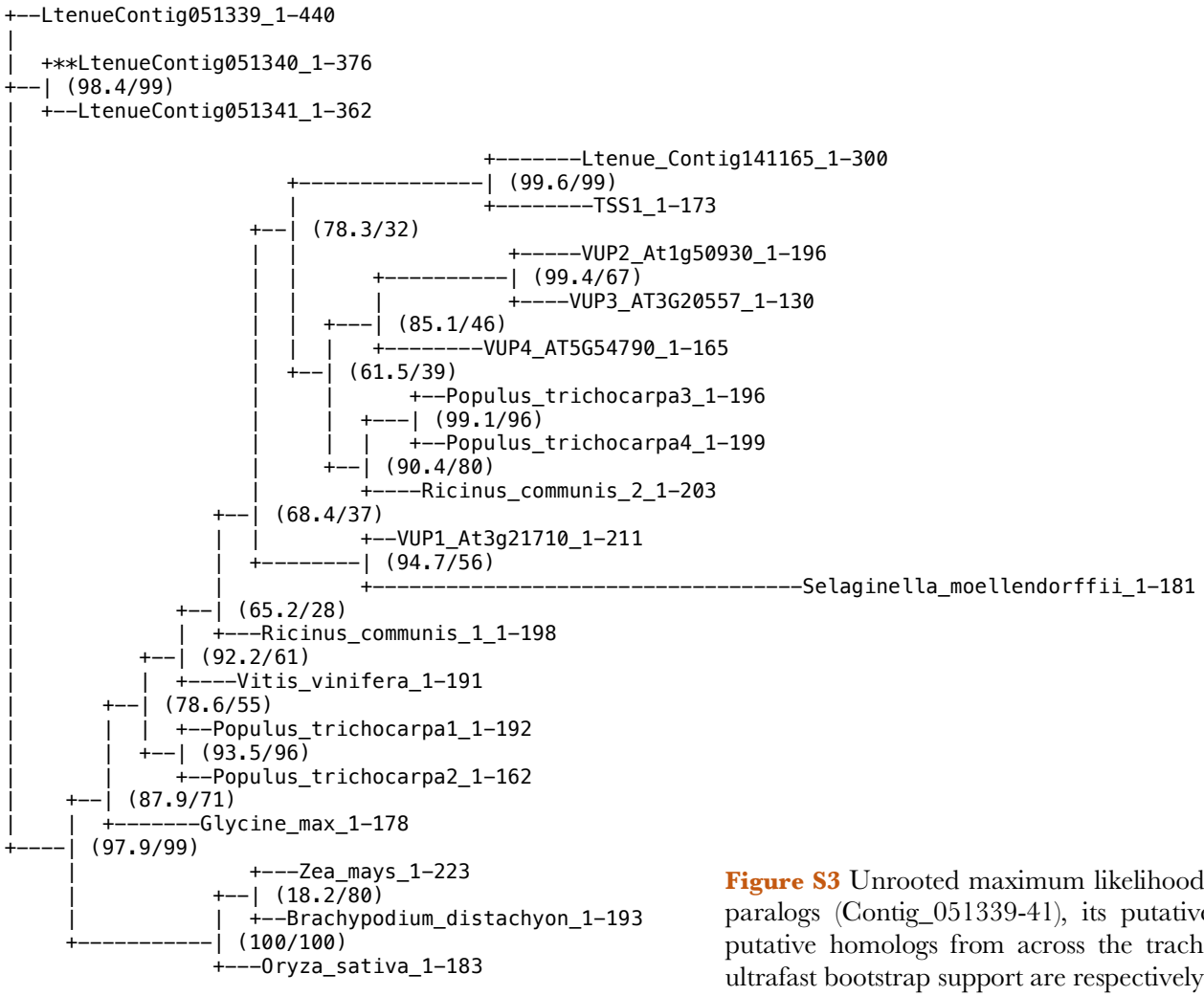


Figure S3 Unrooted maximum likelihood tree of L. tenue Contig_141165, its putative paralogs (Contig_051339-41), its putative ortholog in *L. grandiflorum* (TSS1) and its putative homologs from across the tracheophyta. Percentage SH-aRLT support and ultrafast bootstrap support are respectively shown for each node.

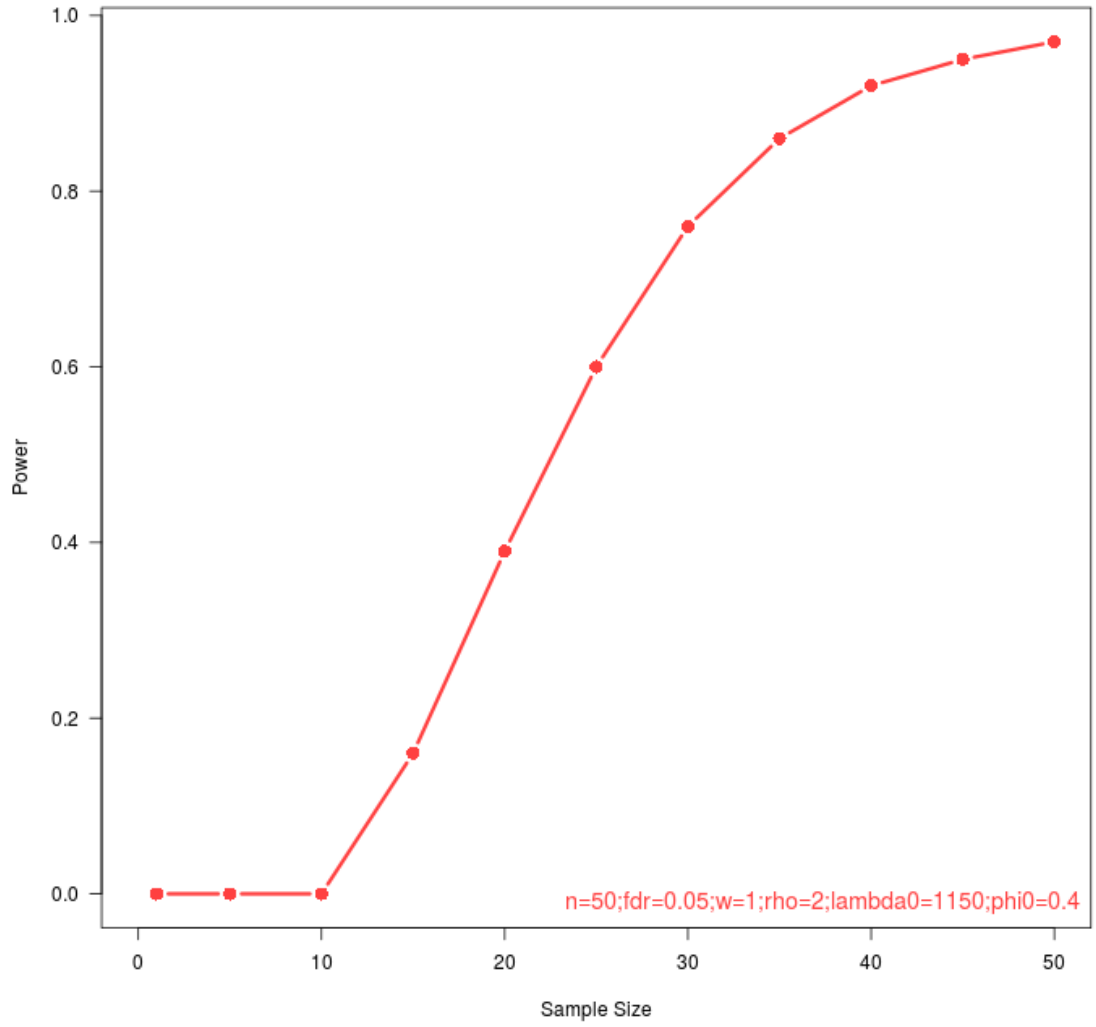


Figure S4 Power curve of the post-hoc power analysis. The curve was constructed using RNAseqPS (Guo *et al.*, 2014) with the following parameters: sample size (n) 50; FDR step-up threshold (fdr) 0.05; total number of genes for testing (m) 170,642; expected number of prognostic genes (m1) 2000; minimum fold changes for prognostic genes between two groups (rho) 2; average read counts for prognostic genes (lambda0) 1150; dispersion for prognostic genes (phi0) 0.4; ratio of normalisation factors between two groups (w) 1. The values for fdr and rho represent the FDR step-up and minimum fold change thresholds that were used for the differential expression; the value of m represents the total number of features in the analysis (number of contigs in the reference); the value for m1 represents the number of differentially expressed features found in our analysis; the value for lambda0 is the (normalised) average of read counts for all differentially expressed features. Other values are used at default settings.