Exploring Candidate Genes for the 'S-locus': The Control of Heterostyly Across Wild and Cultivated Species in the Genus Linum

DESMOND, ELEANOR,LAUREN

Exploring Candidate Genes for the 'S-locus': The Control of Heterostyly Across Wild and Cultivated Species in the Genus *Linum*

Eleanor Lauren Desmond

A thesis presented for the degree of master of science (by research)

Supervised by Dr Adrian Brennan
Department of Biosciences
University of Durham
United Kingdom

March 2018

Exploring Candidate Genes for the 'S-locus': The Control of Heterostyly Across Wild and Cultivated Species in the Genus *Linum*

Eleanor Lauren Desmond

Submitted for the degree of Master of Science
March 2018

Abstract

Heterostyly in *Linum* is characterised by the presence of two floral morphs with their male and female reproductive organs located at different heights. It is an adaptation designed to encourage cross-pollination and minimise self-fertilisation. In *Primula*, heterostyly is controlled by two alleles present at a single locus, designated the *'S-locus'*. The identities of genes responsible for controlling heterostyly in *Linum* remain to be determined. Although *Primula* and *Linum* are evolutionarily distant, the recent elucidation of a 278 kb region conforming to the *S-locus* in *Primula vulgaris* has provided a guide for potential candidate genes.

A series of candidate genes, including GLO^T and CYP^T have been identified from other heterostylous species. RNA-Sequencing analysis has been employed to assess the expression of candidate genes in 44 floral tissue samples, across 10 *Linum* species. Sequence reads were mapped to a reference *Linum tenue* assembly and read data was used to perform morph-specific differential expression analysis between the two heterostylous morphs and between heterostyles and homostyles. Relative expression of candidate genes was validated by qPCR analysis on both long and short-styled morphs of *L. tenue* at various stages of floral growth. It was anticipated that *S-locus* genes found in *Primula* and others would have orthologues in *Linum* and that there would be expression of candidate genes in the 'thrum morph', but not the 'pin morph'. It was also expected that the candidate genes would show differential expression during different stages of floral development.

There was no clear evidence of differential expression of any tested candidate gene between the two floral morphs or between the three different developmental stages, neither was there an observed difference in candidate gene expression between heterostylous and homostylous individuals. The most differentially expressed genes were found to be related to stress-response functions. Differences were observed in the relative expression of *GLO*^{T} and *CYP*^{T} between the qPCR and RNA-Seq methods of differential expression analysis. It thus could not be substantiated that heterostyly behaves in the same way in *Linum* as in *Primula*.

Declaration

The work in this thesis is based on research carried out at the Department of Biosciences, University of Durham, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

Firstly, to my supervisor, Dr Adrian Brennan for his infectious enthusiasm, which inspired me to undertake this project and for his patient guidance, encouragement and advice. To Ali Foroozani for access to the *L. tenue* transcriptome and to his RNA-Sequencing libraries; not to mention his patient explanations of bioinformatics, no matter the time of day or night. To Dr Rocio Pérez-Barrales, James Gardiner and Bea Landoni at the University of Portsmouth for wild sample collection and for valuable data analysis insights. To the members of Lab 19, most especially Alice Rowland and Paige Panter, for answering my endless trivial questions with good humour and succinct advice. To 3$^{\text{rd}}$ Year Project Student Conor Hughes for his work conducting *L. viscosum* extraction, quantification and qPCR.

I would also like to acknowledge the support of my friends and to thank them for putting up with me during the past year. I extend my particular gratitude to Adam Leach for his encyclopaedic and seemingly fathomless knowledge of Linux, Python and computer science in general and without whom I would have struggled greatly with the steep learning curve of computational biology. To Amy Stockdale for her excellent photography skills, for listening to my rants and offering pertinent, biologically-sound advice and for making me many, many cups of tea. To Andrew Messenger for his guidance in thesis formatting. To Rob Battrum, James Marriott and James Bernays for their pep talks which were indispensable to my sanity. I am indebted to Rob in particular for his insightful and meticulous proof-reading, as well as to the many others who have cast an appraising eye over this work. To Beth Lines and Jack Towers for repeatedly providing me with food, shelter and excellent company. To all in the Engineering Department for welcoming me into my adopted home and especially to Liz and Andrea in “the hatch" for enabling my coffee addiction. To DUEM, thank you for everything.

Lastly, and perhaps most importantly, to Mum, Dad and Ciara for their unwavering support, insight, guidance, and willingness to proof-read, coupled with no small amount of sarcasm. There is no way that I can thank you enough for all you have done for me.

Contents

Abstract ii

Declaration iv

Acknowledgements v

List of Figures xiii

List of Tables xvi

List of Abbreviations xviii

1 Introduction 1

1.1 Heterostyly in *Primula* 5

1.2 Heterostyly in *Turnera* 9

1.3 Heterostyly in *Fagopyrum* 12

1.4 Heterostyly in *Linum* 13

1.5 Future Perspectives 16

1.6 Aims and Hypotheses 20

1.6.1 Hypotheses 20

1.6.2 Aims 21

2 RNA-Sequencing Analysis 23

2.1 Introduction 23

2.2 Materials and Methods 26

2.2.1 Sequencing 28

2.2.2 Quality Control and Pre-processing 34

2.2.3 Mapping 42

2.2.4 Post-Processing 54

2.2.5 Differential Expression Analysis 58

2.3 Results 60

2.3.1 Raw Count Data 60
2.3.2 DESeq2 Analysis 70
2.3.3 Stress Response Roles for the most Significantly Differentially Expressed Genes 77
2.4 Discussion 85
2.4.1 *GLO*^{T} 93
2.4.2 Future Work 94

3 Experimental Analysis of Differential Expression 97
3.1 Introduction 97
3.2 Methods 98
3.2.1 Primer Design 98
3.2.2 Post-analysis homology searches at the amino acid level 105
3.2.3 Post-analysis determination of primer binding efficiency for loci in different species 109
3.2.4 RNA extraction and cDNA synthesis 113
3.2.5 qPCR 116
3.3 Results 118
3.3.1 Transcripts from Candidate Primers in *L. tenue* 118
3.3.2 Transcripts from Candidate Primers in *L. narbonense* 120
3.3.3 *CYP*^{T} and *GLO*^{T} expression in *L. viscosum* 122
3.3.4 *CYP*^{T} and *GLO*^{T} expression in homostylous species 123
3.4 Discussion 126
3.4.1 Future Work 133

4 Conclusions and Future Work 137
4.1 Future Work 142

References 145

Appendix 160

A DESeq2 Input 160

B Supplementary Code 168

C T-Coffee Example Output 171

D Post-Analysis Homology Searches 175

E RNA Quantification Tables **184**

E.1 Heterostylous Species 184

E.2 Homostylous Species 187

List of Figures

1.1 The reciprocal arrangement of sexual organs in *L. tenue* floral morphs. Distyly is a mechanism designed to minimise self-fertilisation and consequently there is very little inter-morph transfer. The yellow arrows indicate transfer of pollen between morphs. 3

1.2 The reciprocal arrangement of sexual organs in the long-styled (left) and short-styled (right) morphs of *P. vulgaris* (Huu et al., 2016). 6

1.3 The expression and genomic organisation of *S-locus* genes in *P. vulgaris* (Li et al. 2016). Reproduced with permission from Nature Plants. 8

1.4 Eight of the ten wild species (and one cultivated species, *L. usitatissimum*) sequenced for this project. **a)** Heterostylous *L. narbonense* (adapted from: Salguero Quiles (2004)), **b)** heterostylous *L. suffruticosum*, (adapted from: Le Petit Herboriste (1998)), **c)** heterostylous *L. campanulatum* L. (adapted from: Mrugala (n.d.)), **d)** homostylous *L. bienne* (adapted from: Viatour (2006)), **e)** homostylous *L. tenuifolium* (adapted from: Gotler (2009)), **f)** homostylous *L. strictum* L. (adapted from: Badia (n.d.)), **g)** homostylous *L. usitatissimum* (adapted from: Fenwick (2005)), **h)** homostylous *L. catharticum* L. (adapted from: Devlin (2006)) and **i)** homostylous *L. setaceum* Brot. (adapted from: Ramalho (2012)). *Linum* broadly exists in yellow and blue flowered clades. 17

1.5 The most recent published *Linum* Phylogeny, constructed using the rbcL chloroplast gene. Adapted from Ruiz Martín (2018) 18

2.1 Schematic of the Illumina 2500 NGS Workflow. Adapted from SeqAnswers (2007) 25

2.2 Flow chart depicting the RNA-Seq pipeline used to determine differential expression of candidate genes. Paired Illumina reads were mapped to a reference *L. tenue* assembly using STAR. HTSeq-counts was used to determine raw read count expression data from the aligned RNA-Seq reads. Differential expression analysis was conducted using DESeq2. 27

2.3 Species selected for RNA-Seq analysis, highlighted in yellow on the most recent published *Linum* Phylogeny (adapted from Ruiz Martín et al. (2018)), with the exception of *L. setaceum*, which is not depicted as part of this phylogeny. 28

2.4 Map of the wild species collection sites in Spain. Graphic produced using QGIS 2.6.13 software (QGIS Development Team, 2009) 29

2.5 The four developmental stages sampled for bioinformatic analysis. From left, GS1: immature bud, GS2: maturing bud GS3: flower opening, but has not yet completed development, GS4: open flower. Species depicted is *L. tenue*. Photography by Amy Stockdale. 33

2.6 Excerpt from an example .fastq file: *L. viscosum* x T GS1 2U.fastq file. Unpaired Trimmomatic data was used for this visualisation as the smaller file required less space once uncompressed. 34

2.7 FASTQ value characters in increasing order of quality. 34

2.8 Minimal Example of Trimmomatic input code. 35

2.9 FastQC data both pre- and post- trim for *L. bienne* dor5 at GS1. Only one sequence is included here, by way of example. Quality results was assessed by using the MultiQC amalgamation **Figures a) and b)**: Box plots of the base quality scores at read positions across all reads in a sample for pre-trimmed and post-trimmed samples respectively. Quality scores are depicted on the y-axis. **Figures c) and d)** summarise the GC content across the whole length of each sequence in a file and compare it to a modelled GC normal distribution. **Figures e) and f)** depict the adapter content and **Figures g) and h)** show the k-mer content. This is an analysis of overrepresented sequences that will spot an increase in exactly duplicated sequences. 37

2.10 The per base sequence quality graph for an example poor quality RNA-Seq run. Boxplots display the mean and interquartile ranges of the Phred quality score across reads. Image adapted from: Babraham Bioinformatics (2017) 39

2.11 Post-Trim data for all sequenced individuals; graphs produced using MultiQC. **Panel a)** shows the cumulative percentage count of the proportion of the library which has seen the adapter sequences at each position. **Panel b)** shows the percentage of overrepresented sequences found in each library. **Panel c)** indicates the percentage of base calls at each position for which an 'N' was called; if a sequencer is unable to make a base call with sufficient confidence, it will normally substitute an N. **Panel d)** shows the mean quality scores across each base position in the read. **Panel e)** is a representation of the per sequence GC content for all sequenced individuals and **Panel f)** shows the number of reads which have average quality scores. 40

2.12 Summary statistics produced using MultiQC (Ewels et al., 2016) for post-trim results. A heat map is used to display the highest percentage duplications with the most highly duplicated sequences depicted in red. The left hand side of the table shows the November 2016 sequencing run and the right hand side of the table shows the August 2017 sequencing run. . 41

2.13 Post-trim FastQC data depicting the per base sequence content after the removal of the first 9 bases. 42

2.14 Illumina sequencing produces millions of short reads. These can be aligned to a reference transcriptome, in this case *L. tenue*. Reads can be used to determine the degree of expression of a particular RNA transcript. Here, Sample 1 is more highly expressed than Sample 2. 45

2.15 A script to merge the trimmed output of the November 2016 and June 2017 sequence files, ready for input into STAR. 46

2.16 Minimal example of a script to run various functions within the SAMTools suite of packages 55

2.17 Flagstat output for the *L. campanulatum* thrum morph sequence at GS1. . 56

2.18 Minimal example of a script to run HTSeq-count on .bam files, the output of SAMTools 57

2.19 Raw read counts of tested candidate genes in a series of different heterostylous individuals at tested growth stages. 63

2.20 Multiple sequence alignment, performed using Clustal Omega (Sievers et al., 2011), of *S-ELF3*, *PveGLO2*, *GLO*^{T} and Contig 107032, from the *L. tenue* transcriptome scaffold. Asterisks indicate conserved bases. 66

2.21 Top BLAST hits for Contig 107032. 70

2.22 Putative conserved domains for Contig 107032, showing the presence of a potential MADS-domain (Retrieved from NCBI Blast). 70

2.23 Two separate DESeq2 analyses of heterostylous individuals. The DESeq2 results were not exactly reproducible between runs on the same data. In this case significance was tested at p <0.1, although results are reported at p <0.05. 75

2.24 Volcano plots of heterostylous species at **a)** GS1, **b)** GS2 and **c)** GS3 + 4 depicting Log2 fold change on the x-axis against the -log10 p value on the y axis. Genes significant at p <0.01 are depicted as red circles. Transcripts matching the candidate genes are depicted as pale blue circles. Blue lines show the ±2 log fold change. 76

2.25 Contig 119239 shows significant homology to members of the HSP90 superfamily. 77

2.26 MA plots showing differential gene expression between long-styled and short-styled morphs at **a)** GS1 **b)** GS2 and **c)** GS3 + GS4. Statistically significant differences (with an adjusted p-value of <0.05) in expression are marked by a red dot. Triangles point to points falling outside of the plot area. 79

2.27 PCA plots show the variance in gene expression at GS1 across the different species in the two floral morphs. Points cluster by species, rather than by floral morph type. 81

2.28 PCA plots show the variance between gene expression across the different species in the two floral morphs at all four growth stages. 83

2.29 Heatmap of the count matrix, showing the clustering of the different heterostylous species across all growth stages for the most differentially expressed genes. Red shades indicate more closely related sequences Darker shades of blue indicate more closely related sequences. 84

2.30 PCA plot showing the variance in gene expression between homostylous and heterostylous species between the three morph types 86

2.31 Schematic of the ABC model of floral development. Figure adapted from (Irish, 2017). 90

3.1 Minimal example of a tblastx search querying *S-ELF3* against the *L.tenue* transcriptome. 101

3.2 An example alignment of *CYP734A50* produced using GeneDoc Software (Nicholas & Nicholas, 1997). 102

3.3 Phylogenetic tree showing a line up of the closest blastx hits for *GLO*^{T} against the *L. tenue* transcriptome. The tree was created using MEGA 7 and visualised using Figtree v. 1.4.4 (Rambaut, 2009) 106

3.4 Phylogenetic tree showing a line up of the closest blastx hits for *CYP*^{T} against the *L. tenue* transcriptome. The tree was created using MEGA 7 and visualised using Figtree v. 1.4.4 (Rambaut, 2009) 108

3.5 A script written to prepare primers for BLAST searching against *L.tenue* and *L.usitatissimum* 110

3.6 Species selected for RNA-Seq analysis, highlighted in green on the most recent published *Linum* Phylogeny (adapted from Ruiz Martín (2018)) (with the exception of *L. setaceum*, which is not depicted as part of this phylogeny.) 113

3.7 The three growth stages sampled for qPCR analysis. GS1: closed bud. GS2, maturing bud and GS4 open flower. Species depicted is *L. tenue*. Photography by Ali Foroozani. 116

3.8 Mean relative expression of **a)** *GLO*^{T}, **b)** *CYP*^{T}, **c)** *PUM*^{T} and **d)** *S-ELF3*. *GAPDH* is used as a reference gene. Error bars show standard error of the mean. 118

3.9 Bar charts show mean relative gene expression of **a)** *CYP*^{T} **b)** *GLO*^{T} and **c)** *S-ELF3* in *L. narbonense* at all 3 growth stages. Graphs were produced using R Statistics. Error bars show standard error of the mean. 120

3.10 Bar charts show mean relative gene expression of **a)** *CYP*^{T} **b)** *GLO*^{T} at all 3 growth stages. Graphs were produced using R Statistics. Error bars show standard error of the mean 122

3.11 **a)** *CYP*^{T} relative gene expression in four homostylous species. **b)** *GLO*^{T} relative gene expression in four homostylous species. 123

3.12 Comparison of the relative gene expression of **a)** ***CYP***^{T} and **b)** ***GLO***^{T} in all heterostylous and homostylous species at all three growth stages. . . . 124

3.13 Comparison of the relative gene expression of **a)** ***CYP***^{T} and **b)** ***GLO***^{T} in all heterostylous and homostylous species. 125

C.1 An alternative multiple alignment of LgMYB21 using T-COFFEE (Notredame, Higgins, & Heringa, 2000). The output is colour-coded based on the quality of alignment. Closely aligned bases are coloured in red, whilst weakly aligned bases are coloured in green. 174

List of Tables

2.1 Alternative publicly available mapping tools for each stage of the pipeline. 26

2.2 The species, developmental stage and population of all of the wild flowers sequenced for analysis of heterostyly. The developmental stages are described in Figure 2.5 and Table 2.3 30

2.3 Qualitative Description of Floral Morph Growth Stages. 33

2.4 Initial sequencing depth and the number of reads dropped during the Trimmomatic trimming process. 43

2.5 Results of the STAR Mapping process 48

2.6 Percentage of rRNA contaminant removed using BBDuk from each sequence, prior to mapping 51

2.7 BUSCO gene coverage across a series of sequences from the November 2016 sequence run, assembled using Trinity. 54

2.8 Example data comparing the raw read counts of coordinate-sorted and name-sorted .bam files when analysed using HTSeq-count. Data is from *L. campanulatum* at GS2, based on the November 2016 Sequence Run. This data was not ultimately used in the analysis. 58

2.9 Input Comparisons for DESeq2 58

2.10 The heterostylous comparison groups used in the DESeq2 analysis. Not all growth stages for all species were put forward for comparison. 59

2.11 The homostylous comparison groups used in the DESeq2 analysis. Not all growth stages for all species were put forward for comparison. 59

2.12 Raw count data for candidate genes at GS1 67

2.13 Raw count data for candidate genes at GS2 68

2.14 Raw count data for candidate genes at GS3 and GS4. 69

2.15 Raw count data for candidate genes in homostylous species 71

2.16 The top six most differentially expressed genes based on DESeq2 analysis of GS1 heterostylous species. The Base Mean is the mean of normalised counts of all samples, normalising for sequencing depth. lfcSE is the Log Fold Change Standard Error 73

2.17 The top six most differentially expressed genes based on DESeq2 analysis of GS2 heterostylous species. 73

2.18 The top six most differentially expressed genes based on DESeq2 analysis of GS3 and GS4 heterostylous species. 74

3.1 Alternative *L. usitatissimum* transcriptomes tested. The top sequence is GigaDB and the bottom is LUSFL1AD 100

3.2 Candidate genes determined from a literature search, for which qPCR primers were designed. 104

3.3 Top amino acid blast hits for PUM^T.The numbers in brackets refer to the reading frame. 107

3.4 Top amino acid BLAST hits for the GLO^T primer, when blasted against a database of close search results. The low e values indicate the low quality of the matches. 112

3.5 cDNA Synthesis Program 115

3.6 Qualitative Description of Floral Morph Growth Stages. GS3 and GS4 were usually combined for qPCR experiments. Table repeated from Section 2.2 for ease of reference. 115

3.7 Master Mix for cDNA synthesis. To create a total volume of 10 μl in each well. The SYBR mix additionally includes Taq and dNTPs. 116

3.8 qPCR cycling conditions. Data collection was completed at Stage 3, step 2. 117

3.9 The primer dissociation stage implemented after the qPCR cycling. 117

D.1 Top amino acid blast hits for GLO^T. The numbers after the contigs refer to the reading frame. 176

E.1 *L. tenue* thrum quantification using Nanodrop and Qubit. Values given to 3 significant figures. 184

E.2 *L. tenue* pin quantification using Nanodrop and Qubit.Values given to 3 significant figures. 185

E.3 *L. narbonense* thrum quantification using Nanodrop and Qubit. Values given to 3 significant figures. 185

E.4 *L. narbonense* pin quantification using Nanodrop and Qubit. Values given to 3 significant figures. 186

E.5 *L. viscosum* thrum quantification using Nanodrop. RNA extraction and cDNA synthesis performed by Conor Hughes (3rd year project student). Values given to 3 significant figures. 186

E.6 *L. viscosum* pin quantification using Nanodrop and Qubit. RNA extraction and cDNA synthesis performed by Conor Hughes (3rd year project student). Values given to 3 significant figures. 187

E.7 *L.setaceum* quantification using Nanodrop and Qubit. Values given to 3 significant figures. 187

E.8 *L.catharticum* quantification using Nanodrop and Qubit. Values given to 3 significant figures. 188

E.9 *L.tenuifolium* quantification using Nanodrop and Qubit. Values given to 3 significant figures. 188

E.10 *L.maritimum* quantification using Nanodrop and Qubit. Values given to 3 significant figures. 189

List of Abbreviations

CYP^T	Thrum-specific Cytochrome P450
GLO^T	Thrum-specific *GLOBOSA*
AFLP	Amplified Fragment Length Polymorphism
BAC	Bacterial Artificial Chromosome
BAM	Binary Alignment/Map format
bp	Base Pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
CCM$^\mathrm{T}$	Thrum-specific Conserved Cysteine Motif
cM	Centimorgan
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GABA	Glutamate-derived γ-aminobutyric acid
GC	Guanine - Cytosine complementary base pairs
GFF	General Feature Format
GO	Gene Ontology
GS1	Growth Stage 1
GS2	Growth Stage 2
GS3	Growth Stage 3
GTF	Gene Transfer Format
KFB$^\mathrm{T}$	KFB$^\mathrm{T}$
LFC	Log2FoldChange

lfcSE Log Fold Change Standard Error. Stat (the Wald statistic) is the Log Fold Change divided by its standard error and is used to calculate p-values.

mRNA Messenger RNA

NCBI National Centre For Biotechnology Information

NGS Next Generation Sequencing

PCA Principal Component Analysis

PUM^{T} Thrum-specific Pumilio-like protein

RAPD Random Amplified Polymorphic DNA

RNA-Seq RNA-Sequencing

RPKM Reads Per Kilobase of transcript per Million mapped reads

rRNA Ribosomal RNA

S-ELF3 S-LOCUS EARLY FLOWERING 3

SAM Sequence Alignment/Map format

SSH Suppression Subtractive Hybridisation

STAR Spliced Transcripts Alignment to a Reference

TPM Transcripts Per Kilobase Million

TSS1 THRUM STYLE-SPECIFIC GENE

UEA The University of East Anglia

Chapter 1

Introduction

The generation and maintenance of genetic diversity is of paramount importance to the survival of a plant species, it improves general resilience and adaptability to changing environmental conditions (e.g. Lin, 2011) and can allow plants to mitigate any negative effects imposed by the sessile mode of living. The majority of angiosperms are hermaphroditic, with each flower containing both male and female sexual organs (Renner, 2014) and consequently plants have developed methods to minimise self- fertilisation, a process that depletes genetic diversity. Selfing is often associated with inbreeding depression, which lowers offspring fitness (Charlesworth & Charlesworth, 1987). Amongst others, outcrossing methods include wind-pollination and pollination by insects and other animals. Self-incompatibility comprises a series of genetic systems designed to reject “self" pollen and to receive “non-self" pollen, thus promoting outcrossing (Franklin-Tong, 2008). Another, sometimes secondary, method of promoting outcrossing is heteromorphy: the development of morophological differences between individuals (Smith, 2015). Heterostyly is one such system (reviewed in Ganders, 1979; Barrett, 1992; Barrett & Shore, 2008) and is found in at least 28 angiosperm families (Barrett & Shore, 2008). Heterostylous individuals exhibit one of several different floral morphs where the sexual organs are at different heights. Distylous species exhibit two different floral morphs: those with short styles and long stamens (also termed S-morph, thrum flowers) and those with long styles and short stamens (also termed L-morph, pin flowers); tristylous species are less common and have three floral morphs. Tristyles operate similarly although their transfer dynamics are more complex. Heterostyly is characterised by the reciprocal arrangement of male and female sexual organs within flowers. In distylous plants, the reciprocal arrangement dictates that the length of the pistil on one morph corresponds to the length of the anther on the other morph; all flowers on a given plant share the same morph (Ushijima et al., 2015). Pollen is more successful in producing seed when it meets stigma of the opposite morph than on a morph of its own type (Mather, 1950). The pollen's mating type is sporophytically controlled; it is determined by the genotype of the parent plant.

L-morph flowers have long styles, with stigmas at a high position in the flowers. Concurrently, the flowers have short stamens, with anthers at a low position. In S-morph flowers, the styles are short, with stigmas in a low position and the stamens are long, placing the anthers in a high position. The pollination target of pollen from the L-morph is the S-morph stigma and vice versa (Figure 1.1). The reciprocal arrangement of the flowers means that it is spatially difficult for an insect pollinator to transfer pollen within the same individual and floral morph. It is thought that the reciprocal locations of the sexual organs correspond to the areas on the pollinators' bodies where the pollen is deposited and retrieved (Barrett, 2002).

Heterostyly is a relatively rare phenotypic polymorphism, present in just 4% of angiosperm species (Barrett & Shore, 2008); consequently it is highly likely that its persistence in angiosperm families is significant. The adaptive significance of reciprocal stigma-anther height polymorphisms in distylous species is thought to result from enhancements to pollen transfer efficiency between long and short morphs (Ganders, 1979; Wolfe, 2001). The reciprocal positioning of sexual organs within the flower reduces the conflicts inherently associated with both sexual organs occurring within the same reproductive unit (Keller et al., 2014). Darwin (1862) proposed that these features of heteromorphic species evolved in order to promote outcrossing by insect pollinators, which is crucial to the prevention of inbreeding depression in heterostylous species. Heterostyly promotes disassortative pollination which is a key functional benefit of heterostyly as it promotes more efficient transfer between reciprocal sexual organs, located on opposing floral morphs.

Heterostyly has evolved independently in numerous families and consequently there is considerable diversity of floral morphology across the heterostylous taxa (Ganders, 1979). Furthermore, the recurrence of similar polymorphisms across the various heterostylous species is indicative of their functional significance in the operation of the breeding systems and provided the basis for many of the earliest genetic studies of the system (e.g. Bateson & Gregory, 1905). Heterostylous species are a remarkable exemplar of convergent evolution across floral morphology, genetics and physiology (Ganders, 1979).

The textbook model of the genetics of heterostyly is derived from early classical genetic studies in *Primula* conducted by Bateson and Gregory (1905). These analyses concluded that heterostyly was controlled by two alleles present at a single locus, S

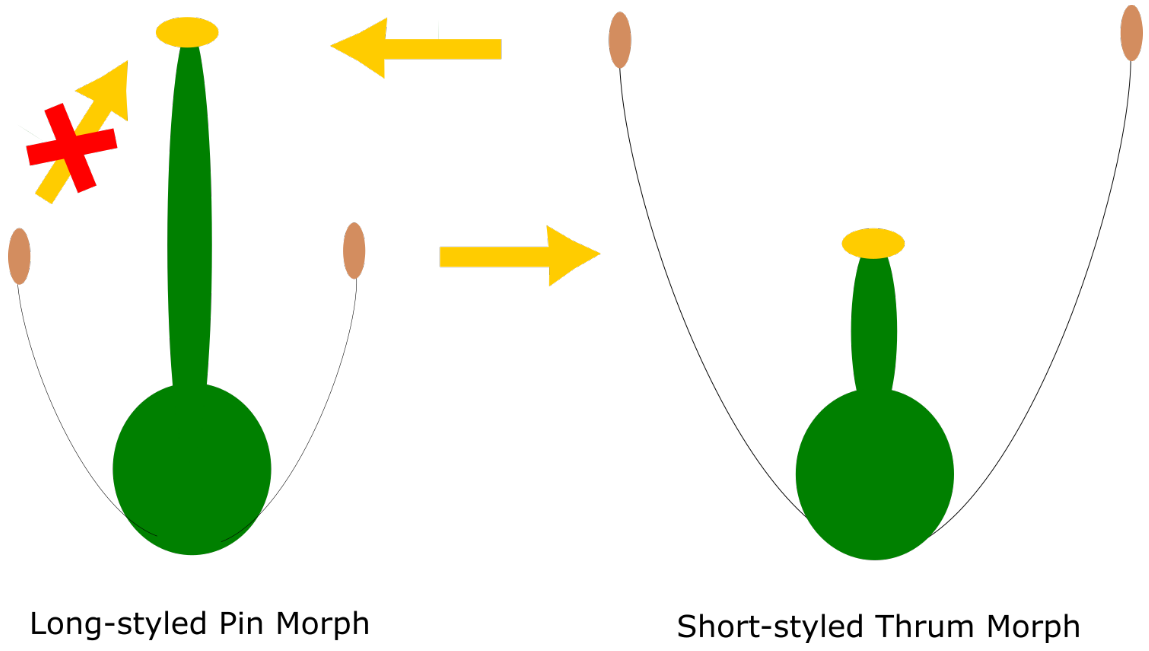


Figure 1.1: The reciprocal arrangement of sexual organs in *L. tenue* floral morphs. Distyly is a mechanism designed to minimise self-fertilisation and consequently there is very little intermorph transfer. The yellow arrows indicate transfer of pollen between morphs.

(short style) and s (long style) (Bateson & Gregory, 1905; Gregory et al., 1923) and led to the interpretation that short-styled flowers (thrum) are heterozygous (S/s) and the long-styled (pin) flowers are homozygous recessive (s/s); homozygous dominant flowers are lethal (Bateson & Gregory, 1905; Kurian & Richards, 1997). However, recent genomic analyses have shown that the locus consists of a single hemizygous allele; disproving the previous conclusions.

The locus responsible for controlling the generation of different floral morphs is a supergene known as the *S-locus* (Lewis, 1954; Barrett, 1992), where the *′S′* denotes Style length. Supergenes are clusters of closely associated genes, always inherited together by progeny (Thompson & Jiggins, 2014). As the *S-locus* controls both the self-incompatibility phenotype and multiple aspects of floral morph physiology, it is anticipated either that the same gene or genes determine both self-incompatibility and morph physiology (Ushijima et al., 2015) or alternatively that both phenotypes are governed by different genes, linked completely at the *S-locus*.

Early work by Ernst established the supergene model; he demonstrated that self-compatible long homostyles were inherited as if determined by additional *S-locus* alleles and initially proposed that two, and later showed that at least three (Ernst, 1955) tightly linked loci are responsible for distyly. These linked sub-loci control style length (*G*), stamen length (*A*) and pollen size (*P*). This work also recognised six subcharacters of heterostyly: position of anthers, position of stigma, size of stigmatic papillae, size of pollen grains, incompatibility reaction of pollen and incompatibility reaction of style which are normally inherited as a single unit. Ernst (1936b) obtained plants in which the female and male characters had been re-associated to give self-compatible homostyly and additionally recorded plants where pollen showed re-association of size and compatibility, i.e. further abnormal plants showed distinctions between pollen grain size and anther height. The isolation of such abnormal plants led to the genetic distinction of three groups, inherited together: style and stigma, and female incompatibility; anther height; pollen grain size and incompatibility. Dowrick (1956) later proposed an extension to this model for *Primula*, suggesting the addition of an allele controlling fertility differences. Dowrick (1956) revised the order of the loci in his survey of homostyly in *Primula*, from Ernst's *GAP* noting that if crossing-over occurred, rather than mutation, as was originally suggested by Ernst, the most likely order of genes would be GPA (i.e. gynoecium, pollen size, anther height). This prompted Ernst to re-examine his initial data . Dowrick's studies of homostylous plants led to the conclusion that the *S-locus* comprises a co-adapted linkage group; recombination within the locus can lead to either long or short homostylous flowers (Dowrick, 1956). This genetic overview has been challenged by Li et al. (2016)'s recent publication and follow-up papers and has now been proven incorrect. Evidence for the new theory for the genetic control of heterostyly will be discussed below, in Section 1.1. In light of the new knowledge proposed by Li et al. (2016) and colleagues, old results can now be re-interpreted.

In the decades since, molecular and transcriptomic work has been conducted in an attempt to characterise gene sequences that form part of the *S-locus* and its surrounding region across a variety of heterostylous plant species. Characterising these sequences should allow identification of the genes responsible for the differential growth of male and female reproductive organs in heterostylous plants. Genes coding for certain biological functions, such as auxin transport and cell-elongation are the best candidates for genes present in and linked to the *S-locus* as they are likely to regulate reproductive organ growth (Labonne et al., 2009). It is now pertinent to examine some of the recent developments in the genera *Primula*, *Turnera*, *Fagopyrum* and *Linum*. Prior to 2016, the genes at the *S-locus* had not been definitively identified in any distylous species (Barrett & Shore, 2008; Labonne & Shore, 2011; Li et al., 2016), however, vast improvements in available molecular techniques over the past decade have enabled high-resolution mapping of the *S-locus* in *Primula*. The recent elucidation of the *S-locus* gene CYP^T (CYP374A50) in Li et al. (2016) predicts an exciting future for the determination of candidates for heterostyly in other species.

As reviewed in (Kappel et al., 2017), the evolution of heterostyly has been explained by two main competing models (Charlesworth, 1992; Lloyd & Webb, 1992a, 1992b). The models differ in their interpretation of the ancestral state of heterostyly and in their inference of the sequence of trait acquisition. Charlesworth (1992)'s model assumes that the common ancestor was homostylous and self-compatible and that a mutation to a novel incompatible pollen type spreads and establishes a polymorphism, if the product of selfing rate and inbreeding depression is above 0.5 (Charlesworth & Charlesworth, 1979). In contrast, Lloyd and Webb (1992a) based their model on the taxonomic observation that heterostyly has only evolved in families with relatively simple 'depth-probed' flowers. Such flowers often show approach herkogamy which led Lloyd and Webb (1992a) to the conclusion that heterostyly evolved from approach-herkogamous ancestors, with at least partial outbreeding. This model requires the invasion of a population by a dominant mutation, shortening the style and then a subsequent mutation which elevates the anthers in the short-styled form to the level of the stigma in the original form; reciprocal herkogamy would evolve before self-incompatibility. The current literature is in favour of the Lloyd and Webb (1992a) model, as the levels of inbreeding depression required by Charlesworth and Charlesworth (1979) are unlikely to be achieved: the majority of non-herkogamous species and their derived homostyles are highly-selfing, which tends to purge the genetic load causing the inbreeding depression. After purging, any further inbreeding is anticipated to be less harmful.

1.1 Heterostyly in *Primula*

Sustained work on the characterisation of the genetic basis of heterostyly has been conducted in members of the genus *Primula* and consequently, members of the *Primulaceae* have contributed a large proportion of the current knowledge base. Classic early 20th century genetic studies by Bateson and Gregory (1905); Bridges (1914); Ernst (1936a, 1936b) and others led to the determination of a linkage group of three genes at the *S-locus*, each controlling a separate aspect of heteromorphic flower development (Mather, 1950). Ernst (1936a) discovered allelomorphs which give rise to anomalous combinations of the sub-characters. Plants in which female and male characters have been re-associated to give self-compatible homostyly were produced, as were plants where pollen shows a re-association of size and compatibility. Since then, genetic studies conducted in *Primula* have demonstrated that the *S-locus* comprises at least 4 genes: S, G, P and A, which control the self-incompatibility phenotype, style length, pollen size and anther height respectively (Barrett, 2002). As of December 2016 (Li et al., 2016) there are

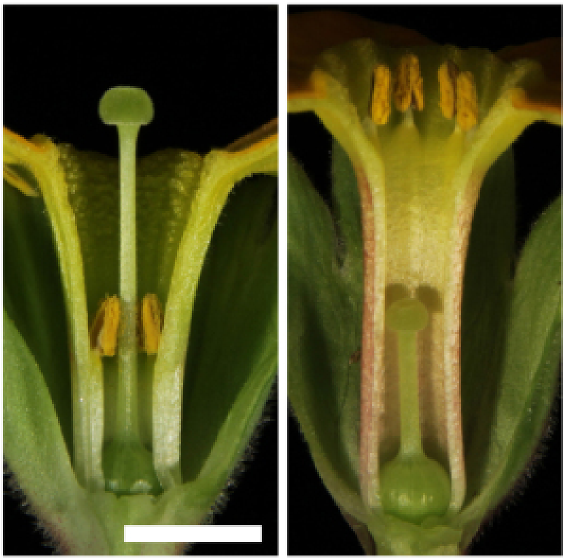


Figure 1.2: The reciprocal arrangement of sexual organs in the long-styled (left) and short-styled (right) morphs of *P. vulgaris* (Huu et al., 2016).

5 confirmed genes within the *S-locus* of *Primula*.

Early molecular attempts to determine the genes responsible for heterostyly involved a comparison of protein profiles of the reproductive organs within different morphs discussed in (Dulberger (1974)). Following these early investigations, studies of *S-locus* mutants such as *Oakleaf* (Cocker et al., 2015) and *Hose in Hose* (Li et al., 2010) assisted in the definition of phenotypic markers for the *S-locus* (Li et al., 2015). The dominant *Hose in Hose* mutant causes the homeotic conversion of sepals to petals in *P. vulgaris* owing to ectopic expression of *PvGLO1*, the orthologue of *GLOBOSA* in *P. vulgaris* (Li et al., 2010). *Oakleaf*, another *S-locus*-linked dominant mutation makes the shape of *P. vulgaris* leaves more deeply lobed than wild-type (Cocker et al., 2015). A high-resolution linkage map was constructed based on these markers placing *GLO1* and *Oakleaf* on opposing sides of the *S-locus* (Li et al., 2015). Rapid advances in our understanding of the genes responsible for heterostyly arose from the publication of the draft genome of distylous *P. veris* L. (Nowak et al., 2015). This represented the first genome assembly of a heterostylous species and was made possible by advances in high-throughput sequencing technologies. The annotated transcriptome produced in this work was combined with restriction-associated (RAD) tag genotyping of large S- and L-morph pools. RNA-Sequencing RNA-Seq) data from *P. veris* and the closely related species *P. vulgaris* revealed 113 genes showing varying levels differential expression between the two floral morphs, making them candidates for the genes controlling heterostyly. One gene in particular, the duplicated-*GLOBOSA* homologue *PveGLO2* is totally silenced in long-styled morph flowers (Nowak et al., 2015). The duplicated gene *PveGLO2* therefore represented an interesting candidate for further study; this thesis hypothesises that there is a conserved basis for heterostyly in *Linum*. However, the short-morph specific duplication of this gene may be unique to *Primula* and therefore is not necessarily transferrable to heterostylous species in other genera.

Transcriptomic work by Huu et al. (2016) identified *CYP734A50* as the gene responsible for controlling style length in several *Primula* species, and also illustrated its mode of operation. Using RNA-Seq to analyse differential gene expression, Huu et al. showed that *CYP734A50* is only present in short-morph plants and is exclusively expressed in their styles. Its gene product is an enzyme which degrades brassinosteroids, a class of plant hormone that promotes cell elongation. Brassinosteroids are being degraded in the short-morph thereby reducing cell elongation to result in a shorter style. Levels of the brassinosteroid castasterone are high in styles of the long-morph and absent in the short-morph (Huu et al., 2016). The reciprocal floral morphs are formed through differences in cell elongation, thus, the presence or absence of *CYP734A50* explains reciprocal organ size in *Primula*. The presence of brassinolide, the biologically active form of castasterone, explains style length, but not anther position (Huu et al., 2016).

Recent results show that the development of the genetically dominant short-morph in *Primula* is controlled by a cluster of five linked genes which are missing in the reciprocal morph (McClure, 2016; Li et al., 2016). This five-gene cluster spans a region of 278kb and contains the following thrum-specific genes: CCM^T (Conserved Cysteine Motif) , GLO^T (short-morph specific *GLOBOSA* gene) , CYP^T (Cytochrome P450) , PUM^T (Pumilio-like) and KFB^T (Kelch repeat F Box) (Li et al., 2016). For the sake of simplicity, the nomenclature from Li et al. (2016) will be used for the remainder of this thesis, nevertheless *CYP734A50* and *GLO2* from *P. veris* correspond to CYP^T and GLO^T from *P. vulgaris*. The regions immediately flanking the insertion contain stretches of DNA without elevated rates of polymorphism between the two morph haplotypes. Kappel et al. (2017) argue that this is suggestive that the recombination has occurred recently in the vicinity of the *S-locus*.

An explanation for reciprocal organ size can be provided by combining these results with the report of *CYP734A50* (Huu et al., 2016), an enzyme encoded by one of the five named genes breaking down a hormone responsible for cell elongation. Huu et al. (2016) and Li et al. (2016) both conclude that CYP^T (*CYP734A50* in Huu et al. (2016)) function is responsible for the control of style length in *Primula*. This correlates with the conclusion of Yasui et al. (2012) that a gene thought to control style-length is absent from long-styled forms in *Fagopyrum esculentum*. In *F. esculentum*, however, the candidate gene for style length is a transcription factor gene which has no obvious connection to floral organ size. This is reviewed in more detail in Section 1.3.

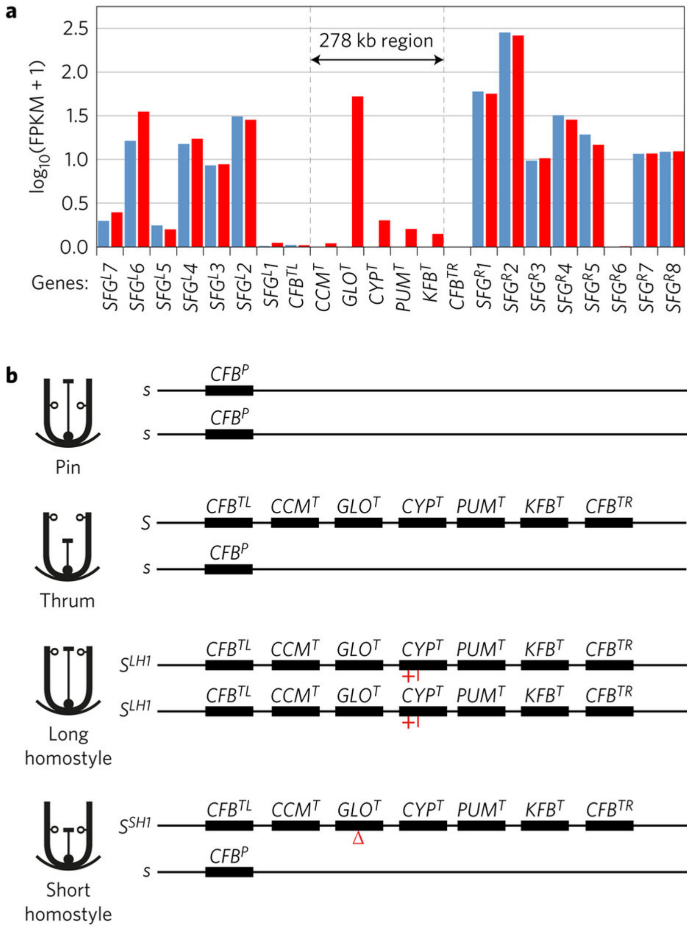


Figure 1.3: The expression and genomic organisation of *S-locus* genes in *P. vulgaris* (Li et al. 2016). Reproduced with permission from Nature Plants.

Li et al. (2016) searched for sequences specific to the thrum by identifying transcripts that were unique to thrum morphs by mapping to the pin genome. From these analyses they concluded no other floral-expressed genes are unique to the morph; the 278kb sequence in Figure 1.3 is the only thrum-specific genomic region. The transcription factor gene GLO^T was initially defined as a thrum-specific allele of *P. vulgaris GLO* in Li et al. (2010); *GLO* is responsible for the *S-locus* linked mutant phenotype *Hose in Hose*. Li et al. (2016) have shown that GLO^T is a distinct locus. Mutations in the CYP^T and GLO^T homostyle alleles suggest that they are candidates for controlling style length suppression (G gene) and low anther height (A gene) respectively (see Ernst (1936b) and Dowrick (1956)). Absence of CYP^T allows for the long-style phenotype and absence of GLO^T forms low anthers. Li et al. (2016) demonstrate very tight thrum-specific linkage between GLO^T and its surrounding region and the *S-locus* in both *P. veris* and *P. vulgaris*. The sequences on the left and right hand sides of the *S-locus* contain genes expressed in both of the morphs of distylous *Primula*, providing evidence of the locus boundary. *PUM*^{T} and *KFB*^{T} are both unique to the 278kb region and have no homologues in the pin genome (Li et al., 2016). Thus, further analysis may demonstrate a role for these genes in defining other short-morph characteristics. The complete absence of the 278kb region in long-styled morphs implies that the short-styled morphs would be better described as hemizygous rather than being caused by dominant alleles as had been the widely held view since Bateson and Gregory (1905). The lack of corresponding sequence at the long-morph s haplotype also explains the lack of recombination, which is required to keep *S-locus* genes together (McClure, 2016).

A further outcome of the recent Li et al. paper was the calculation of the original date of the origin of distyly in *Primula*. The current *S-locus* structure, responsible for heterostyly in the genus *Primula* arises from a duplication event that occurred 57.1 million years ago in a floral homeotic gene, responsible for the identity of flower petals, and its subsequent neofunctionalisation and insertion into the *S-locus*, acquiring the previously absent function of anther-position control. This calculation was performed by isolating *GLO* and *GLO*^{T} sequences from six different *Primula* species and using these along with others to undertake a Bayesian relaxed-clock phylogenetic analysis. (Li et al., 2016).

Ongoing work at The University of East Anglia (UEA) is using gain and loss of function transgenics in mutants of *GLO*^{T} and *CYP*^{T} to further define the role of these key genes in heterostyly and in self-incompatibility. Current analysis of long homostyle and short homostyle mutants has suggested that *CYP*^{T} and *GLO*^{T} are responsible for cross-regulating the other genes comprising the *S-locus*. The UEA group aim to use transgenics and transcriptomics to define this regulation, and subsequently reveal how di-morphic flower development is controlled by the *S-locus* (Gilmartin, 2017) . Since its single origin, heterostyly has broken down at least 30 times in *Primula* to give self compatible homostyles with high anthers and long styles. Studying how the system breaks down should provide valuable insight into its mode of operation.

1.2 Heterostyly in *Turnera*

Distyly is the most common breeding system within the Turneraceae (a family of tropical or sub-tropical shrubs consisting of 120 species in 10 genera) , with 7 of the 10 genera exhibiting the polymorphism (Shore et al., 2006). Many species of the largest, eponymous genus *Turnera*, including *T. subulata* and *T. krapovicasii* are heterostylous, and extensive work by the Shore Laboratory at York University in Ontario, Canada, over several decades has been significant in establishing *T. subulata* as a model for heteromorphy (e.g. Labonne et al., 2009; Labonne & Shore, 2011); discussed in Gilmartin and Li (2010). Owing to the short timespan within which plants may be crossed and brought to flowering the ease of doing so, *Turnera* spp provide a useful system for the identification of genes associated with distyly (Tamari & Shore, 2004). Species in the genus *Turnera* show typical distyly, with the main difference being the relatively open structure of the flowers, which deviates significantly from the more common, tubular shape seen in the majority of heterostylous species (Ganders, 1979).

Labonne et al. (2009) used a mutagenesis screen to identify a series of *S-locus* candidate genes in *T. subulata*; this was the first example of mutagenising a heterostylous species. Two diploid *T. subulata* plants were screened: a long-styled plant (S16L, homozygous recessive, ss) and a short-styled plant (F60SS, homozygous SS at the *S-locus*). Mutants were x-ray generated and of the 3982 progeny screened, 10 long-styled mutants, 1 long homostyle and 1 short homostyle were recovered (Labonne et al., 2010). The x-ray mutagenesis identified a series of deletion mutants. The mutant producing the recovered long-homostyle was discovered, following phenotyping and genotyping, to be within the *S-locus*. This finding implied that the equivalent gene to the *Primula G* locus had been deleted, although the equivalent to *A* was still functional (Labonne et al., 2010).

In a 2011 follow-up study, Labonne and Shore positionally cloned the s haplotype of the long-styled morph in *T. subulata* through the creation of a BAC contig which spanned the *S-locus* (Labonne & Shore, 2011). Through extensive studies of *T. subulata* using inheritance data and mutation of the *S-locus*, Labonne et al. have found heterostyly to be consistent with the *Primula* supergene model. They have positionally cloned the *S-locus* of *T. subulata* and are now in a position to discern the characters responsible for the heterostylous syndrome.

The Shore Laboratory had determined morph-specific proteins in *Turnera* species (Athanasiou & Shore, 1997). Using isozyme markers, it was possible to identify progeny that exhibited recombination adjacent to the *S-locus*. The presence of these proteins guided more specific analyses of the extent of the *S-locus* supergene. Sequencing of polypeptides in *T. subulata* led to the discovery of two genes localised to the style tissue of the short-styled morph. One such protein was annotated as a polygalacturonase (Athanasiou et al., 2003); the other an α-dioxygenase (Khosravi et al., 2004). Subsequent linkage analysis, however, revealed that although the polygalacturonase is located in a region adjacent to the *S-locus* and is not contained within it. Similarly, the α-dioxygenase also does not map to the *S-locus*.

In *Turnera*, the *S-locus* is proposed to localise near the centromere. Centromeric localisation of the *S-locus* was historically considered to be a secondary cause of recombination suppression *S-locus* in both homo- and heteromorphic systems. However, a similar observation in *Primula* (Li et al., 2015) and the recently discovered hemizygous mechanism in *Primula* (Li et al., 2016), with an absent *s* haplotype provides a justification for the apparent lack of recombination, which in the past had been mistaken for recombination suppression. The hemizygous region has nothing with which to recombine and thus, the perceived recombination suppression in *Turnera* like that in *Primula* further implicates a hemizygous system of inheritance.

Labonne et al. (2009) have produced a high-resolution genetic map of the *Turnera S-locus* and proposed three notable candidates for genes involved in the maintenance of heterostyly. The candidates include N-acetyltransferase-encoding *TkNACE*, which flanks the *S-locus* at a distance of 0.35 cM and *TkST1*, which codes for a sulphotransferase and is contained within the *S-locus* itself. *TkST1* in particular shows some differential expression between long and short-styled morphs, indicating possible involvement in distyly. This study has dramatically improved knowledge of the genetic basis of heteromorphy in *T. subulata* and represents the first step towards the identification of associated molecular mechanisms. Upon publication, it was the highest-resolution map of the *S-locus* region in a distylous species. Further investigation *in silico* could determine candidates with similar functions across a range of species to identify potential similarities in the mechanism of heterostyly across these different species.

Several studies have questioned the role of retrotransposition in heterostyly in *Turnera*. Labonne et al. (2009) identified a co-segregating retroelement, *TsRETRO*. However, the presence of at least one STOP codon in the reverse transcriptase domain prevents functionality and suggests that *TsRETRO* is a pseudogene. Nevertheless, the maintenance of *TsRETRO* within the genome suggests that it might not be devoid of functional significance and consequently, the potential for a *TsRETRO*-like gene in other heteromorphic species is worthy of investigation. *S-locus*-associated retroelements have also been found in *Primula* (Manfield et al., 2005), *Brassica* (Cui et al., 1999) and *Antirrhinum* (Lai et al., 2002) amongst others. The comparisons to *Brassica* and *Antirrhinum* must be interpreted with caution however, as the S in the *Primula* refers to style length and not to conventional self-incompatibility. Although similarities are present the *S-locus* in *Primula* cannot be regarded as another *SI* locus.

1.3 Heterostyly in *Fagopyrum*

Fagopyrum esculentum Moench., or common buckwheat, shows a typical form of morphological distyly and is self-incompatible (Darwin, 1877). Like *Turnera* the system conforms to the standard genetic pattern. The short morph is heterozygous (S/s), determined by the dominant S-haplotype and the long morph is homozygous (s/s) (Lewis & Jones, 1992; Kappel et al., 2017). The self-incompatibility system in *Fagopyrum* is expected to be novel based on phylogenetic independence between heteromorphic self-incompatibility and more traditional homomorphic sporophytic self-incompatibility and differences in the timing of pollen rejection between the two systems (Yasui et al., 2012). It has been shown that the flower homostyly of one of the accessions of *F. homotropicum* (a homostylous relative of *F. esculentum*) is determined by a single gene and behaves as a dominant trait for the pin morph of *F. esculentum*. It was also shown later by a recombination test that the homostylous gene of this accession was not an allele to the *S-locus* in *F. esculentum* (Fesenko et al., 2006).

Random amplified polymorphic DNA (RAPD) (Aii et al., 1999) and amplified fragment length polymorphism (AFLP) (Yasui et al., 2004) markers were identified surrounding the *S-locus*. Additionally, Miljuš-Đukić et al. (2004) detected several proteins that were singularly expressed in long or short-morphed individuals using Two-dimensional gel electrophoresis (2D Page). These proteins remained uncharacterised until the advent of Next Generation Sequencing (NGS) technologies.

NGS technologies have proved instrumental in the identification of transcripts exclusive to the S-morph. Yasui et al. (2012) determined four transcripts which were restricted to the short-styled morph and named them *SSG1-4*. This study successfully isolated *S-LOCUS EARLY FLOWERING 3* (*S-ELF3*) , a gene which controls the short-styled phenotype in buckwheat. Yasui et al. (2012) further suggested that there was strong suppression of recombination around *S-ELF3*. Two independent mutations which resulted in a long homostylous and self-compatible phenotype both carry inactivating mutations in *S-ELF3*. This provided substantial evidence that *S-ELF3* is one gene responsible for the development of heterostyly, and potentially also for female self-incompatibility in *Fagopyrum*.

A 610 kb region surrounding *S-ELF3* was defined using a BAC library. This region contained both *SSG2* and another gene, which is yet to be functionally characterised. Many repetitive sequences and transposable elements were also found within this region. The recent publication of the draft genome of *F. esculentum* by (Yasui et al., 2016) furthered the previous sequencing analysis. This search for other S-morph-specific contigs, in addition to *S-ELF3* determined at least 5.4 Mb of sequence which may be absent in the L-morph, but which was present in almost all S-morph individuals tested. Yasui et al. (2016) noted that the identified contigs were largely composed of transposable elements (almost 75%), and that only 32 predicted genes were found in this region. Thus, the region surrounding in the *S-locus* is indicative of a large, non-recombining hemizygous region (as in *Primula*, described above) that accumulates transposable elements.

1.4 Heterostyly in *Linum*

Although the majority of the more recent work has been conducted in other genera, most commonly *Primula* (Li et al., 2016; Huu et al., 2016) and *Turnera* (Labonne et al., 2009), *Linum* has been historically important for the study of heterostyly. Darwin's work in 1862 and 1877 characterised the relationship between fertilisation capacity and different types of floral morph in *Linum* and began a century's interest in the genetic basis of heterostyly. Distyly is common in *Linum*: it is exhibited by over 40% of *Linum* species (Rogers, 1979). Heterostyly in *Linum* is purported to be controlled by the *S-locus* supergene, in a similar manner to *Primula*; the evidence being that all of the self-compatible monomorphic populations that have been identified resemble the L-morph (Nicholls, 1985).

Linum is a diverse genus with a wide geographical distribution (Ruiz Martín et al., 2018); it consists of approximately 180 species. Amongst these species exists wide variation in mating systems, varying from self-pollination to heterostyly (McDill et al., 2009; Armbruster et al., 2006). Unusually amongst heterostylous angiosperms, there is some variation in the degree of differentiation in traits between floral morphs in heterostylous *Linum* (Wolfe, 2001). In contrast to other heterostylous plants, there is no large difference in anther-height between the two floral morphs in *L. grandiflorum* (Darwin, 1877). Some species exhibit full reciprocal herkogamy with heteromorphic pollen and stigmas. Not all species exhibit heterostyly: wild flax, *L. bienne* and its cultivated relative *L. usitatissimum*, are both self-compatible and homostylous, despite other closely-related members of the genus such as *L. grandiflorum* and *L. narbonense* showing heterostyly; the majority lie somewhere in between. The distylous form is strongly self-incompatible (Dulberger, 1992) and homostylous cultivated flax, *L. usitatissimum*, for example, 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 & Patel, 1938). The majority of the variation in species and mating systems is found in the Mediterranean area.

Heterostyly has been generally considered to be the ancestral condition in *Linum* (Rogers, 1982) and to have been “lost" in modern homostylous species. The two main pieces of evidence used to support this claim are that heterostyly is found in four of the five sections of the genus and that heterostyly is known to occur in other genera of Linaceae, notably *Reinwardtia* and several members of the Hugonioideae. However, heterostyly is only found in old world species (Les, 2017). The phylogenetic study by Armbruster (2006) concluded that heterostyly evolved several times independently within *Linum*'s different lineages. McDill et al. (2009) failed to determine with high levels of confidence as to whether heterostyly or homostyly is the ancestral state.

Flowers in *Linum* species have been referred to as “thrum" and “pin" based on differences in style length (Lewis, 1943; Ushijima et al., 2012). Determining heterostyly's mode of inheritance is crucial for the isolation of the *S-gene* in *Linum* since precise phenotyping is hindered by the absence of true reciprocity in *L. grandiflorum* (Barrett, 1992; Wolfe, 2001). Thus, Ushijima et al. (2015) sought to define the precise characteristics of each morph in *L. grandiflorum*. In distylous *L. suffruticosum*, Armbruster et al. (2006) demonstrated that anthers and stigmas show full reciprocity in three dimensions, a further unique example of heterostyly in a *Linum* species. The polymorphism is a 'twisted' form of distyly, where both styles and stamens bend during floral development to produce a three-dimensional arrangement. The long and short morphs also exhibit reciprocal herkogamy, as in conventional distylous individuals.

Work by Ushijima et al. (2015) recently demonstrated that heterostyly in *Linum* is controlled by a single diallelic locus, the S locus. It has yet to be determined whether the *S-locus* seen in *Primula* and *Fagopyrum* bears any structural similarity as the genes present at this locus remain unknown. McCubbin et al. (2006) compared cDNA between flowers of the two distylous *Primula* morphs. They successfully identified 11 classes of differentially expressed genes but despite some differences between pin and thrum, none were found to be linked to the supergene. Ushijima et al. (2012) isolated a candidate for the *S-gene* called the *THRUM STYLE-SPECIFIC GENE* (*TSS1*) from *L. grandiflorum*. *TSS1* is expressed in thrum (short-style) styles only; it was found in half of all tested pollen grains, which had prompted the hypothesis that the *S-locus* and thrum is heterozygous (Ushijima et al., 2015). However, in light of Li et al. (2016)'s proposal that the short-morph in *Primula* is hemizygous, the fact that this concept may also be true in *Linum* deserves further attention. Ushijima et al. (2015) have shown that *TSS1* segregates completely with the thrum phenotype; indicative of strong linkage. In the same paper, Ushijima et al. (2015) demonstrated that petal pigmentation co-segregates with heterostyly in *L. grandiflorum*: pink flowers were found in both morphs, however, red flowers were found only in short-styled morphs and white flowers only in the long-styled morph.

Further potential candidate genes determined by Ushijima et al. (2012) include *LgSKS1*, *LgAP1* (an aspartyl protease) and *LgMYB21* (a MYB transcription factor). These genes are expected to code for *S-locus* controlled downstream components rather than members of the supergene. *LgSKS1* displayed no polymorphism in SNP genotyping. However, given the effect of another *SKS* member on *in vivo* pollen tube growth, it is likely that *LgSKS1* is a downstream component of the *S* factor, which contributes to the pollen inhibition component of the self-incompatibility reaction, through regulating the growth of the pollen tube. The genes do remain noteworthy because the identification of the pathway controlling floral morph-specific post-transcriptional regulation is likely to be key to elucidating the molecular mechanism of heterostyly in *Linum*. The gene products of these two remaining candidates are very different and their mode of action in the pollen tube is thought to be oppositional: *LgAP1* has a toxic effect on pollen tube growth, whereas *LgSKS1* is expected to facilitate pollen tube growth. Thus, if both gene products are involved in heterostyly, they would be anticipated to have different mechanisms of pollen inhibition in the pin and the thrum (Ushijima et al., 2012). Opposing mechanisms are thought to be the case in *Turnera* (Tamari et al., 2001); owing to differences in pollen tube morphology between the long and short morphs. Thus, it is not inconceivable that a similar mechanism exists in *Linum*. It is hoped that the heterostylous species *L.tenue* may share similar *S-locus* genes to those identified in *P. veris*. Bioinformatic analysis of relevant sequence data will assist with the identification of such potential candidate genes. To date, no candidate for the pollen S gene has been identified in *Linum* (Ushijima et al., 2015), although segregating populations which have been developed will be invaluable for future investigations.

Complete genome sequence data for cultivated flax, *L. usitatissimum* is available (Wang et al., 2012) and furthermore, this Master's Thesis describes the transcriptome sequencing of 44 individuals across wild 10 species of *Linum* that has been completed at Durham University. With the advent of this considerable bank of transcriptomic data, it is anticipated that *Linum* will become an even more valuable system for the study of heterostyly. Despite the fact that the flax genus *Linum* is the largest in the family *Linaceae*, little is currently known about the wild species which make up the majority of the genus. It is, however, anticipated that the Durham dataset will provide meaningful insights into the gene expression profile of wild heterostylous species. In addition to the data gathered at Durham University, and analysed within this thesis, European Research Council funding has recently been granted to the Slotte Lab at the University of Stockholm, to perfom transcriptomic analysis of heterostyly at the genus level (Slotte, 2017). Thus, the genes which make up the *S-locus*, in *Linum* might soon be identified with more certainty. This *genus* is of interest for testing competing hypotheses regarding the ancestral condition of heterostyly.

The genus *Linum* has shaped human civilisation both economically and socially over several thousand years (McDill et al., 2009).The earliest archaeological discoveries of wild flax fibres being used by humans to make cords (Kvavadze et al., 2009) date from at least 30,000 years B.C.E. The first evidence of domestication dates from approximately 9,000 B.C.E. (Hillman, 1975; Zohary et al., 2012) at Tell Abu Hureyra in Northern Syria. Flax was the first plant to be cultivated for fibres and, as such, was of significant economic and cultural importance. In modern times, the seeds of cultivated flax, *L. usitatissimum* have superseded *L. bienne* in terms of economic importance. *Linum bienne* is a homostylous species native to the Mediterranean region, although it has a range extending as far north as England and Ireland (PFAF, 2012). *L. bienne* is considered to be the wild progenitor of cultivated flax, *L. usitatissimum* (McDill et al., 2009). *L. usitatissimum* and *L. bienne* are closely related in an evolutionary sense (based on phylograms from McDill et al. (2009)), and observed in the most recent phylogeny conducted by (Ruiz Martín et al., 2018)).

1.5 Future Perspectives

Heterostyly has arisen independently on at least 28 separate occasions making the heterostylous syndrome of both ecological and evolutionary significance. Recent work undertaken in *Primula* (Huu et al., 2016; Li et al., 2016) advances our knowledge to a point at which the potential for a common mechanism across families can be assessed. The species currently receiving the most attention: *T. subulata*, *F. esculentum*, *P. veris* and *P. vulgaris* all fall within separate orders in distinct Eudicot clades. Nevertheless, all have developed similar (predicted or confirmed) *S-locus* architecture under independent evolutionary events (Gilmartin & Li, 2010). Although it remains unknown how widespread the genetic architecture that has been determined in *Primula vulgaris* will be, it is likely that genes encoding equivalent functions will comprise the *S-loci* across a variety of Eudicots. Heterostyly across the genus *Linum* has not been studied to nearly the same extent as in *Primula* species, nevertheless, the discovery of genes comprising the *S-locus* supergene in *Primula* provides a new exciting avenue for research in flax. It is hypothesised, based on Li et al. (2016)'s studies in *Primula*, that *Linum* genes homologous to those comprising the *S-locus* also control the heterostylous polymorphism. It is worth

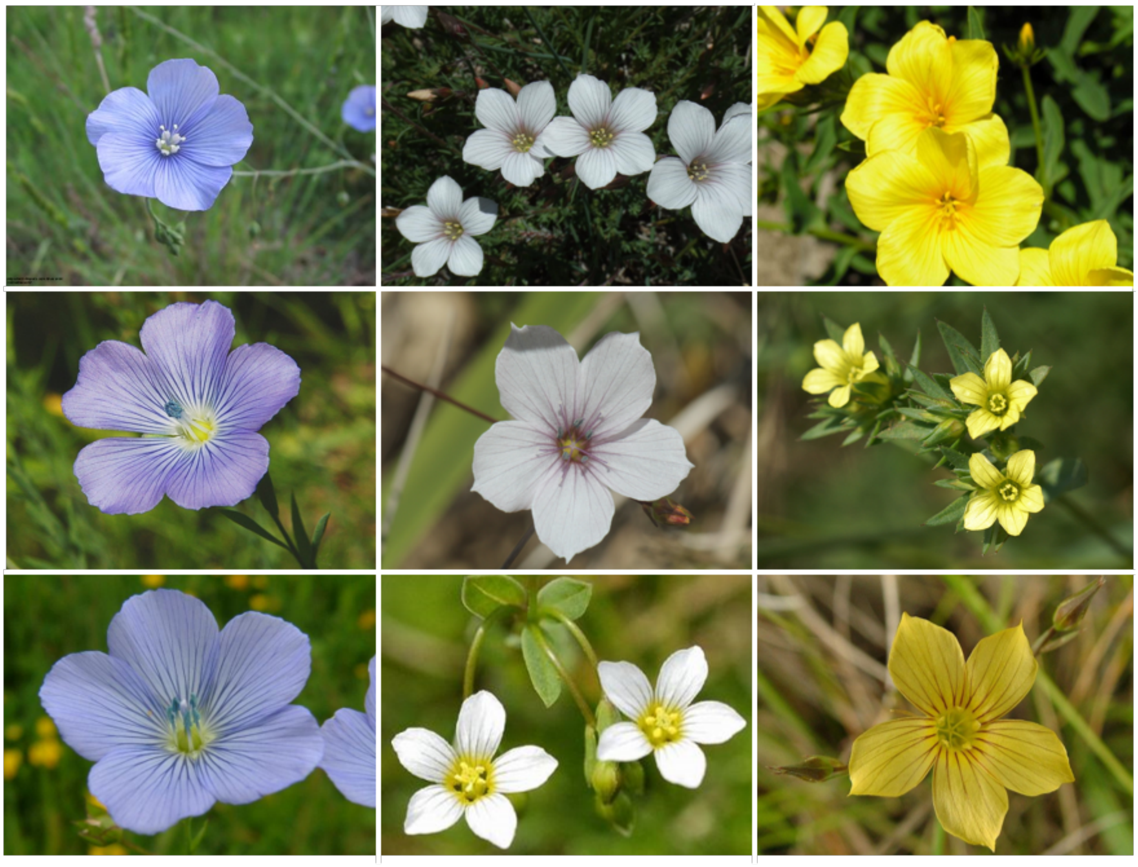


Figure 1.4: The eight wild species (and one cultivated species, *L. usitatissimum*) sequenced for this project. **a)** Heterostylous *L. narbonense* (adapted from: Salguero Quiles (2004)), **b)** heterostylous *L. suffruticosum* (adapted from: Le Petit Herboriste (1998)), **c)** heterostylous *L. campanulatum* L. (adapted from: Mrugala (n.d.)), **d)** homostylous *L. bienne* (adapted from: Viatour (2006)), **e)** homostylous *L. tenuifolium* (adapted from: Godtler (2009)), **f)** homostylous *L. strictum* L. (adapted from: Badia (n.d.)), **g)** homostylous *L. usitatissimum* (adapted from: Fenwick (2005)), **h)** homostylous *L. catharticum* L. (adapted from: Devlin (2006)), **i)** homostylous *L. setaceum* Brot. (adapted from: Ramalho (2012)). *Linum* broadly exists in yellow and blue flowered clades.

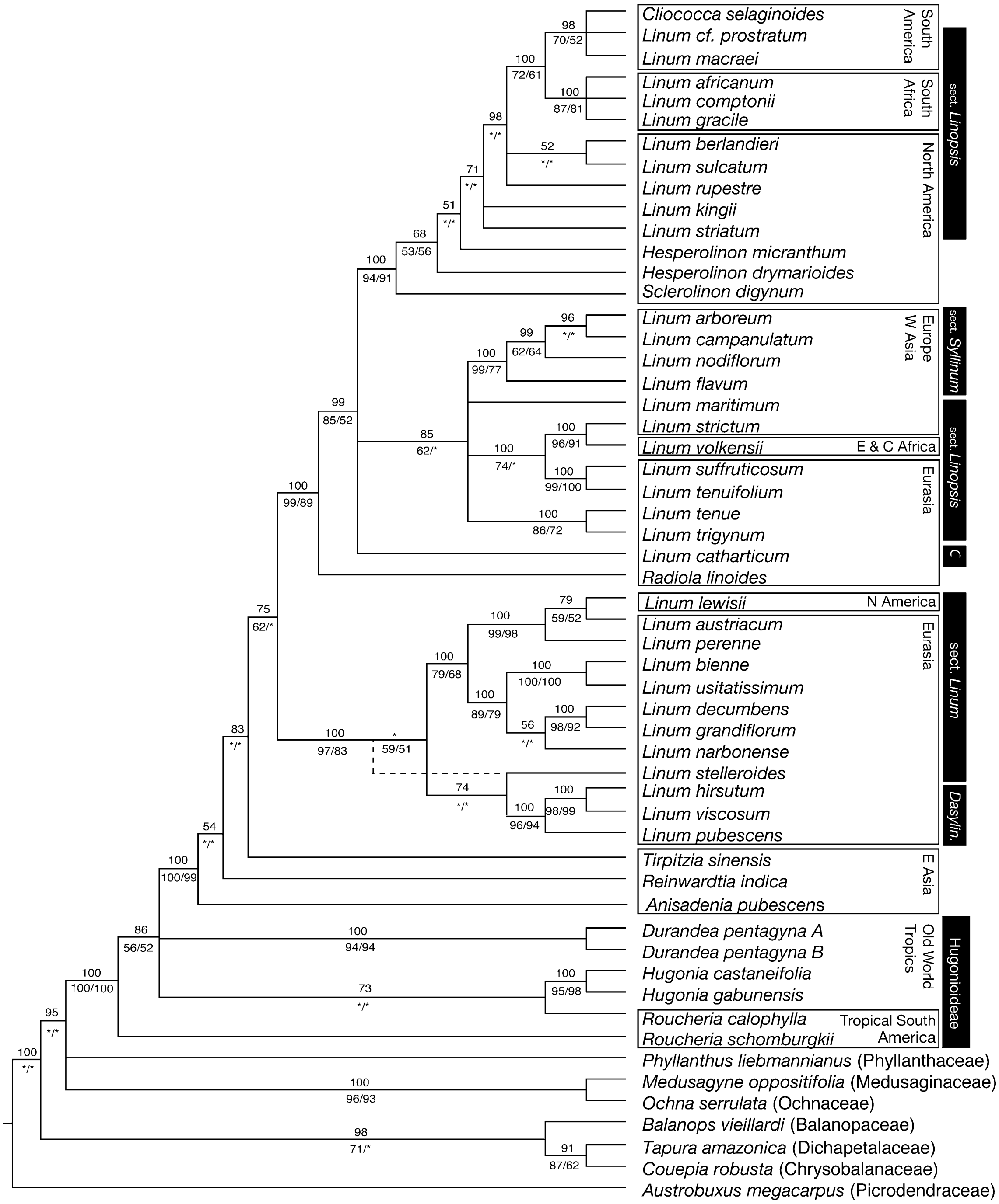


Figure 1.5: The most recent published *Linum* Phylogeny, constructed using the rbcL chloroplast gene (Ruiz Martín et al., 2018)

noting here, however, that although it is the hypothesis of this thesis, there is no reason to assume that the same class of genes will provide equivalent functions; especially given the evolutionary distance between the species involved.

In both *Turnera* and *Primula* it is noteworthy that the formation of distyly involves the absence of one or more genes in the short-styled morph (McClure, 2016). Thus, the approach to identifying the genes at the *S-locus* in a variety of species should include mutation studies and other molecular biological analyses. Furthermore, the creation of bacterial artificial chromosomes (BAC) has recently been completed in a series of heterostylous species. BAC libraries are now being superseded by next-generation sequencing techniques, nevertheless, they remain useful and have led to some of the important insights in *Turnera* and *Fagopyrum* discussed in this review.

Proteomic approaches, like those employed in *Turnera* (Athanasiou & Shore, 1997; Athanasiou et al., 2003) will facilitate finding proteins which segregate with the *S* gene. Through evidence from comparisons between transcript and protein accumulations, Ushijima et al. (2012) hypothesised the existence of morph-specific post-translational modifications associated with heterostyly. It is hoped, too, that proteomic analysis will help to elucidate post-transcriptional modification and regulation. Thus far, however, proteomic approaches towards understanding heterostyly have failed to lead to the identification of genes at the *S-locus*, nor have they characterised any post-translational effects in heterostylous individuals (Ushijima et al., 2012). However, functional analyses of such proteins have enabled confirmation of genes that are not part of the *S-locus*, such as *LgSKS1*.

The annotation of the *P. veris* genome by Nowak et al. (2015) has provided data which will greatly assist bioinformatic analysis of potential candidate genes in other species, including *Linum* and the *S-locus* has been identified using a combination of the data collected by Huu et al. (2016) and Li et al. (2016). The recent determination of the floral architecture of the *S-locus* in *P. vulgaris* (Li et al., 2016) now facilitates the determination of the *S-locus* in other plant species, such as *Linum*. It is possible that orthologues of the 5 thrum-specific genes are also at the *S-locus* in other species. Furthermore, despite advances in determining the genes responsible for style position, the dimorphic anther position remains unconfirmed (McClure, 2016) although Li et al. (2016) present a short homostyle mutant that carries a transposon insertion in *GLO*^{T}. Further characterisation of this mutant remains to be published, however, *GLO*^{T} is presented as the candidate for anther elevation based on sequence analysis of the mutant. Regarding this study in *Linum*, the strong candidate *TSS1* found by Ushijima et al. (2012) in *L. grandiflorum* may be related to one of the 5 *S-locus* genes identified by Li et al. (2016). It is thrum-specific and expressed only in the style, indicative of similarity to the function of CYP^T. In light of the new knowledge of one mechanism controlling heterostyly, it will hopefully be possible to interpret old results and to draw conclusions about the nature of heterostyly in the genus *Linum*.

There is considerable scope to extend the study of the genetic basis of heterostyly to other species of *Linum*. *Linum* in particular is of biological interest owing to the fact that not all members of the genus exhibit heterostyly. It is uncertain whether the *S-locus* will be absent in self-compatible homostyles, such as *L. usitatissimum*, or perhaps be present as a series of pseudogenes. Since some *Linum* species appear to have somewhat different genetic architecture to other heterostylous species; showing reciprocity only in styles, members of the *Linum* genus represent a fascinating candidate for study. With the discovery of this cluster of genes, responsible for reproductive traits in *Primula*, it is hoped that similar genes will be shared across other families. The advancing effects of climate change render an understanding of mating patterns more important than ever as the influence a population's ability to maintain itself. Only through a true appreciation of the genetics governing plant mating patterns can we hope to impact changing biodiversity and food security.

1.6 Aims and Hypotheses

1.6.1 Hypotheses

The following hypotheses were considered as part of this project:

1. Heterostyly in *Linum* is regulated by the *'S-locus'* supergene, as in *Primula* species (Li et al., 2016; Labonne et al., 2009).

2. The genes comprising this *'S-locus'* are orthologues of known *'S-locus* genes from other species, for example *Primula* species from the Li et al. (2016) paper.

3. The genotype of the region is hemizygous.

4. In line with the proposed hemizygous genotype, there is no expression of *'S-locus'* genes in the pin morph; the 278kb *S-locus* region is absent in this morph in distylous *Primula*.

5. The majority of the differential gene expression is at the immature bud (the youngest) stage of development.

6. Homostyles share some expression characteristics of pin morphs and others of thrum morphs, as they have phenotypic features in common with both heteromorphic flowers.

1.6.2 Aims

The experimental focus of the work presented is to identify orthologues of the key genes recently identified in *Primula* (Li et al., 2016), *L. usitatissimum* (Ushijima et al., 2012) and *F. esculentum* (Yasui et al., 2012) as the main regulators of heterostyly from the acquired sequence data using two separate approaches: differential expression analysis (Chapter 2) and qPCR (Chapter 3). RNASeq data from flowers of the genus *Linum* obtained prior to the start of the project was analysed in an attempt to find these sequences. Some insight into the origins of heterostyly may be derived by examining the relationship between homostylous and heterostylous species.

To attempt to determine the genes associated with the '*S-locus*' in the genus *Linum*, a list of candidate genes for heterostyly in *Linum* was compiled from a search of the available literature. From this list, the goal was to design primers by undertaking BLAST searches against the transcriptomes of homostylous and heterostylous *Linum* species and subsequently to determine the efficacy of such primers in the laboratory. Recently-made-available sequence data provided an alternate avenue for assessment of the '*S-locus*' candidate genes using differential expression analysis programs. The sequences were mapped to a reference transcriptome and the mapped sequences of individuals exhibiting different floral morphs and varying mating patterns were compared bioinformatically. Thus, the identification of '*S-locus*' genes was attempted in *Linum* using two alternative differential expression approaches: bioinformatic analysis of RNA-Seq data and quantitative PCR (qPCR). This project assessed ten wild *Linum* species exhibiting a wide variety of mating strategies and it was anticipated that the results achieved from the RNA-Seq data analysis would be cross-validated using qPCR. Based on the experiments conducted by Li et al. (2016), it would be expected that candidate genes for heterostyly will be expressed in the short-styled thrum individuals and not in the long-styled pin individuals. Following from assertions of hypotheses 3. and 4., it was also determined that it would be possible to consider heterostylous individuals from different wild species as biological replicates of one another.

A workflow for the analysis of RNA-Seq data was defined as a byproduct of the bioinformatic analysis of potential '*S-locus*' genes; best practices for RNA-Seq analysis have been hotly debated, and no consensus has been reached. Additionally, no single analysis pipeline is suitable for use in all cases (Conesa et al., 2016). The nature of the data collected for this experiment, namely having sequences from 10 different wild species at different growth stages, as opposed to standard control and treatment cases, makes for a unique experimental analysis set-up, and required sensitive software selection.

The hypothesis that the *S-locus* in *Primula* shares a common evolutionary origin with the purported locus in *Linum* does carry the following caveat: the divergence of *Primula* and *Linum* may be of a similar time frame to that of *Primula* and *Fagopyrum* (118 million years). Therefore failing to find conserved sequences involved in heterostyly may not be surprsising as they are likely to have arisen by convergent evolution from independent origins in different genera. A gene duplication that led to the formation of the *Primula* heterostyly supergene has been dated to 53.7 MYA, long after the divergence of the above three heterostylous species.

It is evident with hindsight that the samples collected prior to the start of the project would be better suited to a comparative sequence analysis of expressed genes between *Linum* and other species rather than the differential expression analysis which was attempted. A comparative analysis would determine the presence or absence of transcripts rather than transcript abundance. Nevertheless, the actual analysis undertaken is reported in the results and the shortcomings associated with this method are addressed in the discussion.

Chapter 2

RNA-Sequencing Analysis

2.1 Introduction

DNA encodes information that can determine the functions and properties of every cell in an organism. Cells dynamically access and translate specific functions through gene expression: i.e. they selectively switch on and off particular combinations of genes (Finotello & Di Camillo, 2015) to achieve specific effects. The information encoded in these genes is transcribed into messenger RNA (mRNA) transcripts which can, in turn, be translated into proteins, or can bind directly to molecules to finely control gene expression. The transcriptome consists of the total set of transcripts in a cell, including messenger RNA (mRNA) , destined for translation into proteins and non-coding RNAs such as ribosomal RNA (rRNA) . These transcripts represent a profile of the genes that are being actively expressed in a cell at a specific point in time, distinct developmental stage or under a particular physiological condition or stress. The transcriptome is dynamic and, unlike the genome, is specific to different tissues at different developmental stages (Wang et al., 2009).

Next Generation Sequencing (NGS) technologies first emerged in the early 2000s and in the decades since, have had a dramatic impact on the scope of genomic and transcriptomic research (Mardis, 2011).Historically, ecological investigations using transcriptomics have been largely confined to model species and their close relatives. The significant barrier imposed by the required infrastructure and financing prevented the analysis of non-model species assemblages (Swenson & Jones, 2017). However, recent rapid reductions in the cost and increases in the availability of such transcriptomic techniques have led to a dramatic increase in the occurrence of such studies in non-model species. The development of such approaches has improved the qualitative and quantitative data obtained, whilst also reducing costs. Consequently, using transcriptomics presents a distinct advantage when assessing the expression of genes across distylous floral morphs and homostyles; RNA-Seq analysis enables highly sensitive and accurate measurements of gene expression across the transcriptome (Illumina, 2017). RNA-Seq gives reads from the ends of a random sample of short DNA fragments from a created sequence library of reverse-transcribed cDNA. Sequencing of sufficient depth should give reads of fragments of all mRNA present in the tissue at the time of analysis; they represent a snapshot of expression. In order to provide useful information, reads must be combined (assembled) into larger fragments, representing the total mRNA transcript. Once combined, these sequences are referred to as contiguous sequences or "contigs". The pipeline designed and utilised in this investigation incorporates publicly available tools and scripts written by the author to evaluate the RNA-Seq output.

As depicted in Figure 2.1, the NGS library is prepared by fragmenting a sample of DNA and ligating specialised adapters to both the 5′ and 3′ fragment ends. The Pippin Prep (Sage Science, Beverly, MA, USA) was used to facilitate library construction and for size selection. The Pippin Prep uses an electrophoresis platform for the preparation of size-fractionated RNA samples. The library is then loaded into a flow cell and the fragments are hybridised to adapter-complementary oligonucleotides attached to the flow cell surface. Single-stranded, adapter-ligated fragments thus become bound to the surface of the flow cell. Each bound fragment is amplified into a distinct clonal cluster through bridge amplification with Taq polymerase and sequencing primers. Sequencing reagents, including fluorescently-labelled nucleotides are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The wavelength and intensity of the emission are used to identify the specific base. This process is repeated 'n' times to create a read length of 'n' bases (Illumina, 2017).

The process of RNA-Seq data analysis begins with quality assessment and control of the raw reads, in this case using a combination of FastQC (Andrews, 2010) and MultiQC (Ewels et al., 2016), ensuring that the sequence reads are of sufficiently high quality to proceed with downstream analyses. Reads are trimmed to remove the Illumina sequencing adapters using Trimmomatic. Once trimmed and re-assessed for quality, the clean reads are aligned to a reference transcriptome using STAR (Dobin et al., 2013). The mapped reads are subsequently converted into coordinate-sorted binary output (BAM) files using SAMtools. Raw gene expression counts of the number of mapped reads per reference transcript, were measured using the python-based package HTSeq-count (Anders et al., 2015). Normalisation and differential expression testing was conducted using the DESeq2 package for R statistics (Love et al., 2014). A list of differentially expressed genes was obtained and raw read counts of contigs containing candidate genes

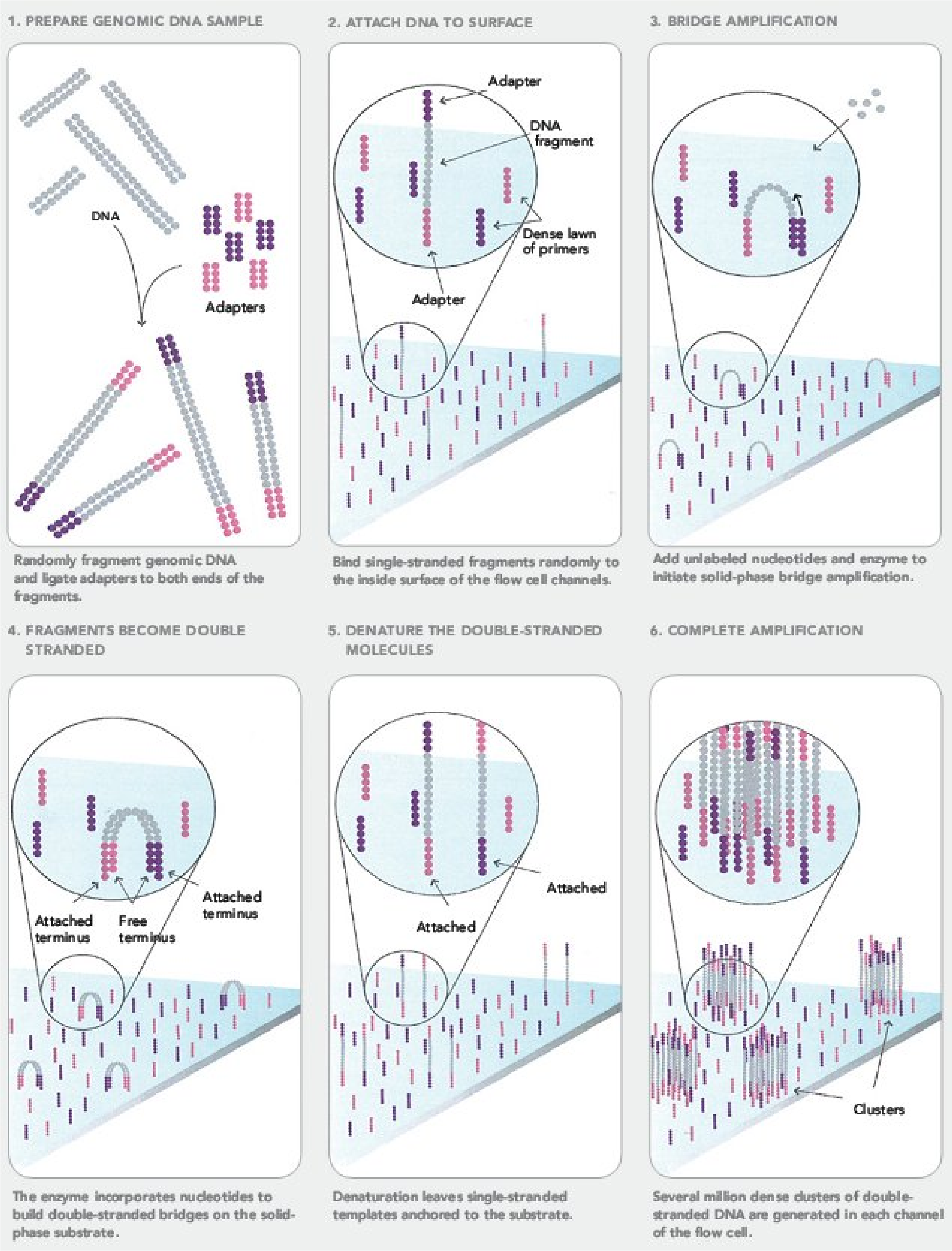

1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 2.1: Schematic of the Illumina 2500 NGS Workflow. Adapted from SeqAnswers (2007)

were extracted from the datasest. The analysis pipeline was developed through a review of recent literature (e.g. Nowak et al. (2015); Huu et al. (2016)) and through taking into account advice gathered from online resources (e.g. SeqAnswers, 2007; Biostars, n.d.). Nevertheless, at every stage of the pipeline there are alternative, publicly available tools to those selected. Some of the most popular substitutes are listed in Table 2.1.

Table 2.1: Alternative publicly available mapping tools for each stage of the pipeline.

Pipeline Step	**Tool Used**	**Alternative Tools**
Quality control	FastQC	HTQC, RSeqQC
Alignment	STAR	Bowtie2, BWA, Tophat, HISAT2
Gene expression level	HTSeq-count	FeatureCounts, summarizeOverlaps
Gene expression profiling	DESeq2	edgeR, Limma

Given the ability of high-throughput sequencers to generate enormous quantities of data in a single run, adequate methods of analysing such sequences must be developed. Analysis of high-throughput sequencing is a rapidly burgeoning area of bioinformatics, and consequently the pipelines used are being constantly updated.

NGS Sequencing techniques are used in this chapter to compare the differential expression of orthologues of identified candidate genes in *Primula* and *L. usitatissimum*. As has been discussed elsewhere in this thesis, owing to the lack of replicates in the initial sequencing run, the differential expression analysis conducted cannot be used to make statements about the significance of changes. However, this, is not to say that the data is not useful. Qualitatively, expression patterns can be observed. Thus, this study can be considered as more of a broad-reaching, fact-finding exercise which could form the basis of future research.

2.2 Materials and Methods

The pipeline used to analyse the RNA-Seq data is shown in Figure 2.2. The Hamilton Cluster at Durham University was used for high performance computing.

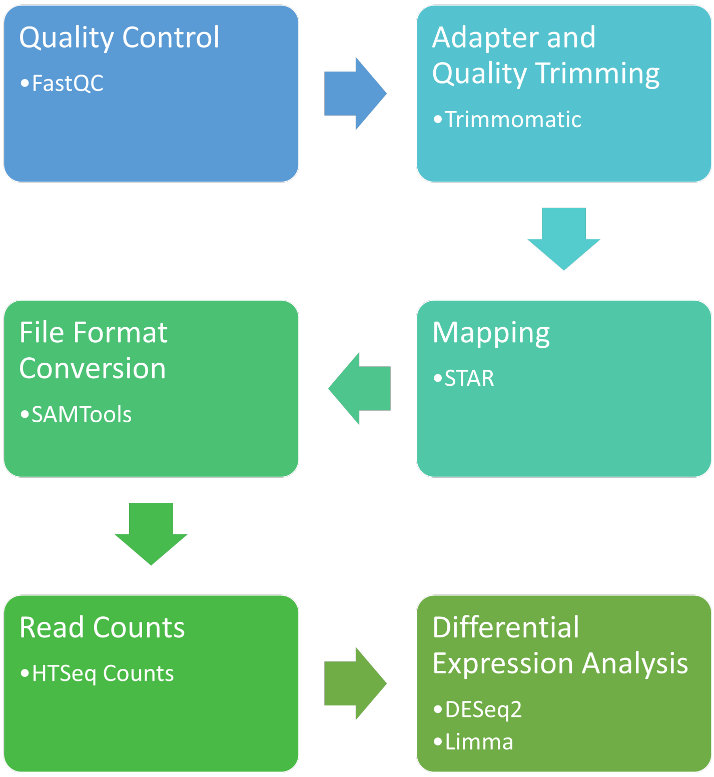


Figure 2.2: Flow chart depicting the RNA-Seq pipeline used to determine differential expression of candidate genes. Paired Illumina reads were mapped to a reference *L. tenue* assembly using STAR. HTSeq-counts was used to determine raw read count expression data from the aligned RNA-Seq reads. Differential expression analysis was conducted using DESeq2.

2.2.1 Sequencing

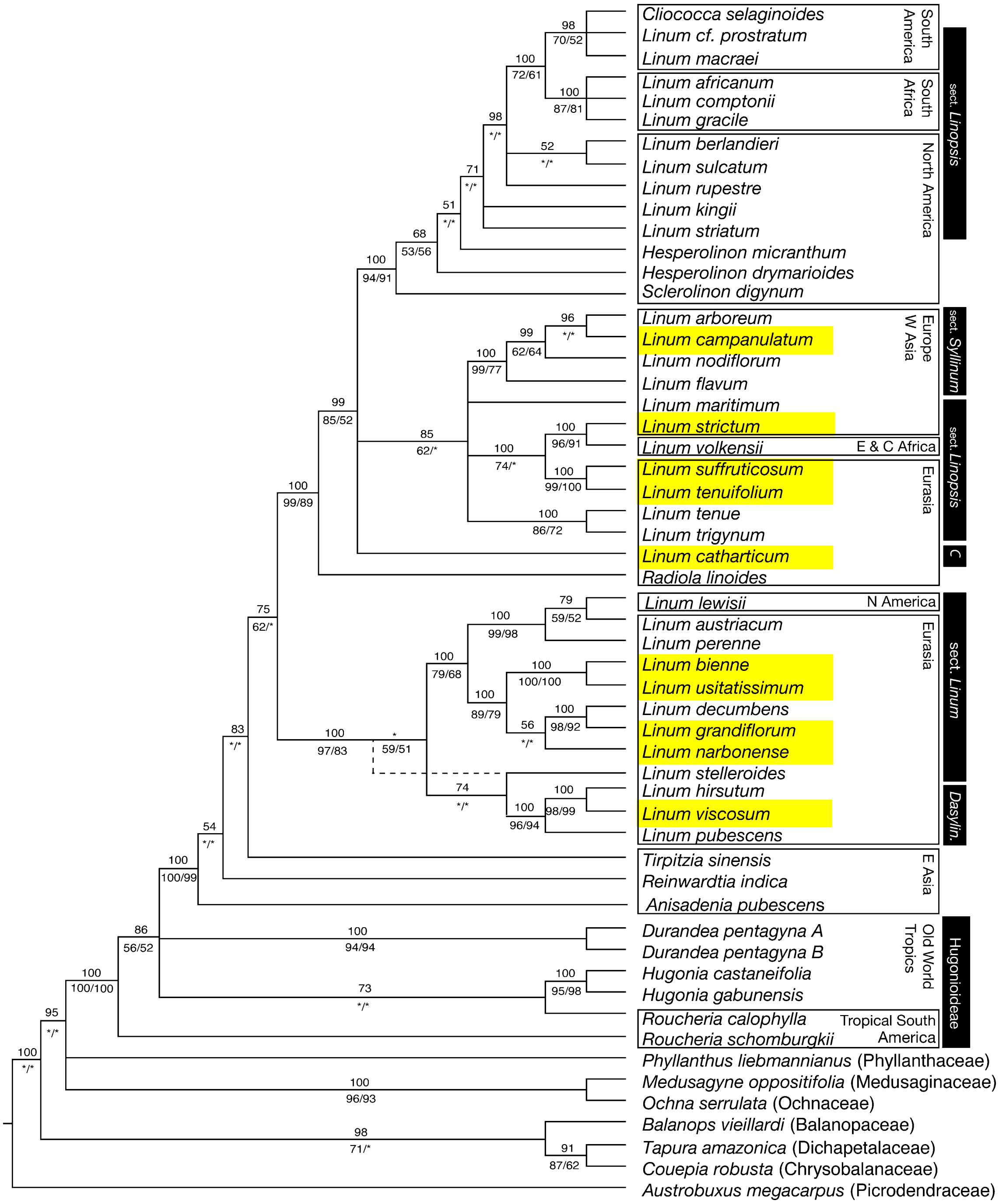


Figure 2.3: Species selected for RNA-Seq analysis, highlighted in yellow on the most recent published *Linum* Phylogeny (adapted from Ruiz Martín et al. (2018)), with the exception of *L. setaceum*, which is not depicted as part of this phylogeny.

Illumina technology was used to sequence total RNA from 44 individuals across 10 different *Linum* species (highlighted in yellow in Figure 2.3). Each individual was sequenced at multiple different developmental stages (see Figure 2.5). Wild *Linum* plants were collected at field sites across Spain and the UK and stored in 1.5 mL RNALater in screwtop tubes and later stored frozen at -80 °C from about three weeks post col-

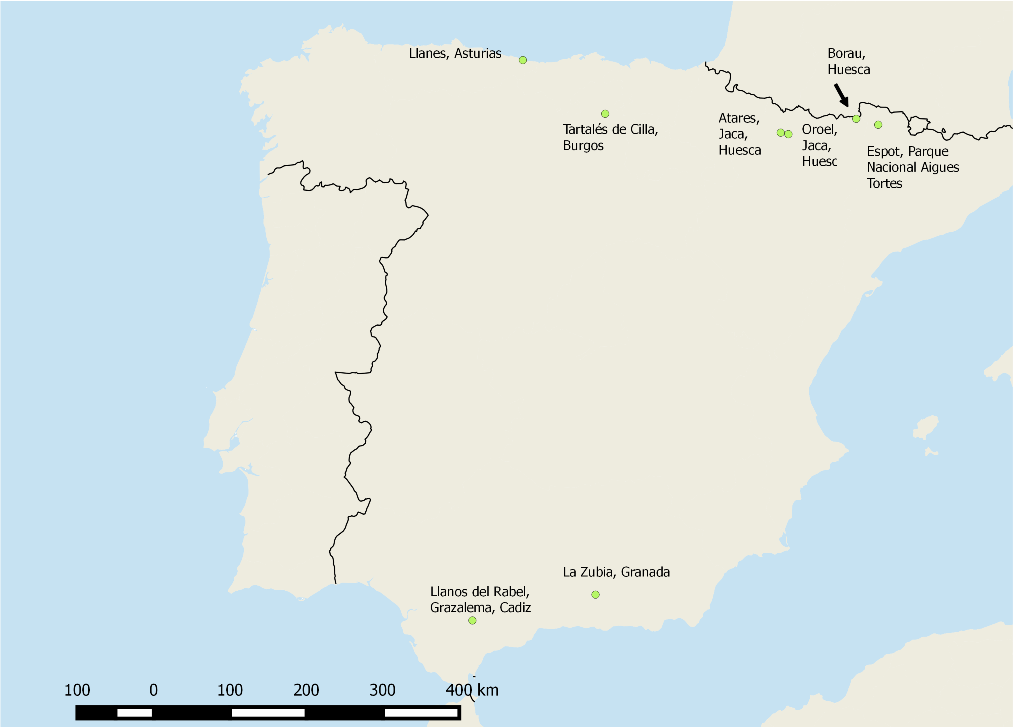


Figure 2.4: Map of the wild species collection sites in Spain. Graphic produced using QGIS 2.6.13 software (QGIS Development Team, n.d.).

lection (Pérez-Barrales et al., personal communication). Spanish field sites are depicted in Figure 2.4. Prior to the start of this project, RNA was extracted from the collected individuals at Durham University according to methods that are described in detail in Chapter 3 and this RNA was used to generate Illumina RNA libraries using TruSeq total RNA library prep kits (Illumina, Cambridge). Each library was sequenced on an Illumina HiSeq 2500 machine at the Durham Genomics Facility. Multiplexing of adapter sequences was employed using the A and B Illumina index sets and the data was demultiplexed as an initial bioinformatics service at the Durham University Genomics Facility. Two sequencing runs were completed for the same set of libraries. The first was undertaken in November 2016, and the second in August 2017. The initial sequencing of the *L. narbonense* libraries was unsuccessful and these results were unable to be mapped or included in downstream analyses. This and other lower sequence read depths led to the decision to conduct the second run. Read length for the August 2017 sequencing was 100 bp , shorter than the 125 bp read length from the November 2016 sequencing. Table 2.2 shows the species and developmental stage of the wild flowers that were sequenced, as well as the population in Spain or the UK from which they were collected.

Table 2.2: The species, developmental stage and population of all of the wild flowers sequenced for analysis of heterostyly. The developmental stages are described in Figure 2.5 and Table 2.3

.

Species	Morph Type	Developmental Stage	Population
L. bienne	Homostylous	GS1	Dorset, UK
L. bienne	Homostylous	GS2	Dorset, UK
L. bienne	Homostylous	GS3	Dorset, UK
L. bienne	Homostylous	GS1	Llanes, Asturias
L. bienne	Homostylous	GS2	Llanes, Asturias
L. bienne	Homostylous	GS3	Llanes, Asturias
L. campanulatum	Short Style	GS1	Borau, Huesca
L. campanulatum	Short Style	GS2	Borau, Huesca
L. campanulatum	Short Style	GS4	Borau, Huesca
L. campanulatum	Long Style	GS1	Borau, Huesca
L. campanulatum	Long Style	GS2	Borau, Huesca
L. catharticum	Homostylous	GS1	Tartalés de Cilla, Burgos
L. catharticum	Homostylous	GS2	Tartalés de Cilla, Burgos
L. catharticum	Homostylous	GS3	Tartalés de Cilla, Burgos
L. narbonense	Short Style	GS1	Oroel, Jaca, Huesca
L. narbonense	Short Style	GS2	Oroel, Jaca, Huesca
L. narbonense	Short Style	GS3	Oroel, Jaca, Huesca
L. narbonense	Short Style	GS4	Oroel, Jaca, Huesca

Table 2.2

Species	**Morph Type**	**Developmental Stage**	**Population**
L. narbonense	Long Style	GS1	Oroel, Jaca, Huesca
L. narbonense AR006	Long Style	GS2	Oroel, Jaca, Huesca
L. narbonense	Long Style	GS3	Oroel, Jaca, Huesca
L. narbonense	Long Style	GS4	Oroel, Jaca, Huesca
L. setaceum	Homostylous	GS1	Llanos del Rabel, Grazalema, Cadiz
L. setaceum	Homostylous	GS2	Llanos del Rabel, Grazalema, Cadiz
L. strictum	Homostylous	GS1	La Zubia, Granada
L. strictum	Homostylous	GS3	La Zubia, Granada
L. suffruticosum	Short Style	GS1	Oroel, Jaca, Huesca
L. suffruticosum	Short Style	GS2	Oroel, Jaca, Huesca
L. suffruticosum	Long Style	GS1	Oroel, Jaca, Huesca
L. suffruticosum	Long Style	GS2	Oroel, Jaca, Huesca
L. suffruticosum	Long Style	GS4	Oroel, Jaca, Huesca
L. tenuifolium	Homostylous	GS1	Espot, Parque Nacional Aigues Tortes

Table 2.2

Species	**Morph Type**	**Developmental Stage**	**Population**
L. tenuifolium	Homostylous	GS2	Espot, Parque Nacional Aigues Tortes
L. usitatissimum	Homostylous	GS1	Marmalade, Canada
L. usitatissimum	Homostylous	GS2	Marmalade, Canada
L. usitatissimum	Homostylous	GS3	Marmalade, Canada
L. viscosum	Long Style	GS1	Oroel, Jaca, Huesca
L. viscosum	Long Style	GS4	Oroel, Jaca, Huesca
L. viscosum	Short Style	GS1	Atares, Jaca, Huesca
L. viscosum	Short Style	GS4	Atares, Jaca, Huesca

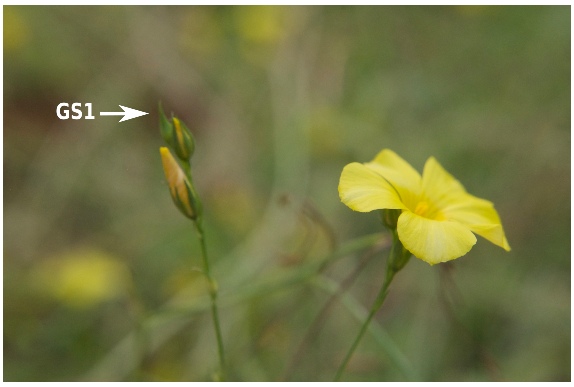


Figure 2.5: The four developmental stages sampled for bioinformatic analysis. From left, GS1: immature bud, GS2: maturing bud GS3: flower opening, but has not yet completed development, GS4: open flower. Species depicted is *L. tenue*. Photography by Amy Stockdale.

Table 2.3: Qualitative Description of Floral Morph Growth Stages.

Growth Stage	**Description**
GS1	Immature buds.
GS2	Maturing buds: beginning to open, petals visible through the top of the bud.
GS3	Flower closed, but has completed growth. Petal length now exceeds that of the sepal.
GS4	Open flowers.

As depicted in Figure 2.5, each individual was sampled at one of four different growth stages (GS). The youngest, immature buds were referred to as GS1; no petal tissue can be seen. A bud at GS1 is depicted on the far left of Figure 2.5. GS2 refers to a maturing bud, depicted adjacent to the immature bud in Figure 2.5, where some petal tissue is visible and the bud is larger. At GS3, the flower is on the point of opening, but has not yet completed development. GS4 represents an open flower. Unfortunately, not all growth stages were able to be collected for every species (Table 2.2) as, at the library preparation stage, only the RNA extractions with the highest yield in terms of RNA concentration were selected for sequencing. The high cost of the RNA-Seq process meant that those individuals with the highest chance of success were chosen.

The RNA-Seq analysis was conducted with paired-end data. It is commonly accepted (e.g. Trapnell et al., 2012; Katz et al., 2010) that using paired-end data, i.e. data consisting of reads from both ends of the mRNA fragments instead of just from one end, is an improvement over single-end data alone, despite the large increase in cost; up to twice the cost of a single-end experiment. Information about mate pairs is used during the read alignment to more accurately determine mappings (Frazee, 2015).

$$Q = -10 \log_{10} P \tag{2.1}$$

The Illumina sequencer outputs a FASTQ file for each library. FASTQ files consist of four lines per read: the first line is the read ID, the second contains the sequence data for that read, the third is redundant and the fourth encodes the Phred quality score, coded as a series of 40 letters and symbols. An example fastq output is depicted in Figure 2.6. Phred quality scores, Q, are logarithmically related to the base-calling error probabilities, P, where P = 1. (Equation 2.2.1) (Ewing et al., 1998). Phred scores are ranked on a scale of 1 - 40; a Phred score of 30 indicates that the chances of this base being called incorrectly are 1/1000 or a 99.9% base-call accuracy rate. Similarly, a Phred score of 40 indicates a 1/10,000 probability of an incorrect base-call.

```
@D00498:66:CBC7BANXX:2:2202:8881:2048 2:N:0:TTACCA
ACTGGATAATCTCAAGGTTGAGACTTGATCCCAGGAAGGACGTTGATTTTGTTCCATTGCCTGGTCCTCTTCTTCGTAA$
+
//<//FBFFFFFFFFF</<<F//</B//<///F///<///B//FFFBF<BFBF///<///<///FB///<BBB/</F</$
@D00498:66:CBC7BANXX:2:2202:9587:2461 2:N:0:TGACCA
GTGAAATGGTTCATACATGTTTGTTGCCGGGGGCAACAAGATCAAGTTGGTAAGGATTTCAAAATTACCTTCATTTTTC$
+
```

Figure 2.6: Excerpt from an example .fastq file: *L. viscosum* x T GS1 2U.fastq file. Unpaired Trimmomatic data was used for this visualisation as the smaller file required less space once uncompressed.

The example FASTQ data shown in Figure 2.6 demonstrates a reasonably low quality sequence. The fourth line is encoded using ASCII quality value characters, with '!' representing the lowest quality, and '~' representing the highest quality. The quality value characters in increasing order, left to right, is depicted in Figure 2.7.

```
!"#$%&'()*+,-./0123456789:;<=>
?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Figure 2.7: FASTQ value characters in increasing order of quality.

2.2.2 Quality Control and Pre-processing

The first stage of the pipeline involves read file processing with FastQC (Andrews, 2010) and MultiQC (Ewels et al., 2016). FastQC is a program written in Java that enables a

```
TruSeq3-PE-.fa:3:30:10:8:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50
```

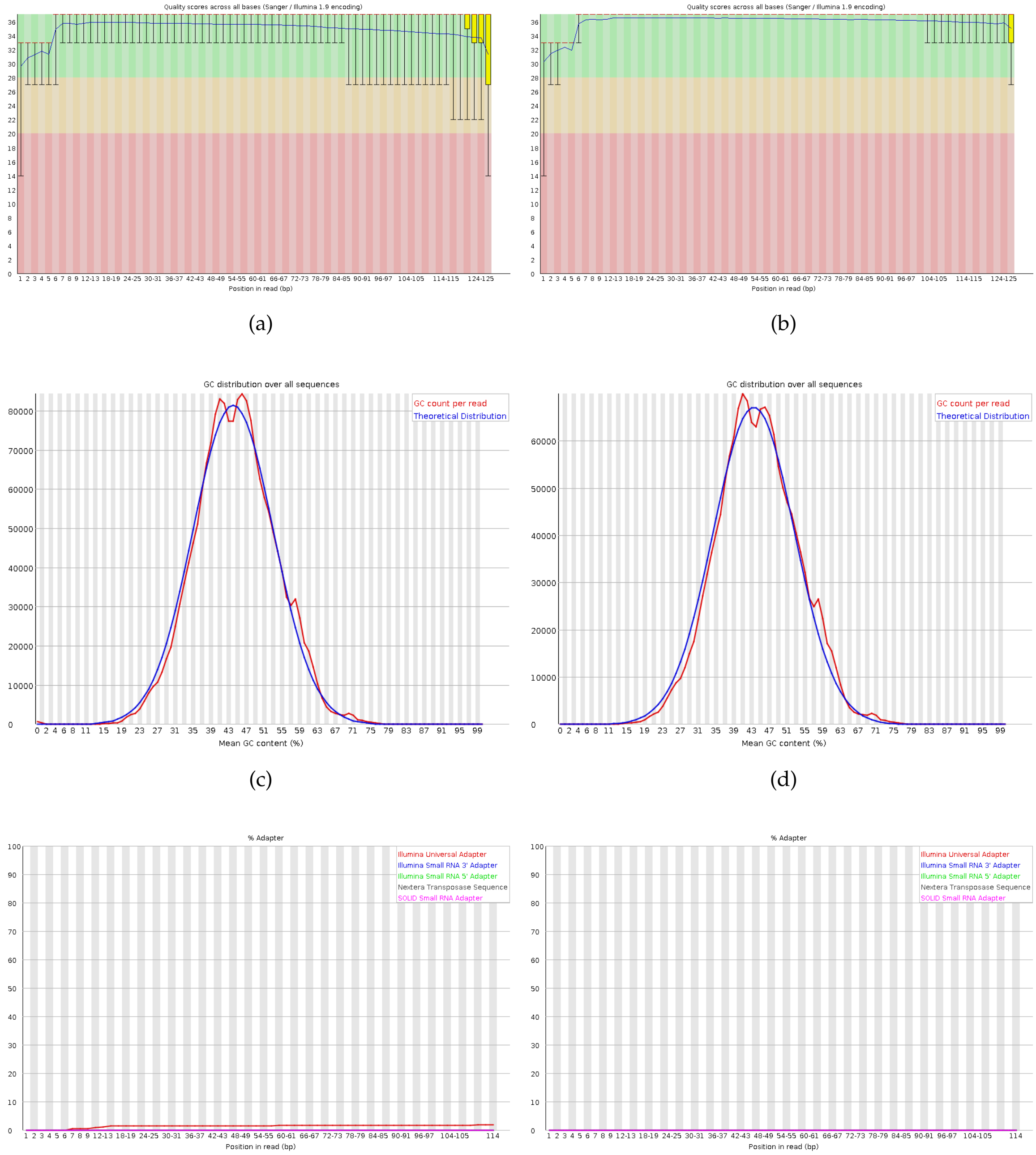
Figure 2.8: Minimal Example of Trimmomatic input code.

qualitative assessment of the quality of the sequence reads. FastQC reports a number of metrics which summarise all of the reads in the file, including: per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and Kmer content (a Kmer is a motif of length k bases, observed more than once in a sequence). MultiQC summarises FastQC analysis results across multiple sequences, thus facilitating the simultaneous visualisation of the quality scores of multiple sequence files and by extension their interpretation. Potential quality issues flagged by FastQC were manually assessed, and if deemed necessary, were rectified by re-trimming with Trimmomatic (Bolger et al., 2014) to trim for specific overrepresented sequences or by omitting the sample from downstream analysis. Issues identified by FastQC may arise from problems with the sequencing run, or from defects in the initial library starting material.

Trimmomatic v.0.36 (Bolger et al., 2014) was used to filter the RNA-Seq data, both for quality control, removing low quality base calls and to remove sequencing adapters. Adapter sequences occur only at the 3′ end if the fragment is shorter than the number of sequencing cycles (eCSeq Bioinformatics, 2016). Trimmomatic is a lightweight Java application capable of removing Illumina adapter sequences and low quality reads. It uses a sliding window to analyse sections of each read (Zhernakova et al., 2013). The adapter sequences, added during Illumina sequencing must be removed from the output, if they have escaped removal during the initial read processing by Illumina's BaseSpace software (https://basespace.illumina.com). The newly trimmed reads were then re-assessed using FastQC to determine the efficacy of the processing. Trimmomatic parameters were set according to Figure 2.8 following initial optimisation tests.

Trimmomatic removes Illumina adapters provided in the TruSeq3-PE.fa file. Trimmomatic initially will look for seed matches, allowing up to 2 mismatches. In this context, a seed represents a short aligned region between the adapter and the read. Short sections of each adapter (up to a maximum of 16 bp) are tested in each possible position within the sequence reads. If the short alignment (“seed") is a sufficiently close match, the entire read-adapter alignment is given a score increment (Bolger et al., 2014). In the case of the Figure 2.8 example, these seeds will be extended and clipped if a score of 30 is reached (Paired End, PE) or a score of 10 is reached (Single End, SE). This will remove leading and trailing low quality or N bases below a quality score of 3. It will also scan the read with a 4-base wide sliding window, cutting when the average quality per base falls below 20. The trimmer will cut the left-most position in the window, beyond the point where the quality drops below the specified threshold (Bushnell, 2015b). The trim also drops all reads which are less than 50 bases long following these steps. Trimmomatic produces four output files: Forward unpaired, Forward paired, Reverse unpaired and Reverse paired. Only the paired data was carried forward to the mapping stage.

Apart from the removal of adapter sequences, it is possible to resolve issues with sequence quality towards the ends of reads through trimming base-pairs off the raw read and additionally to specify particular highly expressed sequences for removal. There has, however, been some controversy over the benefits of removing base-pairs to improve quality, with some advocating only trimming reads to remove adapter sequences, rather than attempting to cut only those reads of low quality (e.g. Yang & Kim, 2015). In this analysis, trimming was undertaken only to remove remaining adapter sequences in order to prevent the unnecessary exclusion of highly expressed sequences which could be of biological relevance.



(a)

(b)

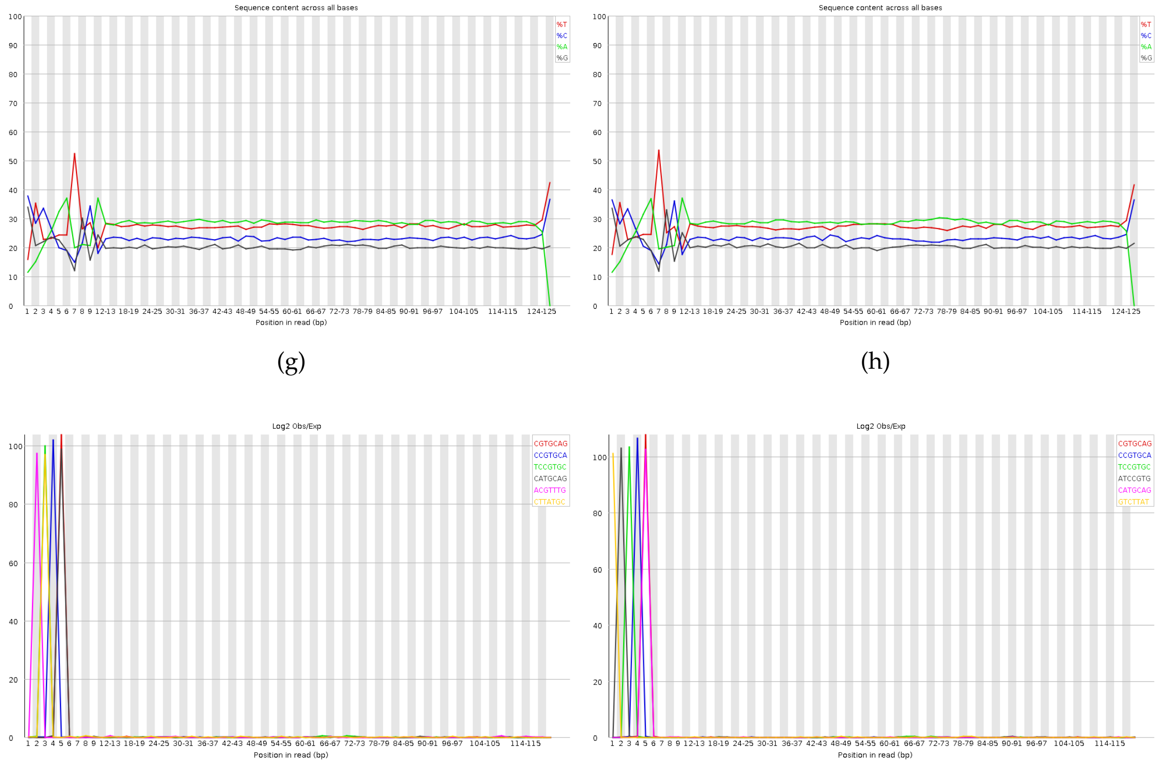
(c)

(d)

(e)

(f)

Figure 2.9: FastQC data both pre- and post- trim for *L. bienne* dor5 at GS1. Only one sequence is included here, by way of example. **Figures a) and b)**: Box plots of the base quality scores at read positions across all reads in a sample for pre-trimmed and post-trimmed samples respectively. Quality scores are depicted on the y-axis. **Figures c) and d)** summarise the GC content across the whole length of each sequence in a file and compare it to a modelled GC normal distribution. **Figures e) and f)** depict the adapter content and **Figures g) and h)** show the k-mer content. This is an analysis of overrepresented sequences that will spot an increase in exactly duplicated sequences.



(g)

(h)

(i)

(j)

Figure 2.9: **Figures g) and h)** indicate the sequence content across all bases **Figures i) and j)** show the k-mer content. This is an analysis of overrepresented sequences that will spot an increase in exactly duplicated sequences.

Overall, the quality of the *Linum* RNA-Seq data was high. Nevertheless, warnings presented on several of the examined quality modules. The most commonly attained warnings were for Sequence Duplication Levels, Per base sequence content (e.g.Figure 2.11d) and Kmer content (e.g. Figure 2.9j). The Sequence duplication levels and Kmer content are largely an artefact of random hexamer priming during library preparation. Libraries which derive from random priming will nearly always show a bias at the start of the library due to an incomplete sampling of the possible random primers (Babraham Bioinformatics, 2018). The biased selection of random primers in the first 12 bp of each run appears to have no impact on the ability to measure expression. Nevertheless, comparison with an example low-quality FastQC output (e.g. 2.10) confirms the quality of the obtained sequence data.

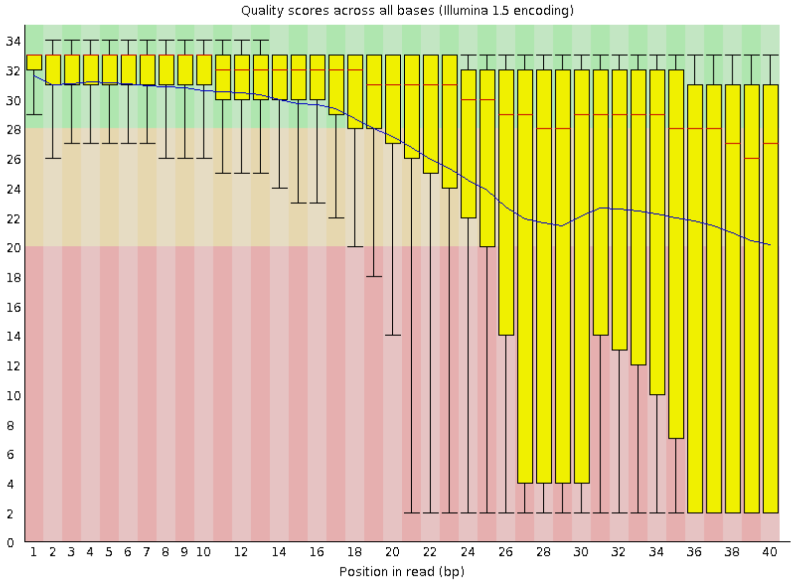
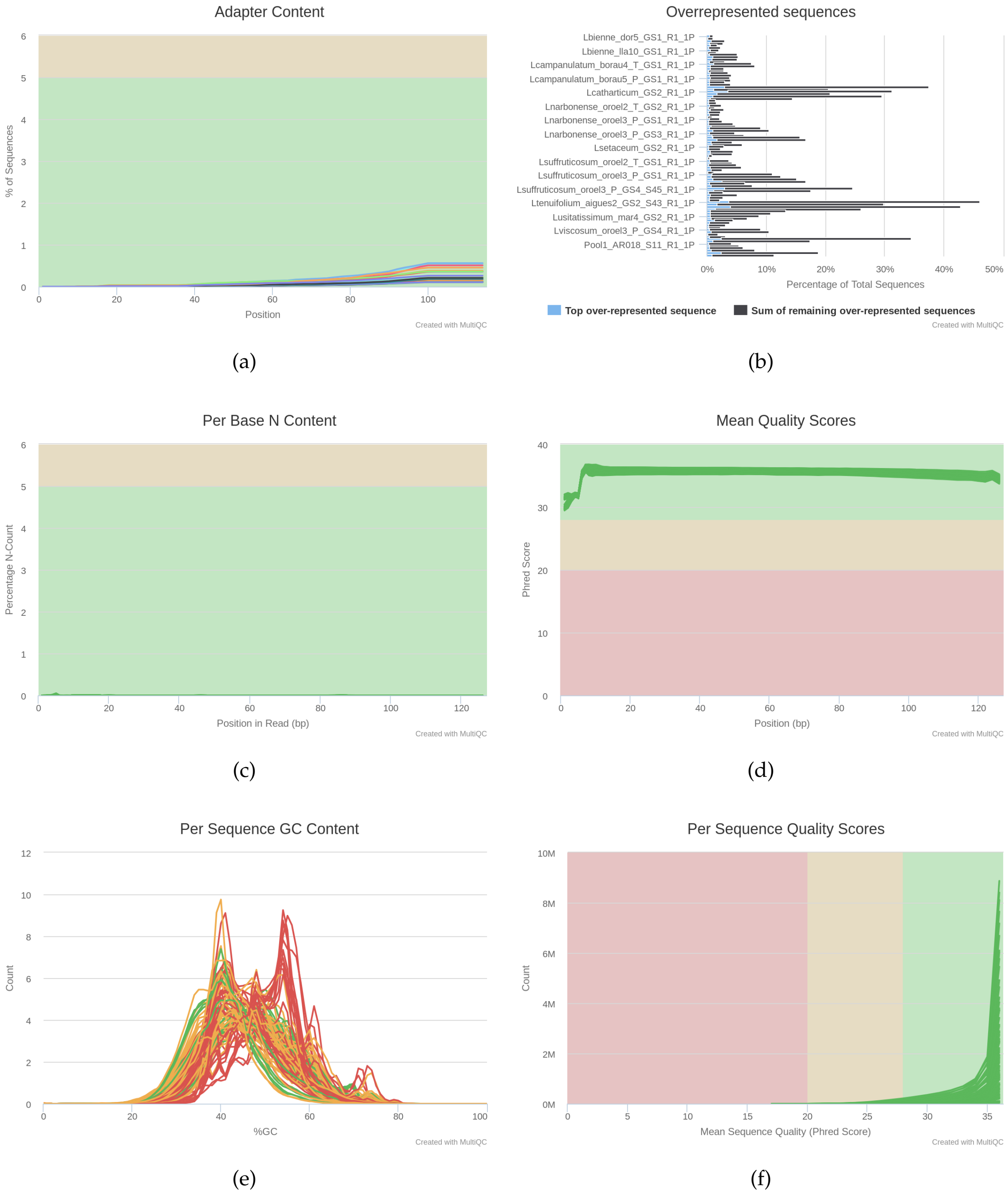


Figure 2.10: The per base sequence quality graph for an example poor quality RNA-Seq run. Boxplots display the mean and interquartile ranges of the Phred quality score across reads. Image adapted from: Babraham Bioinformatics (2017)

Figure 2.9 shows the FastQC output for one individual, *L. bienne* dor5 at GS1 before and after the trimming process. As can be observed from Figure 2.9: prior to trimming, the quality score drops to approximately 25 near the end of the 3'end. After trimming, the quality scores were unanimously high. From Panels 2.9e and 2.9f show the effective removal of adapter content from the sequence during the trimming process and Panels 2.9g and 2.9h are indicative of the overall sequence composition per base. Panels 2.9i and 2.9j show little difference in the Kmer content. This is as expected, as trimming for overrepresented sequences was not undertaken. The top six most positionally biased Kmers are reported to show their distribution and it is evident from these results that the most enriched Kmers all occur at the very beginning of the read. As described above, this is likely to be because of a bias in the selection of random primers in the first few bases and is not a fault that can be fixed by trimming. MultiQC is a versatile tool, capable of aggregating the results from bioinformatic analyses of many samples into one single report. MultiQC was used to compile the pre-trim and post-trim results into a format that could be more easily interpreted, and produced lists of the most overrepresented sequences.



(a)

(b)

(c)

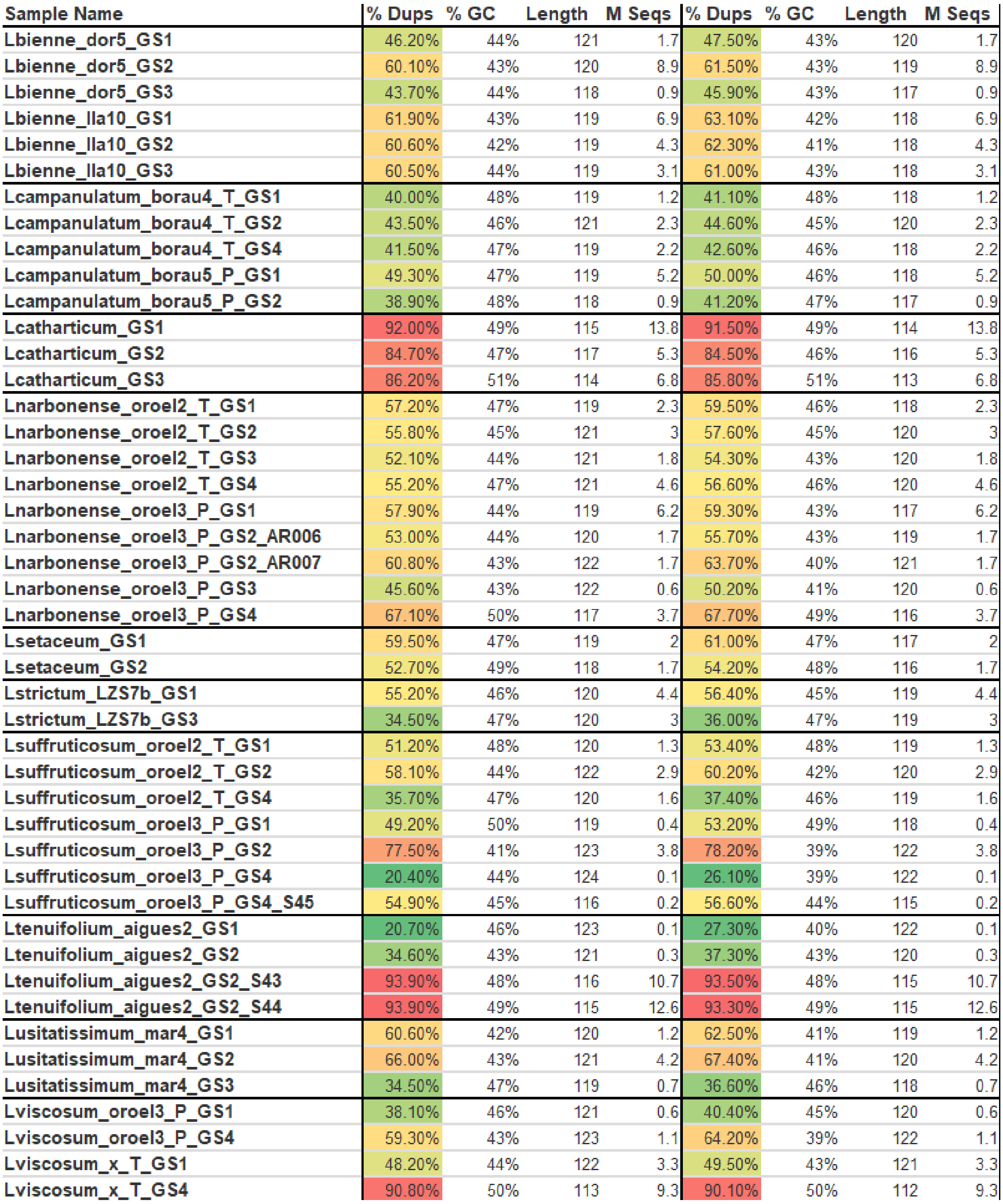
(d)

(e)

(f)

Figure 2.11: Post-Trim data for all sequenced individuals; graphs produced using MultiQC. **Panel a)** shows the cumulative percentage count of the proportion of the library which has seen the adapter sequences at each position. **Panel b)** shows the percentage of overrepresented sequences found in each library. **Panel c)** indicates the percentage of base calls at each position for which an 'N' was called; if a sequencer is unable to make a base call with sufficient confidence, it will normally substitute an N. **Panel d)** shows the mean quality scores across each base position in the read. **Panel e)** is a representation of the per sequence GC content for all sequenced individuals and **Panel f)** shows the number of reads which have average quality scores.

The four different bases (A,C,G,T) were not evenly distributed across the first 14 base pairs (Figure 2.9h). This is commonly observed across NGS data, and is an artefact of



Sample Name	% Dups	% GC	Length	M Seqs	% Dups	% GC	Length	M Seqs
Lbienne_dor5_GS1	46.20%	44%	121	1.7	47.50%	43%	120	1.7
Lbienne_dor5_GS2	60.10%	43%	120	8.9	61.50%	43%	119	8.9
Lbienne_dor5_GS3	43.70%	44%	118	0.9	45.90%	43%	117	0.9
Lbienne_lla10_GS1	61.90%	43%	119	6.9	63.10%	42%	118	6.9
Lbienne_lla10_GS2	60.60%	42%	119	4.3	62.30%	41%	118	4.3
Lbienne_lla10_GS3	60.50%	44%	119	3.1	61.00%	43%	118	3.1
Lcampanulatum_borau4_T_GS1	40.00%	48%	119	1.2	41.10%	48%	118	1.2
Lcampanulatum_borau4_T_GS2	43.50%	46%	121	2.3	44.60%	45%	120	2.3
Lcampanulatum_borau4_T_GS4	41.50%	47%	119	2.2	42.60%	46%	118	2.2
Lcampanulatum_borau5_P_GS1	49.30%	47%	119	5.2	50.00%	46%	118	5.2
Lcampanulatum_borau5_P_GS2	38.90%	48%	118	0.9	41.20%	47%	117	0.9
Lcatharticum_GS1	92.00%	49%	115	13.8	91.50%	49%	114	13.8
Lcatharticum_GS2	84.70%	47%	117	5.3	84.50%	46%	116	5.3
Lcatharticum_GS3	86.20%	51%	114	6.8	85.80%	51%	113	6.8
Lnarbonense_oroel2_T_GS1	57.20%	47%	119	2.3	59.50%	46%	118	2.3
Lnarbonense_oroel2_T_GS2	55.80%	45%	121	3	57.60%	45%	120	3
Lnarbonense_oroel2_T_GS3	52.10%	44%	121	1.8	54.30%	43%	120	1.8
Lnarbonense_oroel2_T_GS4	55.20%	47%	121	4.6	56.60%	46%	120	4.6
Lnarbonense_oroel3_P_GS1	57.90%	44%	119	6.2	59.30%	43%	117	6.2
Lnarbonense_oroel3_P_GS2_AR006	53.00%	44%	120	1.7	55.70%	43%	119	1.7
Lnarbonense_oroel3_P_GS2_AR007	60.80%	43%	122	1.7	63.70%	40%	121	1.7
Lnarbonense_oroel3_P_GS3	45.60%	43%	122	0.6	50.20%	41%	120	0.6
Lnarbonense_oroel3_P_GS4	67.10%	50%	117	3.7	67.70%	49%	116	3.7
Lsetaceum_GS1	59.50%	47%	119	2	61.00%	47%	117	2
Lsetaceum_GS2	52.70%	49%	118	1.7	54.20%	48%	116	1.7
Lstrictum_LZS7b_GS1	55.20%	46%	120	4.4	56.40%	45%	119	4.4
Lstrictum_LZS7b_GS3	34.50%	47%	120	3	36.00%	47%	119	3
Lsuffruticosum_oroel2_T_GS1	51.20%	48%	120	1.3	53.40%	48%	119	1.3
Lsuffruticosum_oroel2_T_GS2	58.10%	44%	122	2.9	60.20%	42%	120	2.9
Lsuffruticosum_oroel2_T_GS4	35.70%	47%	120	1.6	37.40%	46%	119	1.6
Lsuffruticosum_oroel3_P_GS1	49.20%	50%	119	0.4	53.20%	49%	118	0.4
Lsuffruticosum_oroel3_P_GS2	77.50%	41%	123	3.8	78.20%	39%	122	3.8
Lsuffruticosum_oroel3_P_GS4	20.40%	44%	124	0.1	26.10%	39%	122	0.1
Lsuffruticosum_oroel3_P_GS4_S45	54.90%	45%	116	0.2	56.60%	44%	115	0.2
Ltenuifolium_aigues2_GS1	20.70%	46%	123	0.1	27.30%	40%	122	0.1
Ltenuifolium_aigues2_GS2	34.60%	43%	121	0.3	37.30%	43%	120	0.3
Ltenuifolium_aigues2_GS2_S43	93.90%	48%	116	10.7	93.50%	48%	115	10.7
Ltenuifolium_aigues2_GS2_S44	93.90%	49%	115	12.6	93.30%	49%	115	12.6
Lusitatissimum_mar4_GS1	60.60%	42%	120	1.2	62.50%	41%	119	1.2
Lusitatissimum_mar4_GS2	66.00%	43%	121	4.2	67.40%	41%	120	4.2
Lusitatissimum_mar4_GS3	34.50%	47%	119	0.7	36.60%	46%	118	0.7
Lviscosum_oroel3_P_GS1	38.10%	46%	121	0.6	40.40%	45%	120	0.6
Lviscosum_oroel3_P_GS4	59.30%	43%	123	1.1	64.20%	39%	122	1.1
Lviscosum_x_T_GS1	48.20%	44%	122	3.3	49.50%	43%	121	3.3
Lviscosum_x_T_GS4	90.80%	50%	113	9.3	90.10%	50%	112	9.3

Figure 2.12: Summary statistics produced using MultiQC (Ewels et al., 2016) for post-trim results. A heat map is used to display the highest percentage duplications with the most highly duplicated sequences depicted in red. The left hand side of the table shows the November 2016 sequencing run and the right hand side of the table shows the August 2017 sequencing run.

random hexamer priming during the cDNA synthesis step (Hansen et al., 2010). One pipeline was attempted where the first 9 bases were removed from all sequences in an attempt to improve the quality of the data. The outcome is depicted below in Figure 2.13. A slight reduction in the variabilities each base at the first 9 positions is observed, however, the residual variability indicates that not all starting k-mers were removed.

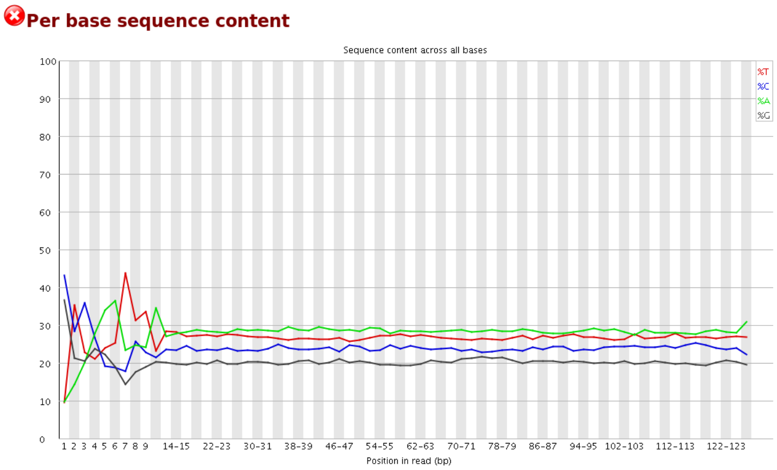


Figure 2.13: Post-trim FastQC data depicting the per base sequence content after the removal of the first 9 bases.

2.2.3 Mapping

After the reads were filtered from the raw cDNA sequence reads, the shorter sequenced fragments were mapped to a reference transcriptome. This processing step is required to determine the location of each fragment on the reference (Figure 2.14). The unpublished *L. tenue* transcriptome was chosen over the published *L. usitatissimum* transcriptome for the mapping procedure, primarily because the species is heterostylous. The candidate genes are expected to be unique to the thrum morph in heterostylous species, and therefore, it is likely that no identification would be possible in a homostylous species. Attempts were also made to utilise an annotated transcriptome of *L. grandiflorum*, as was used in (e.g. Ushijima et al., 2012), however, a transcriptome was not readily available in FASTA format. Mapping to an *L. tenue* reference transcriptome was performed using STAR (Dobin et al., 2013). The maximum mapping was approximately 58% and the minimum mapping just 4% (Table 2.5). These lower mapping values were in accordance with expectations for cross-species maps. A series of settings for the STAR program were investigated in order to optimise mapping by maximising mapping percentages without jeopardising quality. When reads are mapped onto a transcriptome, it is expected that slightly lower mapping percentages will be achieved. These mappings were

Table 2.4: Initial sequencing depth and the number of reads dropped during the Trimmomatic trimming process.

Sequence	Input Read Pairs	Both Surviving	Forward Only Surviving	Reverse Only Surviving	Dropped
L. bienne dor GS1	1842001	1516349 (82.32%)	175472 (9.53%)	49154 (2.67%)	101026 (5.48%)
L. bienne dor GS2	9334607	7987809 (85.57%)	676069 (7.24%)	245084 (2.63%)	425645 (5.48%)
L. bienne dor GS3	951671	793412 (83.37%)	86674 86674 (9.11%)	23995 (2.52%)	47590 (5.00%)
L. bienne lla GS1	7295697	6292127 (86.24%)	519513 (7.12%)	177395 (2.43%)	306662 (4.20%)
L. bienne lla GS2	4568781	3835453 (83.95%)	374706 (8.20%)	123491 (2.70%)	235131 (5.15%)
L. bienne lla GS3	3272470	2771675 (84.70%)	266847 (8.15%)	78436 (2.40%)	155512 (4.75%)
L. campanulatum T GS1	1287714	1082710 (84.08%)	105690 (8.21%)	33679 (2.62%)	65635 (5.10%)
L. campanulatum T GS2	2466647	2083960 (84.49%)	211621 (8.58%)	56262 (2.28%)	114804 (4.65%
L. campanulatum T GS4	2380759	1968072 (82.67%)	218477 (9.18%)	61252 (2.57%)	132958 (5.58%)
L. campanulatum P GS1	5521717	4725591 (85.58%)	436191 (7.90%)	121965 (2.21%)	237970 (4.31%)
L. campanulatum P GS2	942989	777724 (82.47%)	86933 (9.22%)	25925 (2.75%)	52407 (5.56%)
L. catharticum GS1	14647241	12128976 (82.81%)	1230080 (8.40%)	435820 (2.98%)	852365 (5.82%)
L. catharticum GS2	6008735	4484910 (74.64%)	816776 (13.59%)	207541 (3.45%)	499508 (8.31%)
L. catharticum GS3	7310473	5918814 (80.96%)	675234 (9.24%)	218904 (2.99%)	497521 (6.81%)
L. narbonense T GS1	2458313	2026076 (82.42%)	226424 (9.21%)	70937 (2.89%)	134876 (5.49%)
L. narbonense T GS2	3179385	2683803 (84.41%)	258508 (8.13%)	78063 (2.46%)	159011 (5.00%)
L. narbonense T GS3	1974287	1632031 (82.66%)	175835 (8.91%)	55633 (2.82%)	110788 (5.61%)
L. narbonense T GS4	4913341	4091081 (83.26%)	432990 (8.81%)	130065 (2.65%)	259205 (5.28%)
L. narbonense P GS1	6615344	5425210 (82.01%)	617284 (9.33%)	(2.82%)	386545 (5.84%)

Table 2.4

Sequence	**Input Read Pairs**	**Both Surviving**	**Forward Only Surviving**	**Reverse Only Surviving**	**Dropped**
L. narbonense P GS2	3696838	3002683 (81.22%)	374970 (10.14%)	97750 (2.65%)	221435 (5.99%)
L. narbonense P GS3	706092	533229 (75.52%)	100454 (14.23%)	20569 (2.91%)	51840 (7.34%)
L. narbonense P GS4	3887170	3277324 (84.31%)	309748 (7.97%)	98672 (2.54%)	201426 (5.18%)
L. setaceum GS1	2330854	1758546 (75.45%)	345362 (14.82%)	69136 (2.97%)	157810 (6.77%)
L. setaceum GS2	1854720	1449505 (78.15%)	236835 (12.77%)	56157 (3.03%)	112223 (6.05%)
L. strictum GS1	4743459	3797409 (80.06%)	471887 (9.95%)	151967 (3.20%)	322196 (6.79%)
L. strictum GS3	3201215	2618585 (81.80%)	302811 (9.46%)	92713 (2.90%)	187106 (5.84%)
L. suffruticosum T GS1	1404660	1151881 (82.00%)	129527 (9.22%)	39980 (2.85%)	83272 (5.93%)
L. suffruticosum T GS2	3183376	2539000 (79.76%)	337167 (10.59%)	94346 (2.96%)	212863 (6.69%)
L. suffruticosum GS4	1730206	1420686 (82.11%)	158940 (9.19%)	50722 (2.93%)	99858 (5.77%)
L. suffruticosum P GS1	456181	358675 (78.63%)	52112 (11.42%)	12407 (2.72%)	32987 (7.23%)
L. suffruticosum P GS2	4105048	3429023 (83.53%)	362949 (8.84%)	97991 (2.39%)	215085 (5.24%)
L. suffruticosum P GS4	253082	196341 (77.58%)	28689 (11.34%)	6870 (2.71%)	21182 (8.37%)
L. tenuifolium GS1	79949	58610 (73.31%)	8998 (11.25%)	2063 (2.58%)	10278 (12.86%)
L. tenuifolium GS2	24637438	20828494 (84.54%)	1863392 (7.56%)	695108 (2.82%)	1250444 (5.08%)
L. usitatissimum GS1	1262088	1053965 (83.51%)	103821 (8.23%)	35662 (2.83%)	68640 (5.44%)
L. usitatissimum GS2	4486060	3656931 (81.52%)	439101 (9.79%)	120737 (2.69%)	269291 (6.00%)
L.usitatissimum GS3	800631	665566 (83.13%)	68142 (8.51%)	22720 (2.84%)	44203 (5.52%)
L. viscosum T GS1	3641271	2857158 (78.47%)	407635 (11.19%)	116197 (3.19%)	260281 (7.15%)
L. viscosum T GS4	9877760	8138395 (82.39%)	854200 (8.65%)	286195 (2.90%)	598970 (6.06%)
L. viscosum P GS1	658517	524246 (79.61%)	68333 (10.38%)	21710 (3.30%)	44228 (6.72%)
L. viscosum P GS4	1251919	1002829 (80.10%)	138811 (11.09%)	31843 (2.54%)	78436 (6.27%)

performed across species, and hence, it was anticipated that there would be much lower achieved mapping percentages than maps to a transcriptome of the same species.

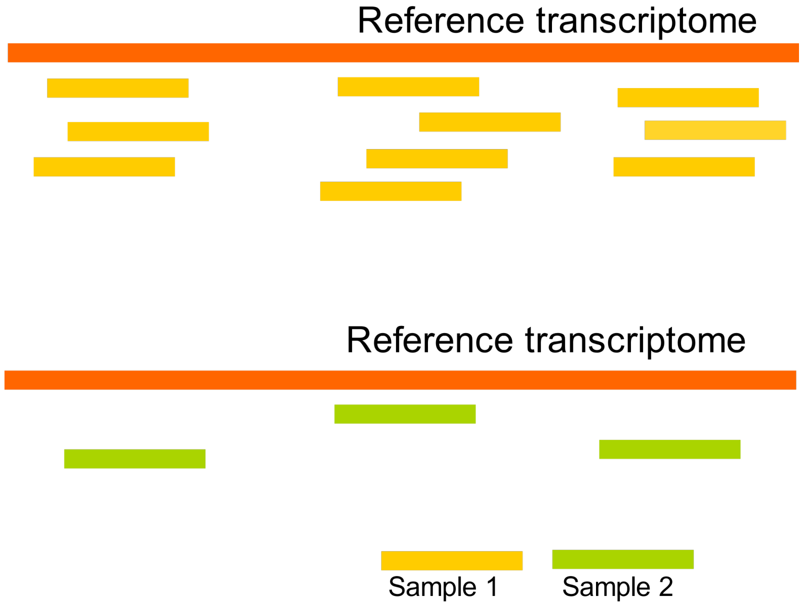


Figure 2.14: Illumina sequencing produces millions of short reads. These can be aligned to a reference transcriptome, in this case *L. tenue*. Reads can be used to determine the degree of expression of a particular RNA transcript. Here, Sample 1 is more highly expressed than Sample 2.

STAR

The ultrafast alignment algorithm 'Spliced Transcripts Alignment to a Reference' (STAR) was utilised (Dobin et al., 2013). A series of settings were tested to achieve the best unique mapping quality, without relaxing the parameters to too great an extent. Unique, in this context, refers to reads mapping to just one place in the reference transcriptome. The mismatch parameters were altered in a series of pilot studies, as using default settings the highest unique mapping percentages obtained were only 19%. In order to increase the sensitivity of novel junction discovery (the boundary between two exons), STAR was run in the 2-pass mode. This enables the detection of more splice reads mapping to novel junctions. The settings –outFilterScoreMinOverLread and –outFilterMatchNminOverLread were both set to 0.2. The default setting for these parameters is 0.66. For outFilterMatchNminOverLread, the output was filtered such that the alignment would only be output if the number of matched bases was equal to or exceeded 0.2, normalised to the read length. For outFilterScoreMinOverLread, the alignment was only output if its score was higher than 0.2, normalised to read length. By relaxing the parameters, there was an increased likelihood of mismatching, however, it also allows for higher read map success. Mapping percentages were further improved by the specification of a .gtf file. Gene transfer format (GTF) is a file format that holds information about gene structures. It is a tab-delimited text file format based on the general feature format (GFF) (Ensembl, 2017). The GTF file was created by Ali Foroozani and was based on mapping to the *Arabidopsis* TAIR database. The overlap parameters, –alignEndsProtrude and –alignIntronMin were also altered from the default. This resulted in a lower percentage mapping, 19.76% for *L. bienne* dor5, but slightly longer reads and a smaller percentage mapped to too many loci. Ultimately, however, it was decided to maintain the default settings for these overlap parameters.

Mapping the sequences using STAR compounded the realisation that some of the original libraries had not been sequenced fully. The mapping percentages described in Table 2.5 show that *L. suffruticosum* and *L. tenuifolium* were particularly low. An indicative minimum read threshold would be approximately 5 million reads. As described in Section 2.2.1, a second sequencing run was undertaken in August 2017 to provide more sequence reads for all samples. In order to maximise the read mapping data, the gzip compressed Fasta files from the August 2017 and November 2016 sequencing runs were combined using the command in Figure 2.15.

```
for Aug in *P.fastq.gz;
do
Nov=${Aug/P.fastq.gz/P_Nov16.fastq.gz}

zcat $Aug $Nov | gzip -9 > Joined_$Aug
done
```

Figure 2.15: A script to merge the trimmed output of the November 2016 and June 2017 sequence files, ready for input into STAR.

The obtained sequence coverage is recorded in Table 2.5, and was obtained from the STAR log files. Only the paired output from Trimmomatic is passed to the STAR program, and consequently this estimation of the sequence coverage is lower than the total output from the Illumina high-throughput sequencer. As discussed further in Section 2.4.2, the use of alternative pipelines such as Salmon (Patro et al., 2017) or kallisto (Bray et al., 2016) might improve this coverage. Aligners can also differ on how they handle reads that map equally well to several locations. Most aligners either discard multimaps (Langmead et al., 2009), allocate them randomly (Li et al., 2008) or allocate them based on an estimate of the local coverage (Oshlack, Robinson, & Young, 2010). The default behaviour of STAR is to report all multimapping regions. In the STAR output, all alignments except one are marked as secondary alignments. The unmarked alignment is either the highest scoring (i.e. the best) or it is randomly selected from alignments of equal quality.

The disadvantage of undertaking cross-species mapping, as in this project was the relatively low obtained mapping percentages. The final achieved mapping percentages for both the November 2016 and August 2017 sequencing runs are recorded in Table 2.5. The average unique mapping percentage was approximately 35%. The *L. narbonense* maps were noticably improved following re-sequencing as were *L. suffruticosum*. However, there was no clear trend of improvement in mapping percentage following the August 2017 re-sequencing. This is unsurprising since the same libraries were utilised. It is not unusual that only 40-50% of the data generated from RNA-Seq are mappable between species (Blow, 2009) and this sentiment is borne out in the data. There is a possibility that these unmapped sequences are biologically relevant, for instance unannotated genes or antisense transcripts. Alternatively, these unmapped fragments may simply be artefacts of the sequencing process. The use of 'paired-end' reads, where sequencing information is obtained from both ends of a DNA fragment helps to significantly improve the quality of the map. Significant variation in the number of input reads for each sequence was also observed. This variation may be partially explained by the quality of the library preparation (Becker, 2015), specifically difficulties in accurate individual sample library quantification prior to pooling for sequencing.

Often the most common reason for a low return of reads are primer dimers in the samples or adapter dimers, poor size distribution or improper quantification. There was a high percentage of unmapped reads in some of the samples and also a high mismatch error percentage up to 3.52% in certain cases. For good libraries, the mismatch error per base would be expected to be in the region of 0.5-0.8% (Dobin, 2013). This may have been a result of contamination or could potentially mark an issue with trimming quality. Some of the main sources of contamination include contaminated starting material from the wild.

The number of reads mapped to multiple loci is high in the final map of every sequence. The mapped sequences were analysed using BBduk,part of the BBTools suite of packages (Bushnell, 2015a), to test for the possibility that these high "aligned to multiple loci" reads were caused by excess ribosomal RNA (rRNA) that ideally should have been depleted from the samples during library prep. The outcome of the BBduk quality control phase is shown in Table 2.6. BbDuk was able to effectively remove kmers, more

Table 2.5: Results of the STAR Mapping process

			August 2017 Sequencing			November 2016 Sequencing		
Species	**Morph Type**	**Growth Stage**	**% Uniquely Mapped Reads**	**% Reads Mapped to Multiple Loci**	**Input Reads**	**% Uniquely Mapped Reads**	**% Reads Mapped to Multiple Loci**	**Input Reads**
L. bienne dor5	Homostylous	GS1	36.54	27.69	1516349	41.44	31.41	5235109
Lbienne dor5	Homostylous	GS2	37.21	27.82	7987809	38.46	28.7	26227392
L. bienne dor5	Homostylous	GS3	34.62	25.78	793412	40.69	31.07	2165451
L. bienne lla10	Homostylous	GS1	40.54	31.00	6292127	42.22	32.22	21056239
L. bienne lla10	Homostylous	GS2	29.30	24.52	3835453	41.29	31.14	6811883
L. bienne lla10	Homostylous	GS3	40.58	28.96	2771675	42.49	26.8	881089
L. campanulatum	Short Style	GS1	43.12	32.82	1082710	46.47	35.58	3376227
L. campanulatum	Short Style	GS2	40.90	27.38	2083960	48.97	33.52	16670782
L. campanulatum	Short Style	GS4	40.63	30.67	1968072	48.8	32.21	3183456
L. campanulatum	Long Style	GS1	46.17	31.63	4725591	51.3	31.21	2059469
L. campanulatum	Long Style	GS2	44.70	27.29	777724	46.1	34.9	5519527
L. catharticum	Homostylous	GS1	32.54	61.89	12128976	34.53	64.77	37357632
L. catharticum	Homostylous	GS2	20.37	60.00	4484910	23.73	75.33	10988444
L. catharticum	Homostylous	GS3	58.11	33.88	3439608	61.77	36.04	17790122
L. narbonense	Short Style	GS1	45.04	23.80	2026076	49.33	26.09	2886478
L. narbonense	Short Style	GS2	41.14	24.84	2683803	33.70	29.33	3947635
L. narbonense	Short Style	GS3	32.25	20.04	526382	43.96	26.65	8231020

Table 2.5

			August 2017 Sequencing			November 2016 Sequencing		
Species	**Morph Type**	**Growth Stage**	**% Uniquely Mapped Reads**	**% Mapped to Multiple Loci**	**Input Reads**	**% Uniquely Mapped Reads**	**% Reads Mapped to Multiple Loci**	**Input Reads**
L. narbonense	Short Style	GS4	32.83	21.63	1343003	36.42	28.70	2850261
L. narbonense	Long Style	GS1	26.58	22.74	5425210	40.4	25.41	650564
L. narbonense	Long Style	GS2	39.72	32.63	2832827	38.07	20.99	4662260
L. narbonense	Long Style	GS3	22.93	11.07	533229	35.06	23.26	12487760
L. narbonense	Long Style	GS4	53.34	32.21	3277324	55.72	33.74	10382387
L. setaceum	Homostylous	GS1	53.22	26.89	1758546	60.7	30.54	3191791
L. setaceum	Homostylous	GS2	48.99	31.72	1449505	53.75	37.00	4720469
L. strictum	Homostylous	GS1	47.61	32.12	1807957	51.35	36.56	10649234
L. strictum	Homostylous	GS3	43.27	32.34	2618585	45.86	34.43	8159885
L. suffruticosum	Short Style	GS1	41.84	40.26	1151881	43.23	42.79	3796652
L. suffruticosum	Short Style	GS2	23.99	16.16	2539000	53.7	35.37	908867
L. suffruticosum	Short Style	GS4	46.67	30.17	1420686	49.21	31.08	2849844
L. suffruticosum	Long Style	GS1	41.23	26.62	358675	1.07	1.63	27780
L. suffruticosum	Long Style	GS2	9.15	13.51	3821967	48.47	31.56	4541597
L. suffruticosum	Long Style	GS4	17.05	35.52	196341	14.83	40.44	31104

Table 2.5

			August 2017 Sequencing			**November 2016 Sequencing**		
Species	**Morph Type**	**Growth Stage**	**% Uniquely Mapped Reads**	**% Reads Mapped to Multiple Loci**	**Input Reads**	**% Uniquely Mapped Reads**	**% Reads Mapped to Multiple Loci**	**Input Reads**
L. tenuifolium	Homostylous	GS1	8.05	5.73	58610	10.72	9.11	2920
L. tenuifolium	Homostylous	GS2	28.20	68.26	14121225	28.29	70.73	66172536
L. usitatissimum	Homostylous	GS1	55.35	22.57	1053965	57.62	24.11	3043814
L. usitatissimum	Homostylous	GS2	41.25	40.65	3656931	55.47	23.11	7278795
L. usitatissimum	Homostylous	GS3	42.27	28.89	665566	47.54	32.01	1852204
L. viscosum	Long Style	GS1	57.27	25.89	524246	68.12	28.86	972961
L. viscosum	Long Style	GS4	9.74	7.89	1002829	65.68	31.06	5375300
L. viscosum	Short Style	GS1	45.68	23.08	2857158	15.45	43.22	28241
L. viscosum	Short Style	GS4	37.43	57.32	8138395	39.2	60.1	15576252

precisely rRNA, from the paired trimmed sequence reads. Although removing rRNA is not strictly necessary, it does speed up the mapping stage. Additionally, having an indication of the proportion of rRNA in the sequence allows judgements of the overall sequence quality to be made. One of the most important steps involved in library preparation is the RNA-extraction protocol used to remove the highly abundant rRNA, which usually constitutes over 90% of the total cellular RNA (Conesa et al., 2016). A correlation certainly exists between the very lowest mapping percentages, and very high proportions of rRNA amongst the tested sequences. One prime example is *L. tenuifolium* GS1, of which only 5.73% mapped to the reference *L. tenue* transcriptome, yet the percentage of rRNA was over 40%. Typically, RNA-Seq libraries are prepared from total RNA using poly(A) enrichment of the mRNA to remove rRNA (Zhao et al., 2014). Thus, the high levels of mRNA may have resulted from problems with the library preparation. Equally, with the example of *L. catharticum* GS1, there were a very large number of repetitive sequences, none of which mapped to the *L. tenue* reference transcriptome, and 90% of which were removed by BBduk. Ideally, *L. catharticum* should be resequenced so that more meaningful results can be drawn from this homostylous species.

Table 2.6 shows the percentage of rRNA contaminant removed from each of the sequences, prior to a second mapping phase. The improvement of *L. narbonense* during the second library sequencing in August is also evident from Tables 2.5 and 2.6. The percentage rRNA removed is still high, however, the number of read counts mapped is overwhelmingly improved.

Table 2.6: Percentage of rRNA contaminant removed using BBDuk from each sequence, prior to mapping

Species	**Morph Type**	**Developmental Stage**	**% Contaminants Removed**
L. bienne dor5	Homostylous	GS1	9.83
L. bienne dor5	Homostylous	GS2	4.86
L. bienne dor5	Homostylous	GS3	10.57
L. bienne lla10	Homostylous	GS1	6.59
L. bienne lla10	Homostylous	GS2	8.71
L. bienne lla10	Homostylous	GS3	9.99
L. campanulatum	Short Style	GS1	9.95
L. campanulatum	Short Style	GS2	12.47
L. campanulatum	Short Style	GS4	10.11
L. campanulatum	Long Style	GS1	12.19
L. campanulatum	Long Style	GS2	15.67

Table 2.6: Percentage of rRNA contaminant removed using BBDuk from each sequence, prior to mapping

Species	Morph Type	Developmental Stage	% Contaminants Removed
L. catharticum	Homostylous	GS1	91.80
L. catharticum	Homostylous	GS2	75.59
L. catharticum	Homostylous	GS3	86.46
L. narbonense	Short Style	GS1	24.46
L. narbonense	Short Style	GS2	15.10
L. narbonense	Short Style	GS3	7.66
L. narbonense	Short Style	GS4	14.62
L. narbonense	Long Style	GS1	4.88
L. narbonense	Long Style	GS2	9.55
L. narbonense	Long Style	GS3	9.94
L. narbonense	Long Style	GS4	31.73
L. setaceum	Homostylous	GS1	29.11
L. setaceum	Homostylous	GS2	21.07
L. strictum	Homostylous	GS1	9.64
L. strictum	Homostylous	GS2	9.55
L. suffruticosum	Short Style	GS1	31.65
L. suffruticosum	Short Style	GS2	8.42
L. suffruticosum	Short Style	GS4	9.84
L. suffruticosum	Long Style	GS1	30.77
L. suffruticosum	Long Style	GS2	10.48
L. tenuifolium	Homostylous	GS1	3.94
L. tenuifolium	Homostylous	GS2	7.80
L. tenuifolium S43	Homostylous	GS2	95.86
L. usitatissimum	Homostylous	GS1	9.24
L. usitatissimum	Homostylous	GS2	8.19
L. usitatissimum	Homostylous	GS3	9.88
L. viscosum	Short Style	GS1	10.50
L. viscosum	Short Style	GS4	92.02
L. viscosum	Long Style	GS1	9.83
L. viscosum	Long Style	GS2	15.44
L. viscosum	Long Style	GS4	5.72

The number of reads unmapped because they were too short was 82.33% for the long-styled *L. viscosum* morph at GS4. Common troubleshooting issues, which may have caused the obtained result include the two paired-end input FASTQ file mates being out of order; the mates are not found at the same line in two of the files.This leads to many improperly mapped read pairs, which are marked as “too short" by STAR (Frech, 2016) . *L. tenuifolium* GS1 suffers from the same problem and in addition has a very low number of input sequence reads. Consequently, the results for *L. tenuifolium* are not sufficiently robust as to be taken into full consideration in the differential expression analysis. Alternative mapping programs, including HISAT2 (Kim, Langmead, & Salzberg, 2015) were audited as mapping tools. Potential benefits of using HISAT2 include that unpaired read data is not discarded during HISAT2 analysis in the same way that it is during STAR analysis. It is consequently possible that valuable read count data is lost during STAR mapping that could be salvaged using HISAT2. However, STAR (Dobin et al., 2013) was ultimately selected, based on its rapid reported mapping speed and the smaller amount of RAM required to sequence. Consequently the required high performance computing time was minimised.

BUSCO

Benchmarking Universal Single-Copy Orthologs (BUSCO) sets were used as a quantitative assessment of transcriptome sequence quality and completeness (Simão et al., 2015). BUSCO are a set of genes that are commonly used to measure genome or transcriptome assembly and annotation completeness. The BUSCO set for eukaryotes comprises 429 well-conserved, single-copy genes.

The *L. tenue* transcriptome assembly, from which the ‘*S-locus*’ candidate gene primers were designed, had 280 complete BUSCOs, 116 complete and single-copy BUSCOs, 164 complete and duplicated BUSCOs, 18 fragmented BUSCOs and 5 missing BUSCOs. This is represented as C:92.4%[S:38.3%,D:54.1%], F:5.9%, M:1.7%, n:303. It is thus evident that the *L. tenue* transcriptome represented an adequate transcriptome assembly for attempting to characterise the *S-locus* across wild *Linum* species.

BUSCO data for individual species indicated that the sequence quality was good, and confirmed the data from FastQC. In order to achieve this BUSCO measure, Trinity (Grabherr et al., 2011) assemblies of the wild-sample RNA-Seq data were created using default parameters. The BUSCO analysis provided a more robust measure of sequence completeness.

As can be seen from Table 2.7, there is considerable variation in sequence completeness, and by extension quality, between the various sequences analysed in this project. Some of these issues were addressed by re-sequencing the libraries in the August 2017

Table 2.7: BUSCO gene coverage across a series of sequences from the November 2016 sequence run, assembled using Trinity.

Sequence	**BUSCO Data**
L. bienne dor5 GS1	C:78.6%[S:23.8%, D:54.8%], F:17.8%, M:3.6%, n=303
L. campanulatum borau4 GS1 T	C:60.7%[S:35.6%,D:25.1%],F:30.7%,M:8.6%,n:303
L. campanulatum borau4 GS1 P	C:96.4%[S:24.8%,D:71.6%],F:2.6%,M:1.0%,n:303
L. narbonense oroel2 GS2 T	C:84.8%[S:40.9%,D:43.9%],F:12.5%,M:2.7%,n:303
L. narbonenese oroel 3 GS2 P	C:9.5%[S:5.9%,D:3.6%],F:30.0%,M:60.5%,n:303
L. suffruticosum oroel 2 GS1 T	C:0.3%[S:0.3%,D:0.0%],F:0.7%,M:99.0%,n:303
L. suffruticosum oroel 3 GS1 P	C:69.9%[S:32.3%,D:37.6%],F:25.4%,M:4.7%,n:303

Sequence Run. The low results for the *L. narbonense* and *L. tenue* sequence runs may have been a factor of encountered mapping difficulties. In fact, by cross-referencing the analysed read-counts in Table 2.5, the correlation between poor mapping, low BUSCO coverage and low read count becomes increasingly apparent.

2.2.4 Post-Processing

SAMTools

SAMTools is a set of utilities capable of manipulating alignments in BAM format . SAMTools imports from and exports to the SAM (Sequence Alignment/Map) format , in addition to performing sorting, merging and indexing operations (Li et al., 2009).

The mapping step creates an output file in SAM format, containing information on both aligned and non-aligned reads (Li et al., 2009). Following RNA-Seq read mapHTSeqping, the data was converted into a suitable input format for the subsequent alignment steps; a BAM binary alignment file, using SAMTools. SAMTools was also used to sort the alignment by mapped co-ordinate. Sorting the contigs by co-ordinate facilitated their input into the raw read sequence counter, HTSeq-count. The BAM file format produces smaller files, facilitating file storage.

```
for STAR_MAP in *;
do

cd "$STAR_MAP" &&\
echo "creating BAM file ###############################################" && \
/ddn/data/ghwq12/Tools/samtools-1.3.1/samtools view -b -o"$STAR_MAP".bam \
Aligned.out.sam

echo "sorting by co-ordinate ###########################################" && \
/ddn/data/ghwq12/Tools/samtools-1.3.1/samtools sort -o\
"$STAR_MAP".sort.bam "$STAR_MAP".bam

echo "sorting by name  #################################################" && \
/ddn/data/ghwq12/Tools/samtools-1.3.1/samtools sort -n -o\
"$STAR_MAP".namesort.bam "$STAR_MAP".bam

echo "indexing ########################################################" && \
/ddn/data/ghwq12/Tools/samtools-1.3.1/samtools index "$STAR_MAP".sort.bam

echo "summary report #################################################" && \
/ddn/data/ghwq12/Tools/samtools-1.3.1/samtools flagstat "$STAR_MAP".sort.bam\
> $STAR_MAP.sorted.flagstat
cd ..

done
```

Figure 2.16: Minimal example of a script to run various functions within the SAMTools suite of packages

Figure 2.16 shows an example script used firstly to create .bam files, and subsequently to sort them both by co-ordinate and by name. The hash marks were for the author's reference and were used to identify output whilst the program was in progress.

The BAM file was indexed. Indexing aims to rapidly retrieve alignments overlapping a specific region without the requirement to search the entire alignment. The BAM must be sorted by reference ID and the left-most co-ordinate before attempting indexing. One of the functions of SAMTools is that it is capable of producing a 'flagstat' output. Flagstat is a convenience function, which collates useful data regarding the sequences. Example flagstat output is depicted below in Figure 2.17.

```
File: Lcampanulatum_borau4_T_GS1_R1_1P.fastq.gz_STAR_MAP.sorted.flagstat


4558154 + 0 in total (QC-passed reads + QC-failed reads)
2340935 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
4558154 + 0 mapped (100.00% : N/A)
2217219 + 0 paired in sequencing
1113555 + 0 read1
1103664 + 0 read2
2080456 + 0 properly paired (93.83% : N/A)
2080456 + 0 with itself and mate mapped
136763 + 0 singletons (6.17% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 2.17: Flagstat output for the *L. campanulatum* thrum morph sequence at GS1.

HTSeq-Count

Typically, after mapping RNA-Seq reads to a reference transcriptome, the number of reads mapping to a particular contig is measured. For RNA-Seq data, this obtained read count has been found to have an approximately linear relationship to the abundance of target transcript (Mortazavi et al., 2008), which therefore makes it a useful basic assessment of gene expression.

HTSeq (version 0.6.1p1) (Anders et al., 2015) is a useful set of packages for analysing high-throughput sequencing data using Python (Python Software Foundation, https://www.python.org/). HTSeq-count is a script integrated into HTSeq; it is designed to count how many sequence reads map to a particular feature on the reference (Anders et al., 2015). Several settings for HTSeq-count were investigated. Figure 2.18 gives example HTSeq-counts code.Initial trials demonstrated that altering the input parameters caused significant differences in the outputted raw read counts. It was discerned, following a period of trial and error, that htseq-counts was interpreting coordinate-sorted BAM files twice, and consequently the obtained read counts were double what in actual fact had been obtained. Evidence of this is shown in Table 2.8.

The settings used to obtain raw read counts for the DESeq2 differential expression analysis are depicted below. Sorting by name, rather than by co-ordinate was ultimately preferred. Examining the count data for the identified candidate genes will provide one method of analysing the likelihood of their involvement in the *'S-locus'*.

```
#!/bin/bash


#Script to run HTSeq-count on SAMtools Output
#Ellie Desmond (February 2017)


for file in *
do
echo "${file}"
python -m HTSeq.scripts.count \
-s reverse \
-r name \
-i transcript_id \
-t CDS \
-f bam \
"${file}" \
~/finalPre-draft.gtf\
> ~/Aug_2017_Sequence_Run/HTSeq-Counts/NovAug_Merged/"${file}".counts

done
```

Figure 2.18: Minimal example of a script to run HTSeq-count on .bam files, the output of SAMTools

Comparison HTSeq-count runs between .bam sorted which were sorted by their coordinate and those sorted by their name, in alphabetical order, were conducted. The result of this brief experiment is depicted in Table 2.8.From Table 2.8 it can be seen that the coordinate-sorted .bam files produced artificially higher read counts than the name-sorted files. Count data serves as a proxy for the magnitude of gene expression since transcripts of greater abundance in the cell should have more reads transcribed into RNA. However, count data is not a failsafe method of gene expression analysis, as the levels of gene expression may vary between samples or individuals according to the experimental design.

Table 2.8: Example data comparing the raw read counts of coordinate-sorted and name-sorted .bam files when analysed using HTSeq-count. Data is from *L. campanulatum* at GS2, based on the November 2016 Sequence Run. This data was not ultimately used in the analysis.

	L. campanulatum T		*L. campanulatum* P	
	Name-Sort	**Coordinate-Sort**	**Name-Sort**	**Coordinate-Sort**
Contig 1	9	15	2	2
Contig 100000	0	0	1	1
Contig 100001	1	1	0	0
Contig 10001	2	4	0	0
Contig 100014	1	1	0	0
Contig 100016	0	0	2	4
Contig 100026	24	47	1	1
Contig 100028	15	29	0	0

2.2.5 Differential Expression Analysis

Differential expression analysis was conducted in heterostylous species to compare long-styled and short-styled floral morphs across each growth stage 2.10. Additionally, analyses were conducted comparing homostylous species to heterostylous species at each growth stage. GS3 and GS4 were combined for one set of analyses to compensate for the reduced number of samples available. For heterostylous species, different analyses were conducted for each growth stage. For example, all heterostylous species at GS1 were compared in the same analysis. At the library preparation stage, only those RNA extrac-

Table 2.9: Input Comparisons for DESeq2

	DESeq2 Input Analyses
GS1	All heterostylous species
	All homostylous species
GS2	All heterostylous species
	All homostylous species
GS3-4	All heterostylous species
	All homostylous species
Heterostylous vs Homostylous	
All species together	

Table 2.10: The heterostylous comparison groups used in the DESeq2 analysis. Not all growth stages for all species were put forward for comparison.

GS1		**GS2**		**GS3**		**GS4**	
Thrum	Pin	Thrum	Pin	Thrum	Pin	Thrum	Pin
L. campanulatum	*L. campanulatum*	*L. campanulatum*	*L. campanulatum*	*L. narbonense*	*L. narbonense*	*L. campanulatum*	
L. narbonense	*L. narbonense*	*L. narbonense*	*L. narbonense*			*L. narbonense*	*L. narbonense*
L. suffruticosum	*L. suffruticosum*	*L. suffruticosum*	*L. suffruticosum*			*L. suffruticosum*	*L. suffruticosum*
L. viscosum	*L. viscosum*					*L. viscosum*	*L. viscosum*

Table 2.11: The homostylous comparison groups used in the DESeq2 analysis. Not all growth stages for all species were put forward for comparison.

GS1	**GS2**	**GS3**
L. bienne	*L. bienne*	*L. bienne*
L. catharticum	*L. catharticum*	
L. setaceum	*L. setaceum*	
L. strictum	*L. strictum*	
L. tenuifolium	*L. tenuifolium*	
L. usitatissimum	*L. usitatissimum*	*L. usitatissimum*

tions with the highest RNA concentrations at the quantification stage were sequenced (sequenced species at each growth stage are depicted in Table 2.10). Consequently, not all growth stages for all species were put forward for sequencing. Fortunately, the August 2017 sequencing run fixed issues where the sequencing had been inadequate for one of the two floral morphs at the growth stage. Therefore, more intra-species floral morph comparisons were possible with the full dataset.

DESeq2 (Love et al., 2014) was used to perform differential analysis. The R code used to perform the differential expression analysis can be found in Appendix A. Raw read count tables were created to estimate expression, i.e. how many raw molecules of mRNA were present in the samples and, prior to differential expression analysis, a visual inspection was performed. A BLAST search against the *L. tenue* transcriptome was performed to identify candidate genes. The top BLAST hit was extracted using the short AccessionGrab.py python program (Appendix B) and the Contigs identified through this as being associated with candidate genes were extracted from the counts table. fastaGrab.py (Foroozani, personal communication; Appendix B) was then used to find and isolate in the annotated *L.tenue* transcriptome, the sequence at the contig that was the top BLAST hit. The transcripts showing the most significant differential expression, based on normalised p-value were noted.

Although pre-filtration of low-count genes was not technically necessary, rows with zero reads were removed in order to reduce the memory size of the dds data object and thus to increase the speed of analysis. DESeq2 and Edge-R only accept un-normalised read counts, therefore, methods of normalising read count data, such as Reads Per Kilobase of transcript per Million mapped reads (RPKM) , Fragments Per Kilobase of transcript per Million mapped reads (FPKM; used by Cufflinks (Trapnell et al., 2012)) or Transcripts Per Kilobase Million (TPM) were not considered as a part of this analysis. Normalising read count data using RPKM and FPKM has fallen out of favour in recent analyses. The model "~condition + species + condition:species" was used. The interaction term is for testing genes which respond differently at each floral morph, across species. A minimal example of the DESeq2 code used can be found in Appendix A. For the comparison where all heterostylous species were included, the experimental design was: condition + species + growthstage.

DESeq2 returned a P-value determined by Wald-statistics and an adjusted P-value to correct for multiple comparisons tested using the Benjamini-Hochberg method to determine the false discovery rate. DESeq2 automatically determines the reference level for factors based on alphabetical order. The level to compare against is usually taken as the control group . In this experiment, the long-syled pin morphs, where the expression of candidate genes was not anticipated, were considered to be the control group. The results function performs independent filtering based on the mean of normalised counts for each gene. This optimises the number of genes which will have an adjusted p-value below a given cut-off. The aim was to be conservative in our observations.

To analyse the results, MA plots were created in addition to Principal component analysis was performed. MA plots display a log ratio (M) vs an average (A) to assist in the visualisation of differences between the long-styled morphs and the short-styled morphs.

Principal Component Analysis (PCA) was also undertaken to visualise overall differences in expression between species and morphs at different growth stages. A variance stabilising transformation, rld was performed on the dds object prior to the PCA.

2.3 Results

2.3.1 Raw Count Data

Raw count data allows for an estimation of expression by providing the number of reads that have mapped to each contig of the *L. tenue* transcriptome. It is worth considering the fact that a direct comparison of raw counts can be misleading, if one group was sequenced at a significantly greater depth than another unless the count data is normalised appropriately.

The first interesting result gleaned from the examination of read counts extracted from the raw count graphs was the fact that three candidate genes shared the same *L. tenue* contig (107032). Contig 107032 was the top Blast hit for *PveGLO2* (*Primula veris*), *S-ELF3* (*Fagopyrum esculentum*) and *GLO*^{T} (*Primula* species). This observation prompted comparison by multiple sequence alignment, shown below in Figure 2.20. Resultantly, The *GLO* genes are collectively referred to as *S-ELF3* in raw read count tables, as they all matched to the same Contig of *L. tenue*.

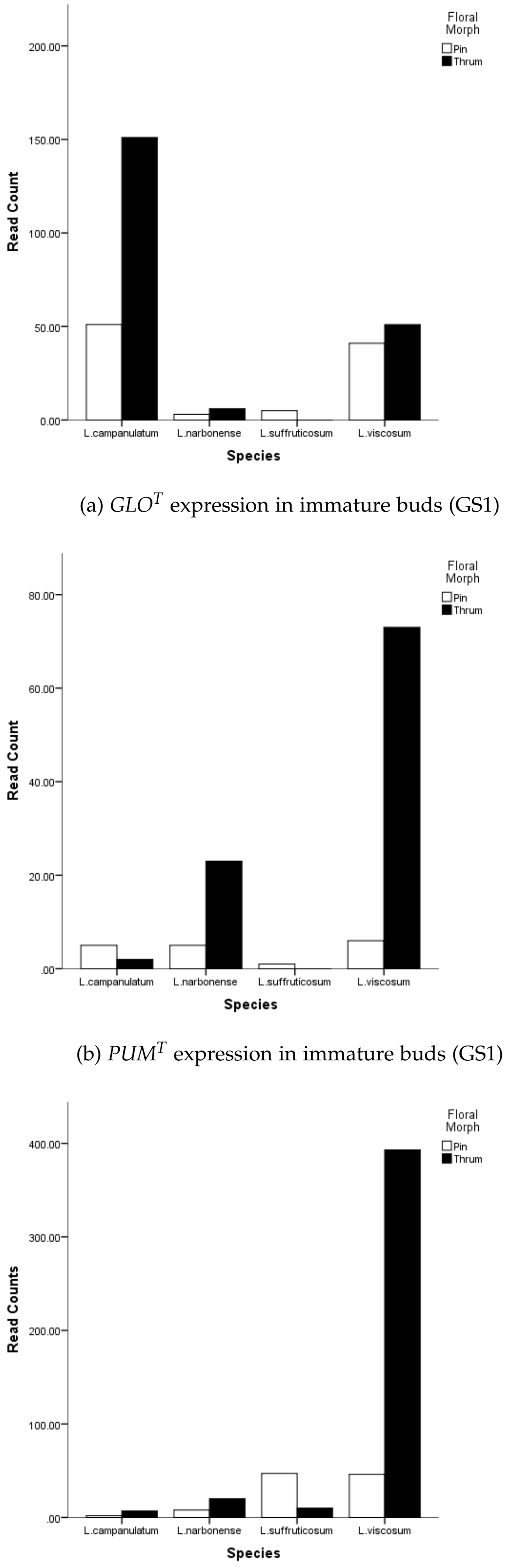
However, subsequent personal communication with Philip Gilmartin reveals the above conclusion to be incorrect. *S-ELF3* and *GLO*^{T} encode completely unrelated proteins of different function (Gilmartin 2018, unpublished data). Furthermore, upon closer inspection there is only 20% nucleotide sequence similarity across all regions being compared. This level of similarity is low and gaps in the sequence introduce frame-shifts into the sequence meaning that Contig 107032 could not encode *S-ELF3*. Phylogenetic analysis of the four strong candidates: *CYP*^{T}, *GLO*^{T}, *PUM*^{T} and *S-ELF3* at the amino acid level have been conducted to determine whether the above nucleotide sequences (especially *GLO*^{T}) code for proteins and to determine whether the contigs represent the thrum-specific genes, rather than one of many other closely-related gene family members.

In this study, the majority of the candidate genes did not show clear patterns of expression, according to expectations from other species. Although, present in the short-styled thrum morph of *L. campanulatum*, *GLO*^{T} was expressed to a lesser extent than GS1 pin morphs. This trend was also observed in the maturing buds (GS2) 2.19d, although the overall read counts were lower.

The expression of *PUM*^{T} was more variable. At GS3 and GS4, the open flower stage, *PUM*^{T} expression was absent in both *L. suffruticosum* and *L. viscosum* (Figure 2.19h). *PUM*^{T} expression was greater in the short morph of *L. narbonense*, although was not absent in the long morph. From Figure 2.19e it can be observed that expression is again absent in *L. suffruticosum*. Raw *PUM*^{T} read counts were, however, greater in the short-styled thrum morph of both *L. campanulatum* and *L. narbonense*, with no *PUM*^{T} reads recorded in the pin morph of *L. narbonense*. In the immature buds at GS1, there was relatively little expression of *PUM*^{T}, with the greatest recorded read count being short of 80 reads.

As is evident from read count tables (Tables 2.12, 2.13 and 2.14), *CYP*^{T} was not highly expressed in any individual; each sample was taken from one single plant. This relationship has therefore not been displayed graphically. At GS1, there is almost no expression of *CYP*^{T}, with only 3 reads recorded in the *L. viscosum* thrum morph (Table 2.12). Similarly, the read count tables for GS2 (Table 2.13) and GS3 + GS4 (Table 2.14) indicate that there was no expression of *CYP*^{T} whatsoever.

Many genes showed no expression at all in either floral morph of any species. Consequently it was deemed irrelevant to display these graphically. There was no expression at all of *LgAP1* in either floral morph of any species (Table 2.13), despite the presence of an *L. tenue* transcript for reads to map against.

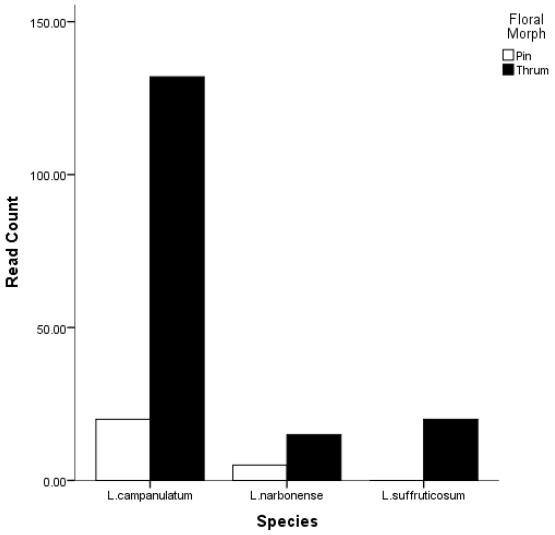


(a) GLO^T expression in immature buds (GS1)

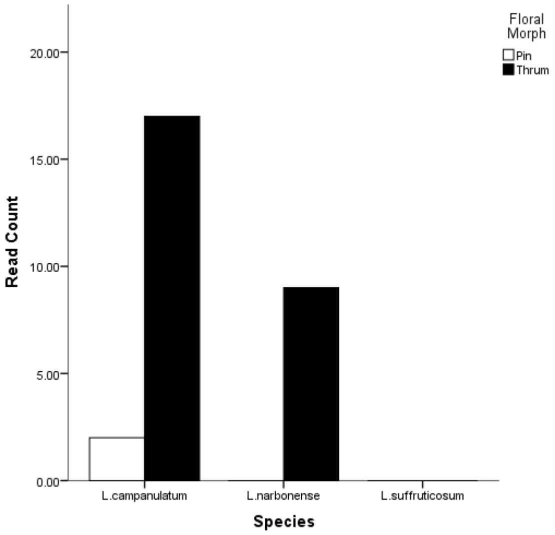
(b) PUM^T expression in immature buds (GS1)

(c) *LgSKS1* expression in immature buds (GS1)

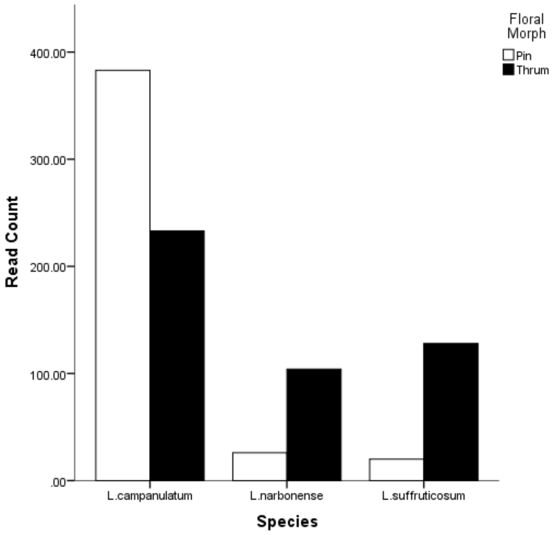
Figure 2.19: Raw read counts of tested candidate genes in a series of different heterostylous individuals at tested growth stages.



(d) *GLO*^{T} expression in mature buds (GS2)

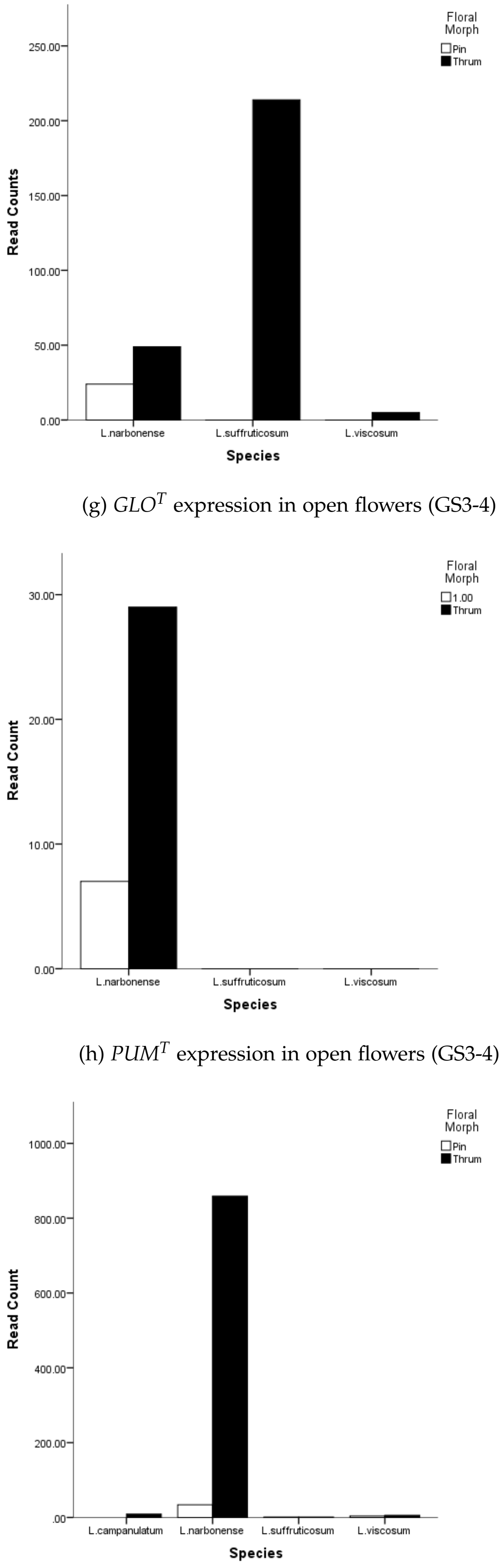


(e) *PUM*^{T} expression in mature buds (GS2)



(f) *LgSKS1* expression in mature buds (GS2)

Figure 2.19: Raw read counts of tested candidate genes in a series of different heterostylous individuals at tested growth stages.

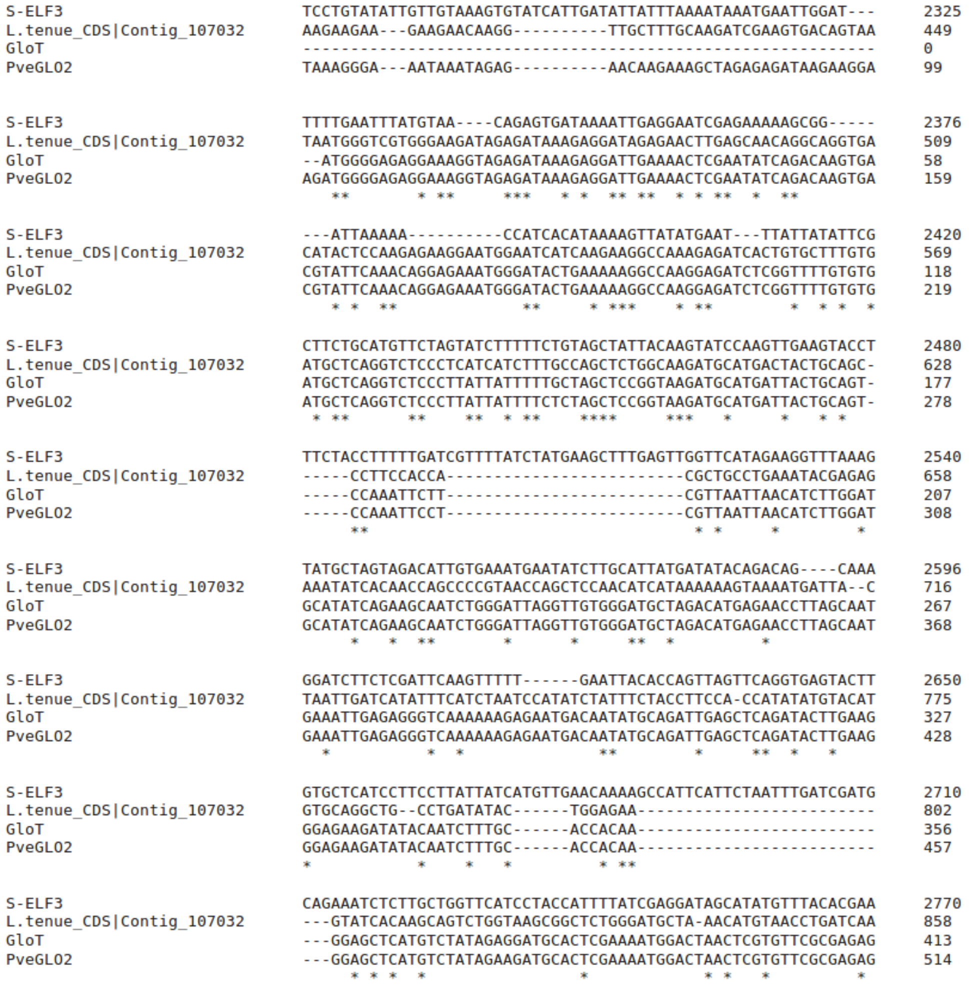


(g) *GLO*^{T} expression in open flowers (GS3-4)

(h) *PUM*^{T} expression in open flowers (GS3-4)

(i) *LgSKS1* expression in open flowers (GS3-4)

Figure 2.19: Raw read counts of tested candidate genes in a series of different heterostylous individuals at tested growth stages.



```
S-ELF3                         TCCTGTATATTGTTGTAAAGTGTATCATTGATATTATTTAAAATAAATGAATTGGAT---     2325
L.tenue_CDS|Contig_107032      AAGAAGAA---GAAGAACAAGG----------TTGCTTTGCAAGATCGAAGTGACAGTAA     449
GloT                           ------------------------------------------------------------     0
PveGLO2                        TAAAGGGA---AATAAATAGAG----------AACAAGAAAGCTAGAGAGATAAGAAGGA     99


S-ELF3                         TTTTGAATTTATGTAA----CAGAGTGATAAAATTGAGGAATCGAGAAAAAGCGG-----     2376
L.tenue_CDS|Contig_107032      TAATGGGTCGTGGGAAGATAGAGATAAAGAGGATAGAGAACTTGAGCAACAGGCAGGTGA     509
GloT                           --ATGGGGAGAGGAAAGGTAGAGATAAAGAGGATTGAAAACTCGAATATCAGACAAGTGA     58
PveGLO2                        AGATGGGGAGAGGAAAGGTAGAGATAAAGAGGATTGAAAACTCGAATATCAGACAAGTGA     159
                                  **       * **     ***   * *  ** **  * * **  *  **

S-ELF3                         ---ATTAAAAA----------CCATCACATAAAAGTTATATGAAT---TTATTATATTCG     2420
L.tenue_CDS|Contig_107032      CATACTCCAAGAGAAGGAATGGAATCATCAAGAAGGCCAAAGAGATCACTGTGCTTTGTG     569
GloT                           CGTATTCAAACAGGAGAAATGGGATACTGAAAAAGGCCAAGGAGATCTCGGTTTTGTGTG     118
PveGLO2                        CGTATTCAAACAGGAGAAATGGGATACTGAAAAAGGCCAAGGAGATCTCGGTTTTGTGTG     219
                                  * *  **             **     * ***    * **        *  * *  *

S-ELF3                         CTTCTGCATGTTCTAGTATCTTTTTCTGTAGCTATTACAAGTATCCAAGTTGAAGTACCT     2480
L.tenue_CDS|Contig_107032      ATGCTCAGGTCTCCCTCATCATCTTTGCCAGCTCTGGCAAGATGCATGACTACTGCAGC-     628
GloT                           ATGCTCAGGTCTCCCTTATTATTTTTGCTAGCTCCGGTAAGATGCATGATTACTGCAGT-     177
PveGLO2                        ATGCTCAGGTCTCCCTTATTATTTTCTCTAGCTCCGGTAAGATGCATGATTACTGCAGT-     278
                                * **      **    **  * **    ****     ***   *     *   * *

S-ELF3                         TTCTACCTTTTTGATCGTTTTATCTATGAAGCTTTGAGTTGGTTCATAGAAGGTTTAAAG     2540
L.tenue_CDS|Contig_107032      -----CCTTCCACCA-------------------------CGCTGCCTGAAATACGAGAG     658
GloT                           -----CCAAATTCTT-------------------------CGTTAATTAACATCTTGGAT     207
PveGLO2                        -----CCAAATTCCT-------------------------CGTTAATTAACATCTTGGAT     308
                                    **                                  * *     *        *

S-ELF3                         TATGCTAGTAGACATTGTGAAATGAATATCTTGCATTATGATATACAGACAG----CAAA     2596
L.tenue_CDS|Contig_107032      AAATATCACAACCAGCCCCGTAACCAGCTCCAACATCATAAAAAAGTAAAATGATTA--C     716
GloT                           GCATATCAGAAGCAATCTGGGATTAGGTTGTGGGATGCTAGACATGAGAACCTTAGCAAT     267
PveGLO2                        GCATATCAGAAGCAATCTGGGATTAGGTTGTGGGATGCTAGACATGAGAACCTTAGCAAT     368
                                    *   *  **       *      *     **  *         *

S-ELF3                         GGATCTTCTCGATTCAAGTTTTT------GAATTACACCAGTTAGTTCAGGTGAGTACTT     2650
L.tenue_CDS|Contig_107032      TAATTGATCATATTTCATCTAATCCATATCTATTTCTACCTTCCA-CCATATATGTACAT     775
GloT                           GAAATTGAGAGGGTCAAAAAAGAGAATGACAATATGCAGATTGAGCTCAGATACTTGAAG     327
PveGLO2                        GAAATTGAGAGGGTCAAAAAAGAGAATGACAATATGCAGATTGAGCTCAGATACTTGAAG     428
                                 *          *  *              **        *     **  *   *

S-ELF3                         GTGCTCATCCTTCCTTATTATCATGTTGAACAAAAGCCATTCATTCTAATTTGATCGATG     2710
L.tenue_CDS|Contig_107032      GTGCAGGCTG--CCTGATATAC------TGGAGAA-------------------------     802
GloT                           GGAGAAGATATACAATCTTTGC------ACCACAA-------------------------     356
PveGLO2                        GGAGAAGATATACAATCTTTGC------ACCACAA-------------------------     457
                               *           *    *   *         * **

S-ELF3                         CAGAAATCTCTTGCTGGTTCATCCTACCATTTTATCGAGGATAGCATATGTTTACACGAA     2770
L.tenue_CDS|Contig_107032      ---GTATCACAAGCAGTCTGGTAAGCGGCTCTGGGATGCTA-AACATGTAACCTGATCAA     858
GloT                           ---GGAGCTCATGTCTATAGAGGATGCACTCGAAAATGGACTAACTCGTGTTCGCGAGAG     413
PveGLO2                        ---GGAGCTCATGTCTATAGAAGATGCACTCGAAAATGGACTAACTCGTGTTCGCGAGAG     514
                                    * * *  *                *            * *   *         *
```

Figure 2.20: Multiple sequence alignment, performed using Clustal Omega (Sievers et al., 2011), of *S-ELF3*, *PveGLO2*, *GLO*^{T} and Contig 107032, from the *L. tenue* transcriptome scaffold. Asterisks indicate conserved bases.

Figure 2.20 depicts the most conserved portion of the multiple sequence alignment. *GLO*^{T} and *PveGLO2* are shown to have identical sequences. This is not entirely unexpected, given that both Nowak et al. (2015) and Li et al. (2016) were working with *Primula* and it is highly likely that these are the same gene.

Table 2.12: Raw count data for candidate genes at GS1

Gene	*L. campanulatum*		*L. narbonense*		*L. suffruticosum*		*L. viscosum*	
	Pin	**Thrum**	**Pin**	**Thrum**	**Pin**	**Thrum**	**Pin**	**Thrum**
S-ELF3	51	151	3	6	5	0	41	154
CYP734A50	1	19	1	8	2	0	0	2
LgMYB21	0	1	0	0	0	0	2	0
LgSKS1	2	7	8	20	47	10	46	393
TSS1	1	0	0	0	0	0	1	0
TkST1	0	8	0	0	0	2	0	0
TkNACE	0	0	2	2	0	0	0	0
PUM^T	0	0	4	13	2	1	0	22
CCM^T	0	4	0	4	0	0	0	0
CYP^T	0	0	0	0	0	0	0	3
KFB^T	0	0	0	0	0	0	0	1
TsRETRO	1	0	1	1	0	0	0	0

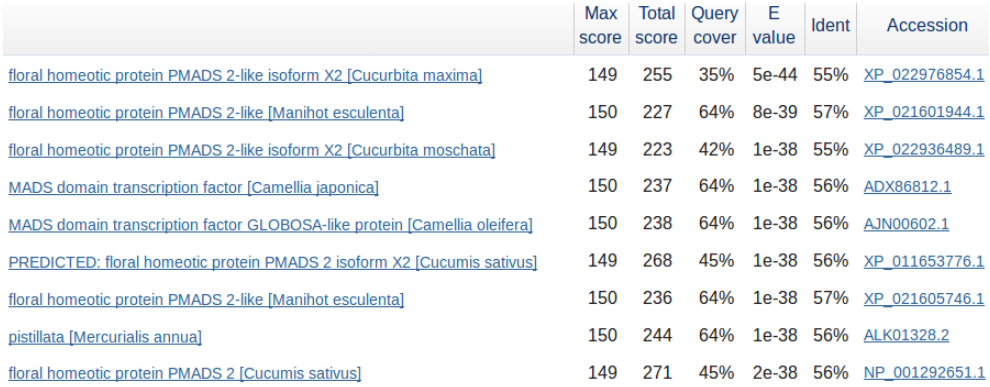
Table 2.13: Raw count data for candidate genes at GS2

Gene	*L. campanulatum*		*L. narbonense*		*L. suffruticosum*	
	Pin	**Thrum**	**Pin**	**Thrum**	**Pin**	**Thrum**
S-ELF3	20	132	5	15	0	32
CYP734A50	0	24	0	1	0	1
LgAP1	0	0	0	0	0	0
LgMYB21	0	11	0	0	0	0
LgSKS1	383	233	26	104	20	128
TsRETRO	0	0	0	0	0	0
TSS1	1	0	0	0	2	0
TkNACE	0	0	0	3	0	1
TkST1	0	0	0	0	0	0
TPP1	165	75	38	16	0	69

Table 2.14: Raw count data for candidate genes at GS3 and GS4.

Gene	*L. campanulatum*	*L. narbonense* GS3		*L. narbonense* GS4		*L. suffruticosum*		*L. viscosum*	
	Thrum	**Pin**	**Thrum**	**Pin**	**Thrum**	**Pin**	**Thrum**	**Pin**	**Thrum**
S-ELF3	90		4	8	17	0	58	0	58
CYP734A50	1	1	0	2	2	4	4	0	0
LgMYB21	1	0	0	6	2	0	1	0	0
LgSKS1	9	18	42	16	817	1	1	4	6
TSS1	0	0	5	1	5	0	1	0	0
TkST1	0	0	0	0	0	0	6	0	0
TkNACE	2	1	1	0	7	0	0	1	0
PUM^T	6	6	7	4	68	0	1	0	2
CCM^T	0	1	3	0	3	0	0	0	0
GLO^T	90	0	4	8	17	0	58	0	5

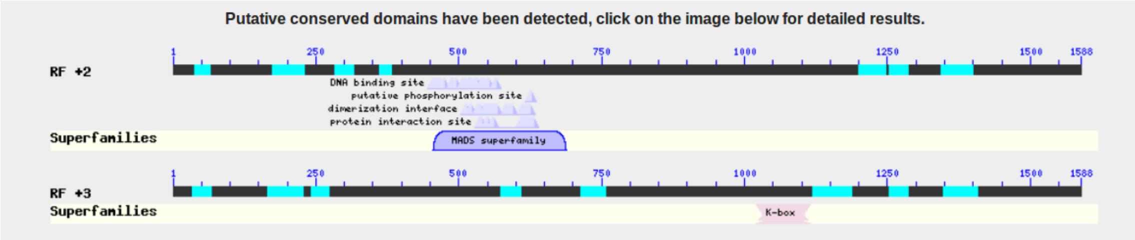
Figure 2.21: Top BLAST hits for Contig 107032.



	Max score	Total score	Query cover	E value	Ident	Accession
floral homeotic protein PMADS 2-like isoform X2 [Cucurbita maxima]	149	255	35%	5e-44	55%	XP_022976854.1
floral homeotic protein PMADS 2-like [Manihot esculenta]	150	227	64%	8e-39	57%	XP_021601944.1
floral homeotic protein PMADS 2-like isoform X2 [Cucurbita moschata]	149	223	42%	1e-38	55%	XP_022936489.1
MADS domain transcription factor [Camellia japonica]	150	237	64%	1e-38	56%	ADX86812.1
MADS domain transcription factor GLOBOSA-like protein [Camellia oleifera]	150	238	64%	1e-38	56%	AJN00602.1
PREDICTED: floral homeotic protein PMADS 2 isoform X2 [Cucumis sativus]	149	268	45%	1e-38	56%	XP_011653776.1
floral homeotic protein PMADS 2-like [Manihot esculenta]	150	236	64%	1e-38	57%	XP_021605746.1
pistillata [Mercurialis annua]	150	244	64%	1e-38	56%	ALK01328.2
floral homeotic protein PMADS 2 [Cucumis sativus]	149	271	45%	2e-38	56%	NP_001292651.1

Figure 2.21 shows that the sequence in *L. tenue* shows some similarity to MADS box genes. All floral B genes characterised to date, as with most floral homeotic genes, are MADS box genes (Riechmann & Meyerowitz, 1997) and encode MADS-domain transcription factors. Thus, the presence of the potential MADS box domain is indicative, at least, of B gene function in this context. However, only ~56% identity has been observed. The *GLO* gene may exhibit a somewhat different structure in *Linum* to equivalents in other species.

Figure 2.22: Putative conserved domains for Contig 107032, showing the presence of a potential MADS-domain (Retrieved from NCBI Blast).



Putative conserved domains have been detected, click on the image below for detailed results.

Table 2.15 shows that candidate gene expression exists in homostylous species in addition to heterostylous species. Expression of GLO^T is reasonably high across the homostylous species. However, the raw read counts of the homostylous species, in general, are lower than in heterostylous species. Interestingly, although read counts are still very low, there was more consistent expression of CYP^T in the homostylous species than in the heterostylous species at any growth stage.

2.3.2 DESeq2 Analysis

DESeq2 analysis was conducted to statistically determine whether genes were differentially expressed between the two floral morphs, and also to determine whether there was differential expression of genes between homostylous and heterostylous species. It

Table 2.15: Raw count data for candidate genes in homostylous species

Gene	*L. bienne* dor5			*L. bienne* lla10			*L. catharticum*			*L. setaceum*		*L. strictum*		*L. tenuifolium*				*L. usitatissimum*		
	GS1	**GS2**	**GS3**	**GS1**	**GS2**	**GS3**	**GS1**	**GS2**	**GS3**	**GS1**	**GS2**	**GS1**	**GS3**	**GS1**	**GS2**	**GS2**	**GS2**	**GS1**	**GS2**	**GS3**
S-ELF3	14	73	1	68	67	10	2	1	3	5	7	13	151	0	1	1	0	0	3	1
CYP734A50	0	2	0	1	8	2	0	0	0	0	2	10	1	0	1	1	0	0	2	2
LgMYB21	0	38	0	0	1	11	0	0	2	12	2	9	1	0	0	0	0	0	3	1
LgSKS1	2	29	2	1	17	13	3	2	5	19	1	5	2	1	3	2	1	19	109	5
TSS1	0	0	0	0	1	0	1	0	0	1	5	0	1	0	0	1	0	0	0	0
TkST1	0	0	0	0	1	1	0	0	0	2	2	0	8	0	0	0	0	0	0	0
TkNACE	0	7	0	0	4	2	1	1	1	1	1	1	1	0	0	0	0	0	1	0
PUM^T	0	0	0	0	2	1	4	5	2	1	2	2	4	0	3	2	1	2	5	1
CCM^T	0	1	0	99	10	4	0	0	0	0	0	0	1	0	0	0	0	4	3	0
GLO^T	14	73	1	68	67	10	2	1	3	5	7	13	151	0	1	1	0	0	3	1
CYP^T	0	1	0	0	1	0	0	1	2	0	1	0	7	0	1	0	1	0	2	4
TsRETRO	1	0	0	1	4	2	0	0	1	1	0	0	1	0	0	0	0	2	3	1
TPP1	1	9	0	0	4	0	0	0	0	15	0	4	1	0	1	0	1	0	3	1

is necessary to note at this stage that the differential expression analysis presented in this chapter was based on a single RNAseq dataset; there are no technical replicates and individuals of different species at each growth stage were utilised as biological replicates. As such, the analyses presented should be interpreted with caution, and cannot be used to draw firm conclusions about differential expression within species. The shortcomings in the dataset have been realised in hindsight and are the product of an attempt to analyse a wide set of wild species and low RNA concentrations from the extraction process preventing analysis of all growth stages for each species. At this stage it is not possible to collect new data samples and the volume of work required to address the discussed issues with the dataset would be more in line with a PhD project.

In general, results were inconclusive. Nevertheless, the largest observed difference in expression was at GS1, the immature bud stage. There was an additional observable difference between the long-styled and short-styled morphs at GS2.

There were such a small number of heterostylous individuals at GS3, the mature bud stage, that these were analysed in the same group as GS4, the open flower. The grouping of GS3 with GS4 rather than with GS2 was based on the fact that, immediately prior to flower opening opening (GS3), the gene expression profile is likely to be more similar to open flowers than to younger, still developing buds. Wellmer et al. (2006) note that as a part of the complex regulatory network underlying flower development, the vast majority of identified floral regulatory genes act during the very early stages of flower formation; in the establishment of floral meristem identity, or in floral meristem patterning as a precursor for the development of petals, stamens, carpels etc. In contrast, comparatively few genes that function specifically at the later stages of flower development have been identified.

Results are displayed graphically and lists of the top most significant differentially expressed genes have been produced. DESeq2 analysis identified 816, 869 and 116659 differentially expressed genes at the overall corrected 0.05 significance level for GS1, GS2 and GS3-GS4 respectively. The very large number of differentially expressed genes at GS3-4 is worthy of note and is replicable. Furthermore, DESeq2's reported values for the top six most differentially expressed transcripts between the short- and long-styled morphs across all the heterostylous species at GS1, GS2 and GS3 are depicted in Tables 2.16, 2.17 and 2.18. An unusual phenomenon of the exact DESeq2 result not being reproducible between runs was detected. This is indicated in Figure 2.23 and considered in detail in Section2.4. Other instances of such a phenomenon have been discussed on the DESeq2 forums (e.g. Yjiangnan, 2017).

Log2FoldChange (LFC) is the effect size estimate; for a particular gene, a Log2 fold

Table 2.16: The top six most differentially expressed genes based on DESeq2 analysis of GS1 heterostylous species. The Base Mean is the mean of normalised counts of all samples, normalising for sequencing depth. lfcSE is the Log Fold Change Standard Error

Contig	**Base Mean**	**Log2 Fold Change**	**lfcSE**	**Stat**	**P Value**	**Adjusted P Value**
L. tenue CDS Contig 2317	159.788	4.269	0.276	15.450	7.441e-54	1.128e-48
TR48232 c0 g1 i1	34924.597	5.043	0.354	4.943e-46	3.746e-41	
TR31923 c0 g1 i6	602.385	-7.245	0.540	-13.420	4.589e-41	2.318e-36
L. tenue CDS Contig 89046	45.217	5.527	0.462	11.958	5.846e-33	2.215e-28
L. tenue CDS Contig 84453	179.647	3.351	0.284	11.772	5.418e-32	1.642e-27
TR66435 | c2 g4 i6	380.32587	4.380	0.401	10.926	8.697e-28	2.197e-23

Table 2.17: The top six most differentially expressed genes based on DESeq2 analysis of GS2 heterostylous species.

Contig	**Base Mean**	**Log2 Fold Change**	**lfcSE**	**Stat**	**P Value**	**Adjusted P Value**
L. tenue CDS Contig 119239	552.071	5.310	0.488	10.880	1.441e-27	9.829e-24
TR32874 c0 g1 i1	41737.857	-9.157	1.015	- 9.019	1.89e-19	6.464e-16
L. tenue CDS Contig 127580	19714.873	-6.335	0.743	-8.526	1.518e-17	3.453e-14
L. tenue_CDS Contig_127381	placeholder	5.749	0.700	8.218	2.073e-16	3.536841e-13
L. tenue CDS Contig 127681	3347.610	-7.761	0.955	-8.130	4.296e-16	5.862e-13
L. tenue CDS Contig 128001	9646.194	-5.279	0.656	-8.050	8.232e-16	9.362e-13

Table 2.18: The top six most differentially expressed genes based on DESeq2 analysis of GS3 and GS4 heterostylous species.

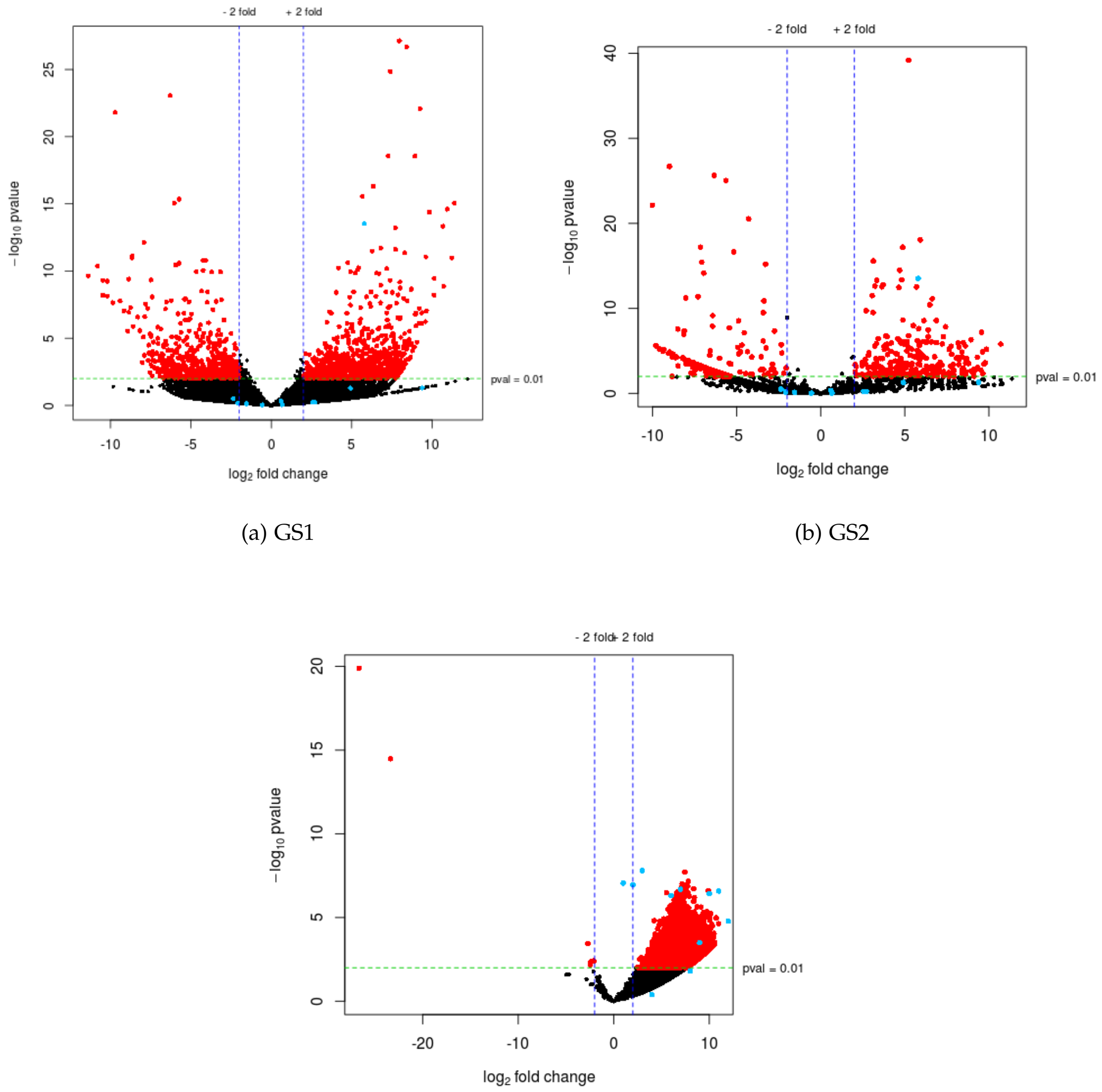
Contig	**Base Mean**	**Log2 Fold Change**	**lfcSE**	**Stat**	**P Value**	**Adjusted P Value**
L. tenue CDS Contig 40109	16.481	-26.651	2.863	-9.308	1.301E-20	1.880E-15
L. tenue CDS Contig 87293	4.676	-23.370	2.966	-7.880	3.281E-15	2.370E-10
TR59037 c0 g1 i1	3356.728	-2.713	0.762	-3.562	0.0004	0.0099
L. tenue CDS Contig 114152	48.468	3.482	1.140	3.055	0.0023	0.0099
L. tenue CDS Contig 128000	331.772	3.530	1.136	3.108	0.002	0.0099
L. tenue CDS Contig 88505	11.627	3.583	1.073	3.338	0.0008	0.0099

Figure 2.23: Two separate DESeq2 analyses of heterostylous individuals. The DESeq2 results were not exactly reproducible between runs on the same data. In this case significance was tested at p <0.1, although results are reported at p <0.05.

```
out of 109305 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 516, 0.47%
LFC < 0 (down)     : 8880, 8.1%
outliers [1]       : 0, 0%
low counts [2]     : 69933, 64%
(mean count < 1)

out of 168648 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 513, 0.3%
LFC < 0 (down)     : 8857, 5.3%
outliers [1]       : 0, 0%
low counts [2]     : 128874, 76%
```

change of -1 for short-styled individuals vs long-styled individuals would be indicative of a change in observed expression level of 2^{-1}, or 0.5.

At GS1, of the 152005 contigs with a non-zero total read count, 464 (0.31%) had a Log2FoldChange greater than 0, indicating up-regulation, and 352 (0.23%) had a Log2FoldChange of less than 0, indiciating down-regulation. At GS2, of the 151553 contigs with a non-zero total read count, 421 (0.28%) were up-regulated and 169 (0.11%) were down-regulated. The most differentially expressed gene, located at Contig 2317, which exhibits a log2FoldChange of 4.27, there is a change in observed expression level of 19.27 fold. At GS3 + 4, 144490 contigs displayed a non-zero read count. There was a much smaller observed difference in expression between the transcripts, which is reflected by the fact that only 88 (0.061 %) were recorded as having read counts that were too low for analysis. At GS3 + 4, 132123 (91%) of the genes were up-regulated and 0.013% of the genes were down-regulated. When all heterostylous species were compared together in the same analysis rather than being separated by growth stage, 272 (0.25%) were up-regulated and 5492 (5%) were down-regulated at the p < 0.05 significance level.



(a) GS1

(b) GS2

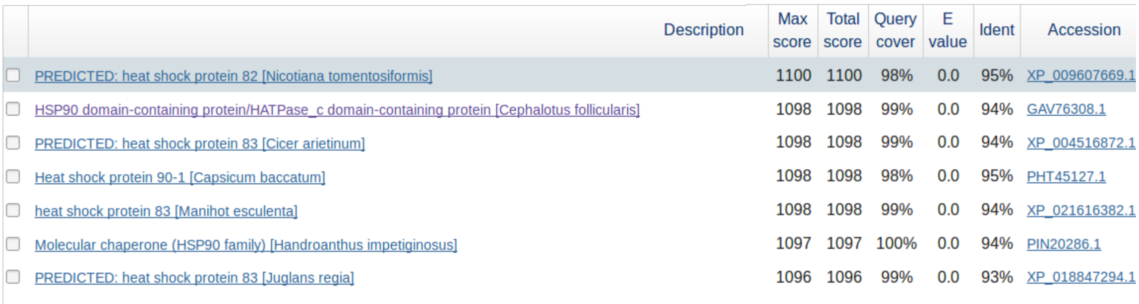
(c) GS3 & GS4

Figure 2.24: Volcano plots of heterostylous species at **a)** GS1, **b)** GS2 and **c)** GS3 + 4 depicting Log2 fold change on the x-axis against the -log10 p value on the y axis. Genes significant at p <0.01 are depicted as red circles. Transcripts matching the candidate genes are depicted as pale blue circles. Blue lines show the ±2 log fold change.

As depicted by the volcano plots, there is a large amount of significant expression (red) at all three tested stages. The tested *S-locus* candidate genes are marked in pale blue on the volcano plots in Figure 2.24. Although not statistically significant, many of the candidate genes show a +2-fold change in expression at GS1 and GS2. Therefore, there is a change in expression profile between the floral morphs. *LgSKS1* is differentially expressed at p > 0.01 at both GS1 and GS2. GS4's plot (Figure 2.24c) is particularly interesting as all of the significantly expressed genes are up-regulated, rather than down-regulated and many of the candidate genes appear to show up-regulation. The candidates, however, are all reported as 'outliers' in the DESeq2 results output. Overall, there is a smaller fold change between pin and thrum at GS4 (Figure 2.24c) than at GS2 or GS3.

2.3.3 Stress Response Roles for the most Significantly Differentially Expressed Genes

Contig 2317, which was the most differentially expressed gene at GS1, showed 92-98% identity to a homeobox-leucine zipper protein (HDG2). These proteins are transcription factors, unique to plants. They are generally involved in abiotic stress responses (Elhiti & Stasolla, 2009). HD-Zip II proteins also have roles in light response and shade-avoidance in *Arabidopsis*. The most differentially expressed gene at GS2 (Contig 119239) shows significant (90-95%) homology to a series of heat shock proteins in the HSP90 superfamily (Figure 2.25). HSP90 is a highly conserved and abundant chaperone protein that assists the correct folding of other proteins in addition to stabilising proteins against heat stress (e.g. Li et al., 2012). Perhaps the significant differential expression of such a heat shock protein is in response to variability in environmental conditions at the point of picking the wild flower, rather than being directly related to floral development. Contig 40109 shows some similarity to a beta-glucosidase; beta-glucosidase enzymes catalyse the hydrolysis of glycosidic bonds.



	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	PREDICTED: heat shock protein 82 [Nicotiana tomentosiformis]	1100	1100	98%	0.0	95%	XP_009607669.1
□	HSP90 domain-containing protein/HATPase_c domain-containing protein [Cephalotus follicularis]	1098	1098	99%	0.0	94%	GAV76308.1
□	PREDICTED: heat shock protein 83 [Cicer arietinum]	1098	1098	99%	0.0	94%	XP_004516872.1
□	Heat shock protein 90-1 [Capsicum baccatum]	1098	1098	98%	0.0	95%	PHT45127.1
□	heat shock protein 83 [Manihot esculenta]	1098	1098	99%	0.0	94%	XP_021616382.1
□	Molecular chaperone (HSP90 family) [Handroanthus impetiginosus]	1097	1097	100%	0.0	94%	PIN20286.1
□	PREDICTED: heat shock protein 83 [Juglans regia]	1096	1096	99%	0.0	93%	XP_018847294.1

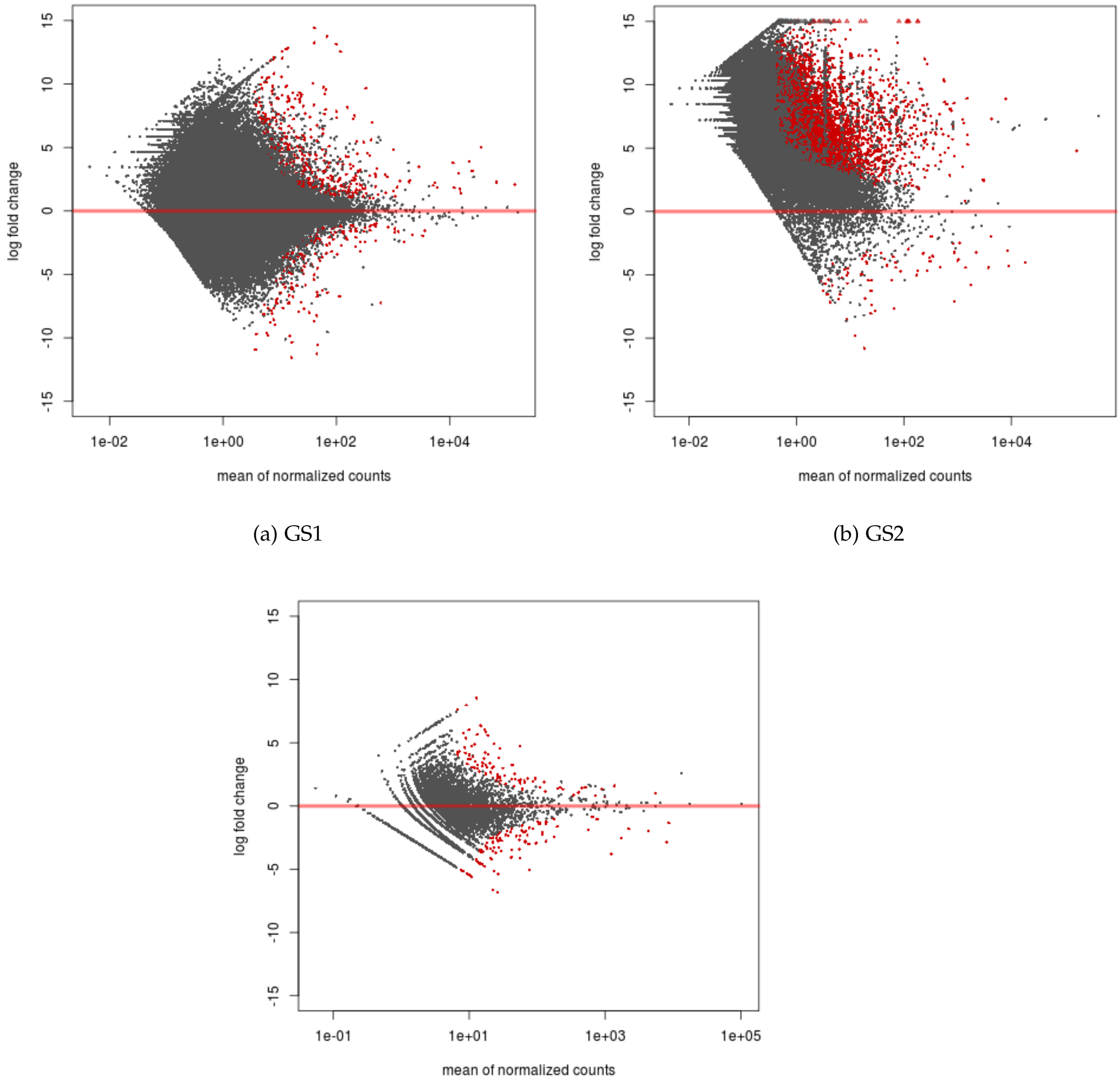
Figure 2.25: Contig 119239 shows significant homology to members of the HSP90 superfamily.

Contig 89046, which is differentially expressed at GS1 resembles stem-specific protein TSJT1. Little is known about the function of this protein (Hu et al., 2016), however, its stem-specific nature possibly suggests that some portion of the stems of some of the buds were included in the extraction and library creation, leading to a difference in gene expression. The first 125 bases at the 5′ end of Contig 84453 show a conserved domain hit of the catalytic domain of a Serine/Threonine Kinase. The BLAST query indicates that no similarity is evident for the remaining 900 bases.

Contig 127580 shows similarity to the GAT-1 superfamily of genes which possess a glutamine-binding domain. GAT-1 is a sodium- and chloride-coupled γ-aminobutyric acid transporter, originally discovered in the rat brain. The binding of glutamine to a glutamine-binding protein is the first step in the active transport of L-glutamine across the cytoplasmic membrane. In plants, glutamate-derived γ-aminobutyric acid (GABA) is thought to be involved at the crossroad between C and N metabolism and is also shown to accumulate under various environmental conditions (Batushansky et al., 2015). GABA has been treated as a metabolite, largely in the context of stress-response (Bouché & Fromm, 2004). The induced changes in protein structure act to increase protein stability: a potential stress response.

Certain contigs could not be matched to hits in the *Arabidopsis* TAIR database. These are designated 'TR', rather than being named based on their hit in the TAIR database (see e.g. Tables 2.16, 2.17 and 2.18), according to a naming convention selected by Ali Foroozani. It is thus more difficult to assign definitive functions to the transcripts aligning to these contigs. Many of the top most differentially expressed contigs have only small regions of putative conserved domains.

Also differentially expressed at GS2 is contig 127381. This gene shows some similarity to the acetyl-CoA carboxylase carboxyltransferase beta subunit, which is located in the chloroplast. This subunit is involved in fatty acid synthesis (Sasaki & Nagano, 2004). Contig 127681 has a sequence corresponding to the ribosomal S4 protein. This is likely to be an abundant protein and perhaps the difference in its expression is related to the size of the sequence file and the high achieved mapping percentage.



(a) GS1

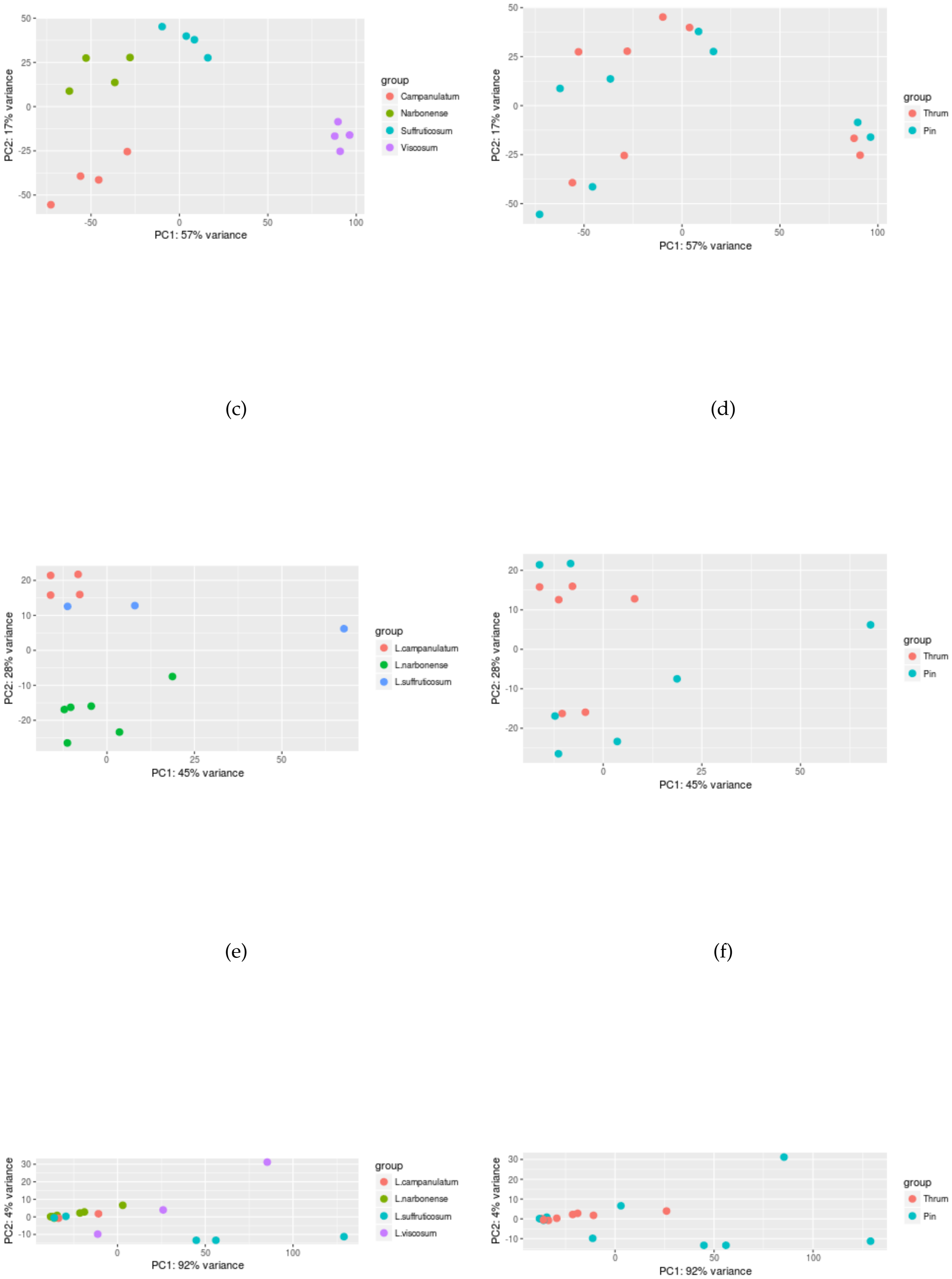
(b) GS2

(c) GS3

Figure 2.26: MA plots showing differential gene expression between long-styled and short-styled morphs at **a)** GS1 **b)** GS2 and **c)** GS3 + GS4. Statistically significant differences (with an adjusted p-value of <0.05) in expression are marked by a red dot. Triangles point to points falling outside of the plot area.

The MA plot in Figure 2.26a displays a log ratio (M) vs an average (A) in order to visualise the differences between the long-styled morphs and the short-styled morphs. The x-axis represents the average 2log ratio of expression over all samples and the y axis the log2 fold change of normalised counts (i.e. the average of the counts, normalised by a size factor) between long-styled and short-styled morphs. In general, the gene expression remains consistent between conditions, most points reside on the y-intercept at 0. The MA plot for GS1 shows significant differences in gene expression. Figure 2.26a demonstrates that in general, only genes with large average normalised counts contain sufficient information to yield a significant result. This phenomenon is also observed at GS2 (Figure 2.26b) and at GS3-GS4 (Figure 2.26c).

Figure 2.27: PCA plots show the variance in gene expression at GS1 across the different species in the two floral morphs. Points cluster by species, rather than by floral morph type.



(a)

(b)

(c)

(d)

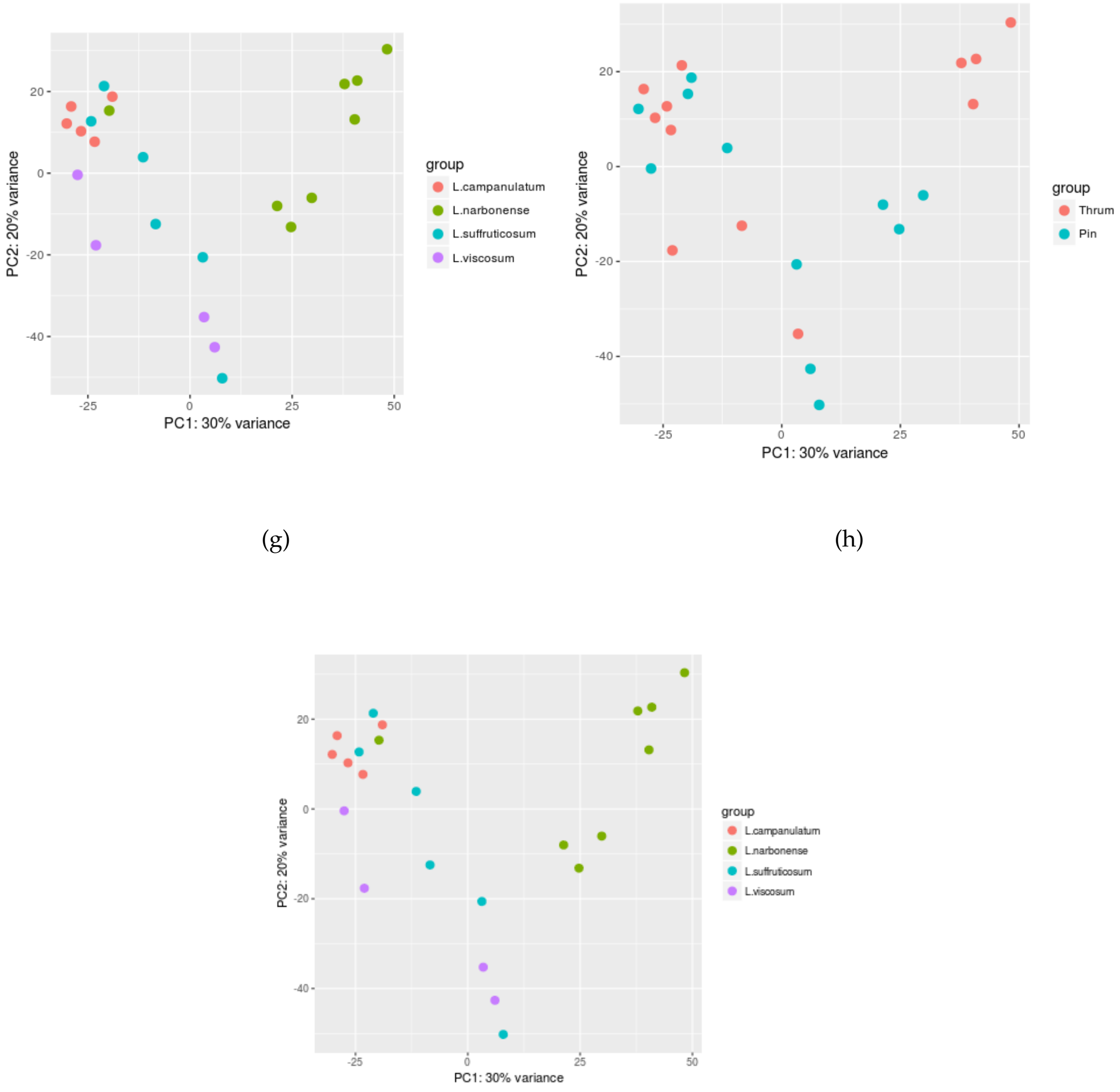
(e)

(f)

Interpretation of the Principal Component Analysis (PCA) is based on determining which of the transcripts contribute to the variance of each component, i.e. which of these numbers are large in magnitude, the furthest from zero in either positive or negative direction. The PCA of GS1 expression showed the separation of gene expression profiles by species, and by floral morph type (Figure 2.27a). PC1 explained more variance than PC2 (17% vs 57%). As evidenced by the PCA, *L. suffruticosum* species cluster closely to *L. narbonense* species. *L. viscosum* species are further removed from the other species. Hence, at GS1 there is clear clustering by species. It is also possible to argue that there is some separation between pin at thrum, however, this effect is not nearly as obvious as that of species.

At GS2, the PCA shows that the difference between species is most explained by PC2 (Figure 2.27c). The difference between the individual thrum morphs is largely explained by GS2. There is only 28% variance in PC2 compared with 45% variance in PC1. There is considerably more variance between the pin morphs. In PC1, *L. campanulatum* is reasonably closely clustered with *L. narbonense*. The variance in *L. suffruticosum* is mostly explained by PC1, there is very little variation at PC2. *L. campanulatum* clusters by species. The variance in *L. narbonense* is most explained by PC1.

At GS3 and GS4, only 4% variance exists in PC2 whereas there is an enormous amount of variance (92%) in PC1. Therefore, almost all of the difference is explained by PC1. The thrum morphs are closely clustered together, whereas the pin morphs display wide variance. Almost all of the variance in *L. suffruticosum* is explained by PC1 and *L. suffruticosum* is far removed from other species. Additionally, there is an *L. viscosum* outlier. This individual did not map well onto the *L. tenue* transcriptome (Table 2.5).



(g)

(h)

(i)

Figure 2.28: PCA plots show the variance between gene expression across the different species in the two floral morphs at all four growth stages.

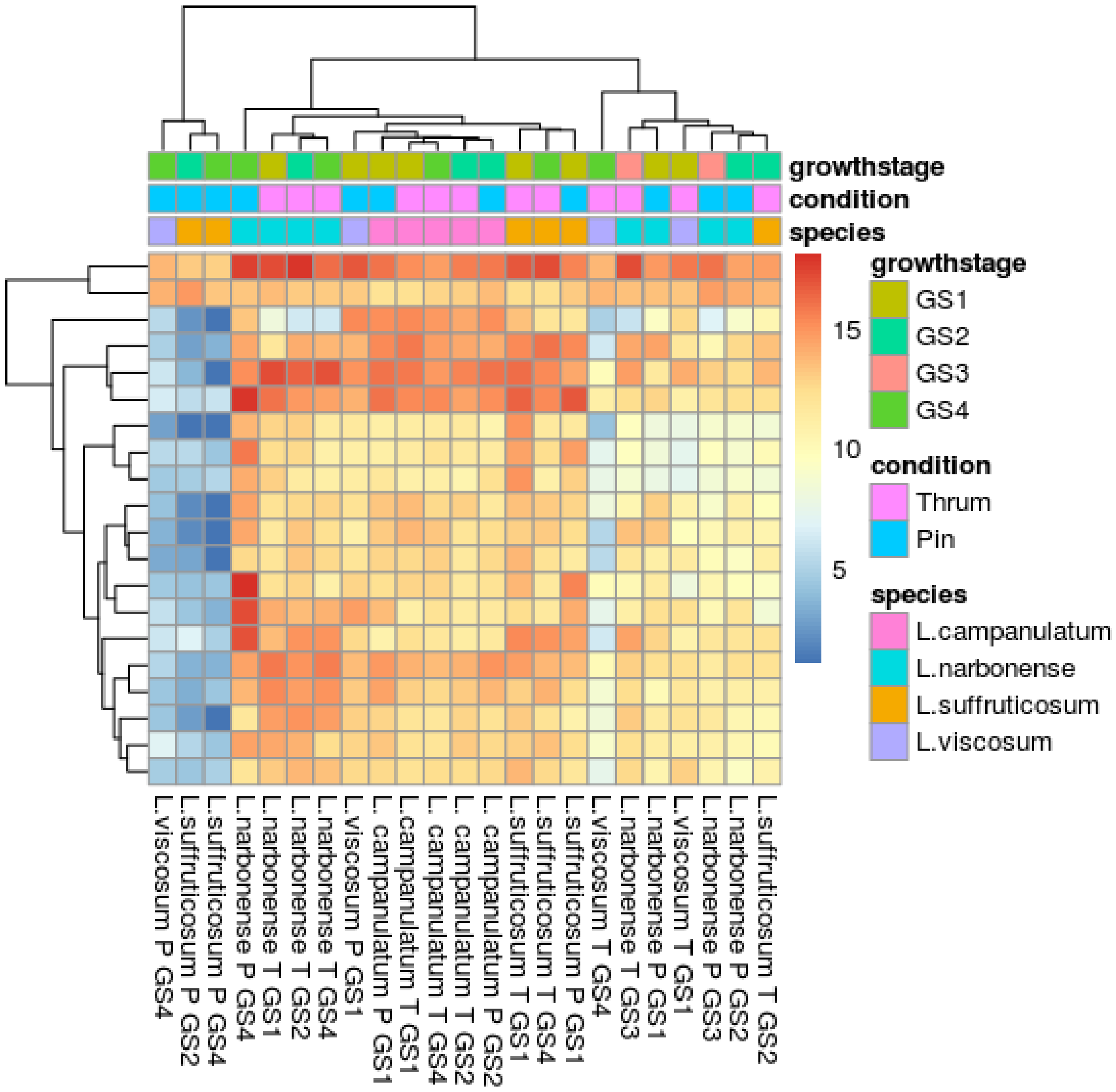


Figure 2.29: Heatmap of the count matrix, showing the clustering of the different heterostylous species across all growth stages for the most differentially expressed genes. Red shades indicate more closely related sequences Darker shades of blue indicate more closely related sequences.

In Figure 2.28 all heterostylous species are plotted together on the same axes. This figure shows the variance in gene expression across species in the two floral morphs at all four growth stages. It can be observed from Figure 2.28g that clustering is largely by species. There appears to be no clustering correlation within the growth stages. A similar trend is depicted in the heatmap (Figure 2.29). The heatmap enables to determine which samples are the most similar to one another. Blocks of samples which are similar in colour show more closely related trends to one another than to the samples in the other blocks. Again, it is evident that the closest clustering is by species. The heatmap does not reveal any clear expression pattern based on condition or on growth stage. Low expression of the top 20 differentially expressed genes in *L. suffruticosum* pin morphs at GS2 and GS4 is evident, which may be related to the low achieved mapping percentages

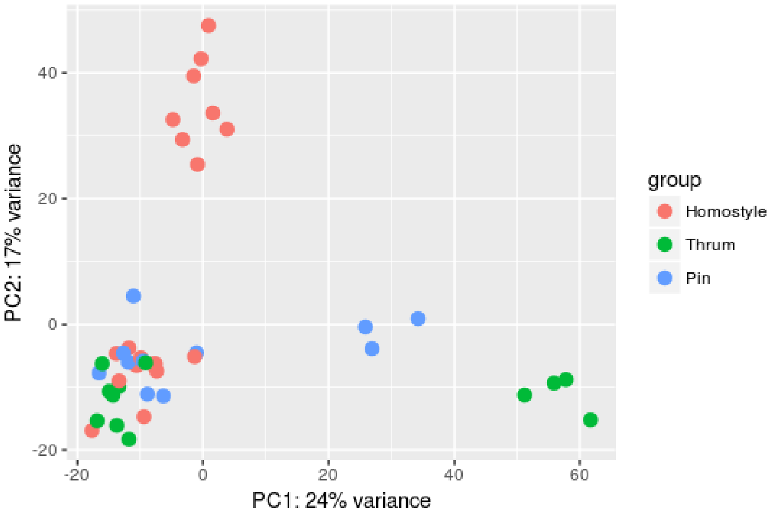
(Table 2.5).

An analysis comparing heterostylous and homostylous individuals was also conducted. From the PCA in Figure 2.30a, it is clear that there is some clustering of individuals depending on morph type. However, many other individuals show very little variance. Cross-referencing with Figure 2.30b shows that the source of the differential morph expression is all from *L. narbonense*. There is, overall, a far smaller effect of species 2.30b. No individuals were differentially expressed at the $p < 0 .05$ level when comparing homostylous and heterostylous individuals together.

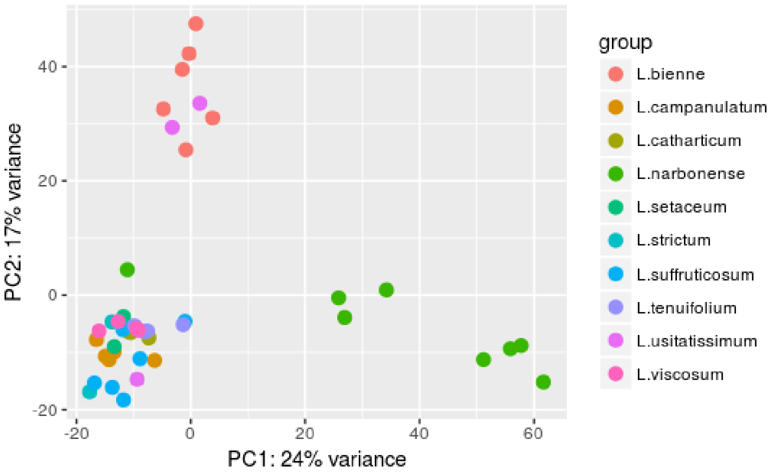
2.4 Discussion

Two approaches to determining differential expression between the *Linum* species were employed. Firstly, raw counts of the candidate genes identified in the literature search were obtained and compared and secondly, differential expression analysis of all observed gene expression using DESeq2 was conducted. The functions of the top 6 most upregulated and downregulated genes, for each short-to long-styled morph comparison at each growth stage as determined by DESeq2 were investigated. The fold change and p-values of the candidate genes, when subjected to DESeq2 analysis were also scrutinised and reported. This two-pronged approach was necessitated by certain aspects of DESeq2's algorithm and by the initial experimental design. DESeq2 tends to underestimate changes in gene expression without biological replicates (Robles et al., 2012), and consequently analysing the raw read counts provides insight into the changes in gene expression in the identified candidate genes independently of whether DESeq2 was able to detect them as significantly different. Equally, as has been explained in the introduction section, a lack of technical replicates and absence of strict biological replicates mean that the differential analysis using DESeq2 should not be used to draw statistical conclusions. Although it is not an ideal experimental design, it is valid to accept heterostylous individuals from different species as biological replicates of a sort. The *'S-locus'* is highly conserved and the inheritance pattern thought to be hemizygous and therefore, it is reasonable to assume that any signal of differential expression between pin and thrum morphs may be detectable — even between species.

Firstly, an analysis of the raw read count data showed that *GLO*^{T} was expressed to a greater extent in the short-styled thrum morph than the long-styled pin morph in all species at GS2 and GS3+4 (Figures 2.19d and 2.19g); in young and mature buds as well as in open flowers. At GS1, greater expression was observed in the thrum morph in all species apart from *L. suffruticosum* (Figure 2.19a). There was no *GLO*^{T} expression



(a)



(b)

Figure 2.30: PCA plot showing the variance in gene expression between homostylous and heterostylous species between the three morph types

at all in the pin morphs of *L. suffruticosum*, nor was there in the pin morphs of *L. viscosum*. In general, the recorded read counts were not high. These results largely accord with expectation. If heterostyly in *Linum* follows the hemizygous model outlined in *Primula* (Li et al., 2016), no expression in the pin morph would be expected. The differences in the number of raw read counts can also be explained by the number of reads in the RNA-Seq data as it was sequenced. Depth of sequencing can cause problems when interpreting raw read count data; hence the normalisation methods included in programs such as DESeq2, and consequently raw read count data should be interpreted with caution.

Differences in the number of raw read counts were considered for all of the candidate genes. However, as can be observed in Tables 2.12, 2.13 and 2.14 many of the contigs matching the candidate genes for heterostyly in *Linum* showed no expression at all. This is likely to have been a related to the low mapping percentages achieved by the cross-species maps (refer to Table 2.5). Nevertheless, some patterns of differential expression were observed in the count data.

No count reads could be attributed to the *CYP*^{T}-matched contig (Tables 2.12, 2.13, 2.14). This could possibly suggest that there is no role for a *CYP*^{T} orthologue in the co-ordination of heterostyly, although this does seem unlikely. There was some very small expression of *CYP734A50* at GS2 (Table 2.13),which perhaps indicates that the orthologue of this gene in *Linum* does not show close similarity. Given the results attained in Chapter 3, there is a strong chance that the *CYP*^{T} reads were simply unable to map to the *L. tenue* transcriptome. It could be possible that the primers designed for qPCR were too broad and amplified other, non *S-locus*-specific sequences, however, without repetition of the experiment it would be difficult to make any definitive judgments. At GS2 *PUM*^{T} expression largely accorded with expectation, with the exception of *L. suffruticosum*. The unusual behaviour of *L. suffruticosum* can be explained by poor mapping of this individual to the *L. tenue* transcriptome. *PUM*^{T} expression was otherwise fairly constant across morphs and growth stages, with the exception of GS3+ GS4 (Table 2.14), where there is very high expression in the *L. narbonense* thrum morph. This particular individual did not have an especially high mapping percentage (32.83% and 36.42% in August and November respectively). It may thus have been expressing the *PUM*^{T} gene to an unusually high extent.

There was evidence of high expression of *LgSKS1* at all tested growth stages in both floral morphs. At GS1, there were a greater number of raw counts in the thrum morph in *L. suffruticosum* and *L. viscosum*. There was greater expression in the thrum morph than the pin morph in *L. campanulatum* and *L. narbonense*, however, the overall count numbers were much lower in these species. Although *LgSKS1* is believed to be downstream of the *S-locus*, greater expression in the thrum morph would still be expected. Without the presence of the *S-locus* supergene, there would be no associated upregulation of *LgSKS1*. The high expression of *LgSKS1* relative to other genes may be a factor of superior mapping. These results could not be confirmed using qPCR (Chapter 3), as it proved impossible to develop primers for the consensus of the *Linum* species' transcriptomes. *LgSKS1* is thought to play a role in pollen inhibition and encodes a thrum-specific pollen protein, homologous to the *NTP303* (Ushijima et al., 2012) of tobacco, linking it tenuously to a downstream component of the self-incompatibility locus. It is unlikely that *LgSKS1* is a member of the *S-locus*, but it is possible that an orthologue could be cross-regulated as a part of heterostyly and self-incompatibility within morphs or cross-compatibility between morphs.

No raw read counts were extracted for *TsRETRO* at any growth stage, in any species (see section 2.3). Therefore, evidence from differential expression analysis suggests that *TsRETRO* is not involved in the *S-locus* in *Linum* or in its downstream functions. The *TsRETRO* primers were designed from a gene from *Turnera subulata*. Work from the Labonne et al. (2009) group has indicated a role for retrotransposition in the control of heteromorphy in *T. subulata*. It seems unlikely that at least *T. subulata*'s particular brand of retrotransposition controls heterostyly in *Linum*. Equally, (Li et al., 2010) note that a retrotransposon insertion in the *PvGLO* (*GLOBOSA*) promoter is associated with the homeotic conversion of sepals to petals in *Hose in Hose* mutant *Primula* species and that *PvGLO* is associated with the *S-locus*. It is possible that identification of similar *Linum* mutants may reveal some similar function.

The DESeq2 results showed that *LgSKS1* was differentially expressed at GS1 and GS2 (Figure 2.24). At GS3-4, a series of candidate genes are differentially expressed , however, these are marked as outliers in the program. The most highly differentially expressed genes at all three growth stages were stress response genes. This is perhaps unsurprising as the plants were collected from the wild.

At GS3, *GLO*^{T} is very highly expressed in homostylous *L. strictum*. As discussed in section 2.4.1, this expression may be related to the primers detecting other *GLO*^{T} species, in addition to the thrum-specific gene. For GS2, 95% of the reads had low read counts. The low read counts obtained in so many of the RNA-Seq analyses indicate that there is room for improvement in the RNA-Seq analysis pipeline.

PCA analysis (Figure 2.27a) showed that the points tended to cluster by species, rather than by floral morph identity. Nevertheless, the PCA results are very interesting. PCA is a measure of the overall signal in the data. The close clustering by species, especially at GS1 (Figure 2.27a) indicates that the effect of species must be taken into account when analysing the data. There may also be some evidence of separation of the principal components by floral morph (i.e. pin and thrum). However, there are only relatively few data points per species, and consequently it would be more appropriate to draw conclusions after more extensive sampling has been undertaken. The PCA plots also show that *L. viscosum* and *L. suffruticosum* exhibit large variances in expression. compared to the other heterostylous species. It might be prudent to drop these individuals from the analysis to prevent skewing of the results. The relatively low read counts of these samples may be a factor which have resulted in these differences (refer to Tables 2.12, 2.13, 2.14). Some correlation between the number of read counts and clustering behaviour has been noted; the species which had the lowest mapping percentages, and consequently the lowest read counts was *L.suffruticosum*. *L. suffruticosum* exhibits a 3D reciprocal form of heterostyly not observed in any other tested species. Thus, this species may show variation in its gene expression profile, associated with heterostyly.

At GS1, the two floral morphs of *L. suffruticosum* have clustered together. *L. campanulatum* and *L. narbonense* are also closely located, despite the fact that they are members of different clades. These results might be explained by the efficacy of the mapping of each individual sample, especially given the noise in the data. Equally, however, the clustering may be a reflection of the fact that the expressed polymorphism is very well conserved.

Based on the pattern observed in *Primula* (Li et al., 2016), it was hypothesised that there would be increased expression of candidate genes for heterostyly in the short-styled, thrum morph compared to the long-styled pin morph. DESeq2 analysis revealed differential expression between long styled and short styled floral morphs, at the bud stage (Growth Stage 1). It was expected that the most significant differential expression would occur at GS1, the earliest bud stage. At this stage, many floral homeotic genes are active (e.g. Ryan et al., 2015). In the ABC model of floral development, a combination of B and C genes are responsible for the development of the male sexual organs in the third whorl, and the function of C genes alone is responsible for the development of the carpel (female organs) in the fourth whorl. Petals are developed in the second whorl based on a combination of A and B gene function (Coen & Meyerowitz, 1991).

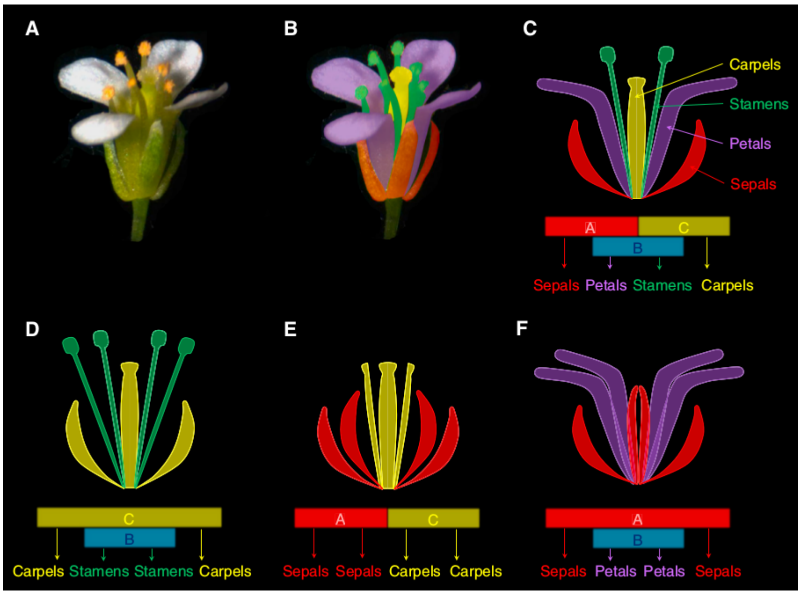

A
B
C
Carpels
Stamens
Petals
Sepals
A
C
B
Sepals Petals Stamens Carpels
D
C
B
Carpels Stamens Stamens Carpels
E
A
C
Sepals Sepals Carpels Carpels
F
A
B
Sepals Petals Petals Sepals

Figure 2.31: Schematic of the ABC model of floral development. Figure adapted from (Irish, 2017).

RNA-Seq analysis holds enormous advantages over conventional expression profiling techniques. It is high-throughput and also enables quantification of transcripts as well as providing overview information of structure and sequence. Currently, there is no accepted or defined method for analysing RNA-Seq data and the production of a coherent differential expression analysis from RNA-Seq count data requires consideration of the interaction of numerous factors, including the hypothesised biological effect size, the number of replicates and the specific interaction used to make differential expression calls (Khang & Lau, 2015). To compound the variation in analysis methods, this study also featured non-conventional sampling, favouring breadth over depth, in an attempt to maximise the number of wild species tested. In future, it might be advisable to limit the number of species examined, in favour of collecting more samples of specific species at each growth stage.

The lack of sufficient replicates in this study has caused fundamental problems as regards data analysis; concerningly, (Khang & Lau, 2015) concluded in their study that when the biological effect size is weak, no meaningful result can be obtained in unreplicated experiments. Additionally, Simon Anders on the Bioconductor support forum stated in response to a questioner: ‘The real reason that you have so few hits is your lack of replicates. In this situation, DESeq reports ... only those hits that are strikingly obvious... You cannot expect to get useful results with a flawed experimental design". This may have contributed to the discrepancies in differentially expressed genes between DESeq2 runs (see Figure 2.23) and to the inconsistently reported number of reads that are too short. Furthermore, the library sizes per sample vary considerably, with a difference of over 100-fold between the lowest read depth and the highest read depth (illustrated in Table 2.5), despite attempts to re-run illumina sequencing on the libraries. The combination of these factors contribute variance that reduces the power of the comparison, although data normalisation using a variance stabilising transformation (vst) and a regularised log transformation (rld) was employed as one defence against the lack of replicates. In addition to the differences in gene expression between the *Linum* species, the absence of technical replicates meant that it was vital for sequence depth to exceed approximately 5 million reads per sample (Sims et al., 2014). Incidentally, above a threshold point (typically 10 million reads) it is considered beneficial to focus on the generation of biological replicates than sequencing depth to ensure the success of RNA-Seq experiments (e.g. Schurch et al., 2016; Liu et al., 2014). In particular, (Schurch et al., 2016) recommend using a minimum of six biological replicates in RNA-Seq experiments. In this experiment, the mean sequence read after the initial sequencing was just short of 3 million reads (2,823,126; refer to Table 2.4), although this value was higher in some better-sequenced individuals.

To attempt to combat the shortcomings in experimental design, a second sequencing run was performed and the two data sets were combined to provide further analysis. These analyses were performed both combining the datasets and using each dataset as a technical replicate. The results obtained between the two methods showed some differences. There are of course vast numbers of different pipelines that could have been employed at every stage of the RNA-Seq analysis pipeline (see Conesa et al. (2016)). Consequently, the selections made during this project are open to significant improvement. One of the perpetual issues caused by the large amount of expression data across the whole transcriptome produced by RNA-Seq analysis is that differences related specifically to heterostyly can be relatively difficult to identify. Despite count normalisation measures, there can often be a large amount of noise which can hide differential expression patterns. Dealing with wild species, and with cross-species maps has led to environmental factors showing the most significant differential expression. Perhaps this is unsurprising, given that the majority of those sampled are wild species. In addition, despite the presence of large amounts of sequence data, the fact that sequencing libraries were not made for all species at all four developmental stages hampered the data analysis. For example, the thrum morph of *L. campanulatum* was sequenced at GS4, but its equivalent at GS3. *L. viscosum* was sampled at GS1 and GS4, but not at GS2 or GS3. In fact, the only heterostylous species for which both floral morphs were sampled at all four growth stages was *L. narbonense*. Furthermore, the small sample sizes on some of the analyses have prompted error messages noting that the dispersion trend was not well captured by the parametric function and a local regression fit was automatically substituted. The non-parametric statistical analysis renders the obtained measures of gene expression differences less statistically powerful. The small sample sizes were a factor of the cost of sequencing, and therefore with more time and funding, it would be possible to collect sufficient replicates, and hopefully draw meaningful results from the conclusion. It would be advisable in future to match up the developmental stages assessed for library preparation with those for the qPCR analysis. Sampling of floral tissues at particular developmental stages in different species is relatively subjective and is dependent on the individual performing the sampling and RNA extraction. The development of a definitive guide, or perhaps a blanket 'days since flowering' would be vastly helpful in terms of gaining an accurate comparison of the transcriptome profile at various growth stages. This suggestion will be discussed in further detail at the conclusion of Chapter 3 (Section 3.4). Ultimately, the experimental design was non-standard and less than ideal. This presented large issues with analysis of the data, a fact which was discovered retroactively. The experimental design was selected as a trade-off to test candidate genes across a larger number of growth stages and species.

Although p values cannot be used as an indicator of statistical significance, they do give an indication of the magnitude of change at species level, and can be interpreted as such with caution. They should not, however, be evaluated in the same way as standard differential expression analyses. Equally, despite this caveat, it is still worthy of argument that different heterostylous species at same growth stage can be treated as biological replicates. The conserved nature of the *'S-locus'* and the fact that the region was hypothesised to be either present or absent depending on the floral morph in question makes the possibility that a signal could have been observed all the more plausible. Of course the data presented has not shown this, however, the method should be recognised as having some validity.

Recent determination of the *L. usitatissimum* plastome found that some RNA-editing sites in flax appear to be unique to the family Linaceae. This discovery opens up the possibility that certain aspects of developmental control, such as the growth of floral organs may be mediated by Linaceae-specific RNA editing sites. As is discussed in further detail in Chapter 3, the PUM^T candidate gene encodes a Pumilio-like transcription factor, which is known (in animals and yeast) to be involved in RNA binding (e.g. Abbasi et al., 2011; Zamore et al., 1997). Perhaps there is a Linaceae-specific role for PUM^T.

As previously discussed, the data included in this chapter cannot be used to infer differential expression, given the lack of biological and technical replicates. However, the variety of species sequenced means that the data could be used to undertake a meaningful screen for sequences from the range of *Linum* species that are related to the key *Primula* and *Fagopyrum* species, including phylogenetic analyses of encoded proteins.

2.4.1 *GLO^T*

The results gained from *GLO^T* posed some of the most interesting questions in this project. *GLO^T* emerged from the literature search as a strong '*S-locus*' candidate. It has been characterised in *Primula* as a duplication of the *GLOBOSA* B-gene, present only in the thrum morph. Glo is known to be the 'G' portion of the *S-locus*, controlling stamen height, based on work in *L. grandiflorum* and *F. esculentum* and therefore also a strong candidate for the control of distyly in *Linum*. No expression was anticipated in the L-morph, corresponding to the hemizygous model found in *Primula* and strongly suggested in *Fagopyrum*. The results were not as anticipated however, and expression was consistently detected in both floral morphs as well as in homostylous species. There are several possible explanations for this obtained result.

Heterostyly has evolved independently in *Linum*, which goes some way towards explaining the lack of consistency in differential expression profile of *GLO^T*. However, one outcome of the DESeq2 analysis was the discovery of *S-ELF3*, a gene in Buckwheat *F. esculentum* which was observed to be similar to *GLO^T*. *S-ELF3* was initially erroneously presumed to be a homologue of *GLO^T* based on a nucleotide sequence alignment, however, personal correspondence with Philip Gilmartin, in addition to an alignment conducted at the amino acid level has since proven this to be incorrect. This finding would have been suggestive of a common mechanism; surprising, given the wide evolutionary distance between *Fagopyrum* and *Linum*, which, according to www.timetree.org diverged approximately 118 MYA. Nevertheless, further investigation would be required to determine whether these mechanisms evolved completely independently of one another.

In an attempt to elucidate the mode of operation of *GLO^T*, a BLAST enquiry of the best contig match for *GLO^T* was conducted against the *L. tenue* transcriptome. The Contig in question, 107032, exhibited an 82.61% sequence match. This represents a high degree of similarity, however, there were no clear observed differences in these sequences which could be attributed to the expected difference between the thrum-specific *GLO^T* and the *GLOBOSA* B-gene from which it was duplicated (Li et al., 2016) (Figure 2.21). This homology analysis was conducted using nucleotide sequence data only. A subsequent analysis carried out with amino acids led to the observation that contigs 107031 and 107033 are in fact equally (if not more) similar to *GLO^T*.

It may be possible to differentiate between several closely related *GLO* genes in the same family by looking at the untranslated region immediately preceding the gene (the 5′ UTR); such methods have proved successful in the literature (e.g. Eveland, McCarty, & Koch, 2008). Using the Interactive Genomics Viewer (IGV) (Robinson et al., 2011) will allow the transcriptome to be viewed graphically, hopefully facilitating the visualisation of small differences between such closely related *GLO* family members. Many genes within a family may have very similar sequences, especially if they have resulted from an evolutionary duplication and subsequent recruitment for an entirely separate function. Another way of determining if the designed primers are amplifying more than one gene family member could be to run GLO^T out on a gel. In this case, if the yield is two separate gene products, Sanger sequencing of each amplified band separately may assist with teasing apart the two variants of GLO^T. The BUSCO gene expression profile (Section 2.2.3) provided some reassurance of adequate transcriptome coverage. Improvements to the unpublished *L. tenue* transcriptome are currently underway by others in the research group at Durham University, and will hopefully clear up uncertainties regarding the adequate isolation of the thrum-specific GLO^T gene product. As regards the analysis,normalisation of the raw data in GLO^T using one of several available methods, including FPKM and RPKM may help in the correct detection of significant differential expression *GLO*. Normalising these counts may provide a more statistically robust method of comparing changes in gene expression between the different floral morphs.

2.4.2 Future Work

To strengthen the reliability of the results produced by the above pipeline, it would be advisable to process a new dataset of additional samples. This would increase the number of replicates for each wild species and would also enable the inclusion of technical replicates, in addition to biological replicates. The production of a new dataset would involve additional fieldwork trips to Spain and extra sequencing resources and would therefore run at an increased cost. A second sequencing run using the same libraries was conducted in response to low sequence yields, particularly in wild heterostylous species such as *L.narbonense*.

Conclusions that could be drawn were impacted by relatively poorly sequenced samples such as *L. narbonense* presenting as outliers (e.g. PCA plots 2.27a, 2.27c and 2.27e). Further insight could also be gained from determining the reason for the poor sequencing of the *L. narbonense* individuals. Knowledge of flaws in the experimental procedure should be beneficial in preventing the reproduction of such flaws in the future by further optimisation of the RNA extraction or RNA-Seq experimental protocols.

New programs and data analysis methods are rapidly becoming available and stricter guidelines for good practices in the analysis of RNA-Seq data have been published even since the beginning of this project (e.g. Conesa et al., 2016). It seems only to be a matter of time before the publication of an accepted set of best-practise guidelines for RNA-Seq analysis practices (similar to the MIQE guidelines (Bustin et al., 2009) for effective qPCR analysis). A new, recommended pipeline involves using fast transcript abundance quantifiers upstream of DESeq2, and then to create gene-level count matrices (Love et al., 2015). This approach corrects for potential changes in gene length across samples, which may, for example, have occurred from differential isoform usage. Salmon (Patro et al., 2017) and kallisto (Bray et al., 2016) are subtantially faster, and require significantly less memory than methods that require the usage of BAM files. It is possible to avoid discarding fragments that align to multiple genes, which is the case in STAR. This increases sensitivity. An overhaul of the pipeline could involve these new methods. The most useful extension to the bioinformatic portion of this work would be to perfect the cross-species transcriptome mapping phase. Cross-species mapping has led to the universally low achieved mapping percentages; losing much data. It is well documented that the ability to detect and quantify rare transcripts is obscured by the dynamic range of mapped reads (Tarazona et al., 2011).

Creating a custom .GTF of the reference file may help the mapping program to accurately place contigs. GTF files Testing analogous mapping tools, such as HISAT2 or the legacy software Bowtie2 may be beneficial. The accuracy of transcript quantification is related to read mapping uncertainty; sequencing error rates, repetitive elements and inaccuracies in transcript annotation (Conesa et al., 2016). All programs employ slightly different algorithms, and therefore a more suitable alternative to STAR may be found.

One way of improving mapping percentages and data gleaned from this RNA-Seq analysis would be to use a *de novo* assembler, such as Trinity (Grabherr et al., 2011) to assemble all 48 RNA-Seq outputs. Assembled sequence data from the 10 wild species could be subjected to multiple sequence alignment and this alignment might ultimately facilitate the identification of orthologous *S-locus* genes between species. This would require significant computational time and power, however, the results obtained would be interesting. A new, updated version of the *L. tenue* transcriptome is currently under construction by Ali Foroozani at Durham University. The release of this improved transcriptome is likely to improve the quality and resolution of the mapping phase significantly. It may be possible, using the updated reference transcriptome file, to distinguish more clearly between genes of the same family, and consequently to solve the mystery which underlies GLO^T expression. A further avenue of research would be to analyse expression of clusters of transcripts via network visualisation using a program such as BioLayout Express (3D) (Theocharidis et al., 2009), or a more modern alternative such as the as the program has now been discontinued. BioLayout(3D) has been superseded by commercial product Graphia Professional (Kajeka, 2014). Alternatively, some other form of functional enrichment analysis, which aims to assign gene function based on the function of similar genes in another species, such as *Arabidopsis* might be insightful as part of further analysis.

It would have been fascinating to consider the role of the *S-locus* in heterostyly from a more evolutionary standpoint. It was beyond the scope of this project to *de novo* assemble wild species sequences, however, although cursory initial investigations proved fruitless, it should be possible using the data at hand to employ a combination of programs such as the mpileup feature of SAMTools (Li et al., 2009) and GATK walker (McKenna et al., 2010) to create a consensus sequence of the mapped sequence reads and subsequently convert the sequence .bam files to .fa files. This should give a rudimentary sequence, thus enabling the building of phylogenetic trees for candidate genes using MEGA7 (Kumar et al., 2015). Consensus sequences were also created to assist with phylogenetic analyses. For each of the candidate genes for which primers could be designed, a consensus of all reads at every contig was produced using Clustal Omega (Sievers et al., 2011).

However, again, time constraints were the most significant limitation in this project and prevented more than a superficial investigation into the creation of candidate-gene based phylogenies. These time constraints prevented further data acquisition meaning that the differential expression analysis was undertaken using a single RNA-Seq dataset and consequently that not all species were represented at all growth stages; complications which have made statistical gene expression analysis impossible. With the publication of the most up-to-date *Linum* phylogeny (Ruiz Martín et al., 2018), perhaps the time is right to undertake a more phylogenetically-led analysis.

Chapter 3

Experimental Analysis of Differential Expression

3.1 Introduction

In addition to the computational analysis, it was hoped to identify '*S-locus*' candidates using experimental techniques. As determined in Chapter 1, one of the main hypotheses was to determine whether the genes from Li et al. (2016)'s paper were more significantly expressed in the short-styled thrum morph than in the reciprocal pin morph in heterostylous *Linum*. This was examined both in greenhouse-grown *L.tenue* and in wild species collected in Spain in 2016. Quantitative PCR (qPCR) provides a method for identifying specific gene transcripts, and for quantifying their expression within a tissue sample. The differential expression analysis which was conducted computationally in Chapter 2 relied on candidate genes found from a comprehensive literature search. These candidate genes were used to design primers to identify orthologues in *Linum* species. In addition to their use in examining raw differential expression data, orthologues were used to design primers that were used to attempt to amplify candidate genes in the laboratory. qPCR is a sensitive technique, and there is therefore a requirement for reference genes.

To accord with the hypothesis of a hemizygous model for the *S-locus*, it would be expected for tested candidate genes to be amplified in the thrum morph, but not the pin morph. In the model presented by Li et al. (2016), the *S-locus* represented a 276 kb region which was entirely absent in the long-styled pin morph. This analysis also hopes to determine whether expression differences exist between different floral developmental stages and between species. It might be that the greatest differential expression is observed in the youngest bud stage, as evidence in *Arabidopsis* has suggested that many floral homeotic genes are active at the earliest stages of floral growth (Ryan et al., 2015). There are, of course, significant morphological differences between the various *Linum* species, and therefore expression of the growth genes, such as *CYP*^{T} is likely to vary across these species. However, as the overall hypothesis is for the *S-locus* to be controlling heterostyly, the candidates for the SPGA genes would be expected to be expressed in thrum but not in pin. Comparisons between homostylous and heterostylous species will also be undertaken. The expectation would be for homostylous species to show no or little expression of these genes if the *S-locus* is absent or degenerate.

As mentioned in the previous chapter, it is acknowledged that without certainty over whether the candidate *Linum* sequences are indeed the orthologues of *Primula* and *Fagopyrum* genes, it is difficult to interpret the data effectively. Consequently there is a likelihood that the extracted sequences represent the combined expression of several closely related genes. The phylogenetic tree presented in Figure 3.4 indicates that there is extensive nucleotide similarity between the thrum-specific candidates. However, it is not entirely certain that the thrum-specific sequence has been successfully identified. The results are presented in this chapter as they were obtained, however, it is vital to reiterate that conclusions cannot be drawn about differential expression between species without confidence that the assessed sequences are indeed the correct ones.

3.2 Methods

3.2.1 Primer Design

Primers were designed from pre-existing *Linum* data prior to obtaining the RNA-Seq output. The consensus sequences of an alignment between heterostylous *Linum tenue* and homostylous *Linum usitatissimum* were used for the design. Table 3.2 contains the primer sequences tested.

FASTA files of the sequences were obtained from the NCBI website. NCBI's command line BLAST toolbox was used to search the transcriptomes both of *Linum usitatissimum* and *L. tenue* for candidate genes. These species alone were selected as their transcriptome sequences were readily available in FASTA format, and represent an example of a distylous species (*L.tenue*) and a homostylous species *L.usitatissimum* BLAST searches were conducted of both the genome and the proteome, in both DNA reading frame directions. The low number of hits obtained upon initial searching, led to the development of this more stringent search strategy.

The *L.tenue* transcriptome used was annotated from previous work in our laboratory (Foroozani, Unpublished Data). The annotations were performed using a combination of tblastx and blastx searches (DNA versus protein) against a series of orthology databases. These databases were FastAnnotator (Chen et al., 2012) and its successor FunctionAnnotator (Chen et al., 2017); the Clusters of Orthologous Groups (COG) tool (Tatusov et al., 1997); the Plant Protein Annotation Suite database (Plant-PrAS) (Kurotani et al., 2015); four specific species from the Plant Transcription Factor Database (PlantTFDB) (Jin et al., 2015): *L. usitatissimum*, TAIR10 (*Arabidopsis*), *Salix purpurea* and *Poplar trichocarpa*; The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000), chosen for its ability to look at pathways and the Conserved Domains Database (CDD) (Marchler-Bauer et al., 2011). Additionally, a reverse position iterative Blastn (RPSTblastn), part of the suite of BLAST+ packages was performed against the CDD database.

The *L. usitatissimum* transcriptome was downloaded from the Phytozome database (Goodstein et al., 2012). Low BLAST hits using the Phytozome *L. usitatissimum* transcriptome prompted attempts to perform the BLAST search using an alternative *L.usitatissimum* transcriptome. Two alternative transcriptomes were investigated, the first from GigaScience Database (Cloutier et al., 2014) and the second, referred to as LUSFL1AD (Venglat et al., 2011) was again sourced from the NCBI website. These alternative transcriptomes proved to be even less effective, as detailed in Table 3.1. The quality of the match was also confirmed by performing tblastx of candidate genes searches against *P. veris* and *Arabidopsis* transcriptomes. Consequently, the Phytozome transcriptome was chosen to generate a consensus sequence for primer design. Candidate genes were selected after conducting a thorough review of the literature. Based on this review, any paper which highlighted genes potentially involved with the development of distyly was investigated and genes which were identified through expression or molecular biology were selected.

Local tblastx searches were conducted of sequences of other heterostylous species from the literature, retrieved from the NCBI against the two reference transcriptomes. A tblastx query searches translated nucleotide databases using a translated nucleotide query. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. tblastx therefore enables the identification of very distant relationships between nucleotide sequences (BLAST® Command Line Applications User Manual).

% Identical Matches	Alignment Length	Mismatches	No.gap openings	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
30.23	43	30	0	199	71	238	110	0.28	29
42.31	26	15	0	250	173	127	50	0.27	29.5

Table 3.1: Alternative *L. usitatissimum* transcriptomes tested. The top sequence is GigaDB and the bottom is LUSFL1AD

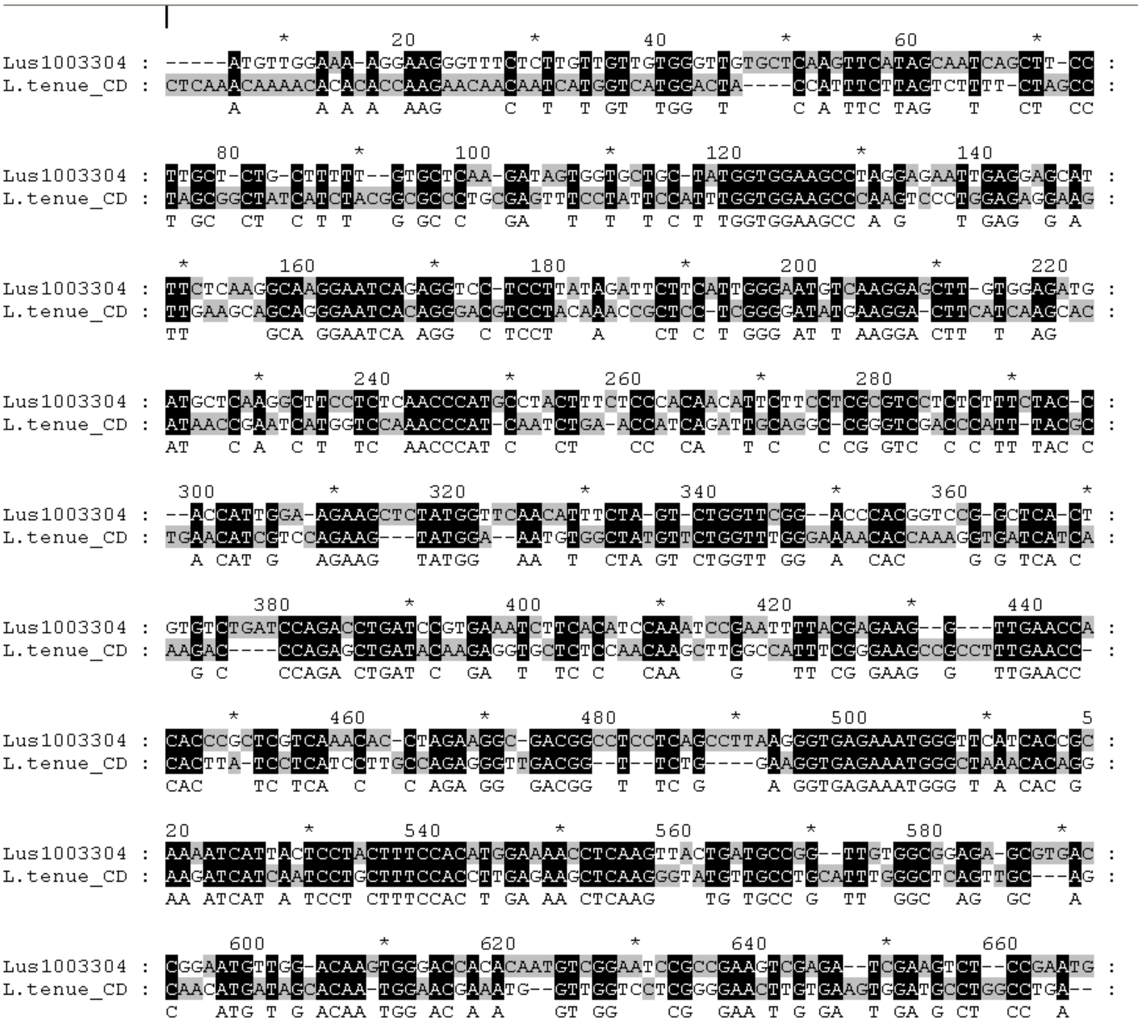
```
tblastx \
-query ~/Candidate_Genes/S-ELF3/S-ELF3.fa \
-db ~/Databases/Ltenue/Ltenue_BLASTdb/Ltenue_BLASTdb/ \
-out ~/Primer_Design/Candidate_Genes/S-ELF3/S-ELF3_Ltenue.outfmt6 \
-outfmt 6 -num_threads 8
```

Figure 3.1: Minimal example of a tblastx search querying *S-ELF3* against the *L.tenue* transcriptome.

A series of scripts were written in Python 3.0 to identify the sequences at contigs with the most significant E values, a value indicating the number of hits expected to be seen by chance. The first script, shown in Appendix B, extracted the accession number of the top BLAST hits and saved them to a file. The second script, which was modified from FastaGrab.py (Foroozani, 2014) extracted the contig which matched the accession numbers of the top BLAST hits to the *L. tenue* transcriptome (Appendix B). Clustal Omega (Sievers et al., 2011) was used to align each sequence against *L. usitatissimum* and *L. tenue*. Other alignment softwares, using different matching algorithms were also tested, including T-COFFEE (Notredame et al., 2000). The differences in their various alignment outputs are detailed in Figure C.1.

The example blastx search depicted in Figure 3.1 queries *F. esculentum* candidate gene *S-ELF3* against a BLAST database created from the *L. tenue* transcriptome. It uses -outfmt 6, which outputs the top BLAST hits in a table. Other information shown in the table reflects that shown in Table 3.1 and includes the percentage of identical matches, the alignment length, the number of mismatches, the number of gap openings, the start of the alignment in the query, the end of the alignment in the query, the E-score and the bit score; the final two being indicators of closeness.

GeneDoc software (Nicholas & Nicholas, 1997) was used to check and, where necessary, manually improve the sequence alignments from Clustal Omega. This improvement was deemed necessary where the differences between species led to poor alignment.Once a satisfactory gene alignment had been achieved, GeneDoc was also used to calculate and create a consensus sequence across the three species. Another Python script was written to replace all gaps in the consensus sequence with 'N' to facilitate insertion into the Primer3 web-based software (http://primer3.ut.ee/). Primers were designed using the consensus sequences of *L. usitatissimum* and *L. tenue*. Designing around the consensus of *L. grandiflorum* was also considered as public sequence resources exist



```
                      *        20         *        40         *        60         *
Lus1003304 : -----ATGTTGGAAA-AGGAAGGGTTTCTCTTGTTGTTGTGGGTTGTGCTCAAGTTCATAGCAATCAGCTT-CC :
L.tenue_CD : CTCAAACAAAACACACACCAAGAACAACAATCATGGTCATGGACTA----CCATTTCTTAGTCTTTT-CTAGCC :
                  A      A A A  AAG     C  T  T GT  TGG  T     C A TTC TAG   T   CT  CC

                 80         *       100         *       120         *       140
Lus1003304 : TTGCT-CTG-CTTTTT--GTGCTCAA-GATAGTGGTGCTGC-TATGGTGGAAGCCTAGGAGAATTGAGGAGCAT :
L.tenue_CD : TAGCGGCTATCATCTACGGCGCCCTGCGAGTTTCCTATTCCATTTGGTGGAAGCCCAAGTCCCTGGAGAGGAAG :
             T GC  CT  C T T   G GC C   GA   T  T  T C T TGGTGGAAGCC A G    T GAG  G A

               *       160         *       180         *       200         *       220
Lus1003304 : TTCTCAAGGCAAGGAATCAGAGGTCC-TCCTTATAGATTCTTCATTGGGAATGTCAAGGAGCTT-GTGGAGATG :
L.tenue_CD : TTGAAGCAGCAGGGAATCACAGGGACGTCCTACAAACCGCTCC-TCGGGGATATGAAGGA-CTTCATCAAGCAC :
             TT      GCA GGAATCA AGG  C TCCT   A    CT C T GGG AT T AAGGA CTT  T  AG

                    *       240         *       260         *       280         *
Lus1003304 : ATGCTCAAGGCTTCCTCTCAACCCATGCCTACTTTCTCCCACAACATTCTTCCTCGCGTCCTCTCTTTCTAC-C :
L.tenue_CD : ATAACCGAATCATGGTCCAAACCCAT-CAATCTGA-ACCATCAGATTGCAGGC-CGGGTCGACCCATT-TACGC :
             AT   C A  C T  TC  AACCCAT C   CT    CC  CA   T C   C CG GTC  C C TT TAC C

              300         *       320         *       340         *       360         *
Lus1003304 : --ACCATTGGA-AGAAGCTCTATGGTTCAACATTTCTA-GT-CTGGTTCGG--ACCCACGGTCCG-GCTCA-CT :
L.tenue_CD : TGAACATCGTCCAGAAG---TATGGA--AATGTGGCTATGTTCTGGTTTGGGAAAACACCAAAGGTGATCATCA :
               A CAT G   AGAAG   TATGG   AA  T  CTA GT CTGGTT GG  A  CAC     G G TCA C

                    380         *       400         *       420         *       440
Lus1003304 : GTGTCTGATCCAGACCTGATCCGTGAAATCTTCACATCCAAATCCGAATTTTACGAGAAG--G---TTGAACCA :
L.tenue_CD : AAGAC----CCAGAGCTGATACAAGAGGTGCTCTCCAACAAGCTTGGCCATTTCGGGAAGCCGCCTTTGAACC- :
               G C    CCAGA CTGAT C  GA  T  TC C   CAA    G    TT CG GAAG  G   TTGAACC

                  *       460         *       480         *       500         *       5
Lus1003304 : CACCCGCTCGTCAAACAC-CTAGAAGGC-GACGGCCTCCTCAGCCTTAAGGGTGAGAAATGGGTTCATCACCGC :
L.tenue_CD : CACTTA-TCCTCATCCTTGCCAGAGGGTTGACGG--T--TCTG----GAAGGTGAGAAATGGGCTAAACACAGG :
             CAC    TC TCA  C   C AGA GG  GACGG  T  TC G     A GGTGAGAAATGGG T A CAC G

             20         *       540         *       560         *       580         *
Lus1003304 : AAAATCATTACTCCTACTTTCCACATGGAAAACCTCAAGTTACTGATGCCGG--TTGTGGCGGAGA-GCGTGAC :
L.tenue_CD : AAGATCATCAATCCTGCTTTCCACCTTGAGAAGCTCAAGGGTATGTTGCCTGCATTTGGGCTCAGTTGC---AG :
             AA ATCAT A TCCT CTTTCCAC T GA AA CTCAAG    TG TGCC G  TT  GGC  AG  GC   A

                  600         *       620         *       640         *       660
Lus1003304 : CGGAATGTTGG-ACAAGTGGGACCACACAATGTCGGAATCCGCCGAAGTCGAGA--TCGAAGTCT--CCGAATG :
L.tenue_CD : CAACATGATAGCACAA-TGGAACGAAATG--GTTGGTCCTCGGGGAACTTGTGAAGTGGATGCCTGGCCTGA-- :
             C   ATG T G ACAA TGG AC A A    GT GG    CG  GAA T G GA  T GA G CT  CC  A
```

Figure 3.2: An example alignment of *CYP734A50* produced using GeneDoc Software (Nicholas & Nicholas, 1997).

for this species, however an annotated *L. grandiflorum* transcriptome in FASTA format could not be located. Screenshots of the GeneDoc alignments show that sequences from the multiple species were quite different, and that often a satisfactory alignment for primer design could not be reached for all candidate genes (Figure 3.2).

To maximise the success of primer binding, the parameters in Primer3 were set as described below. Mixed initial results meant that the settings were relaxed to improve the likelihood of primer selection. From the Primer3 interface, the desired parts of the gene were marked using square brackets. Patchy areas of the sequence could be ignored using angle brackets (< >). The complementary strand was also taken and the accepted GC range was stretched to 20-80% from its optimum setting of 40-60%. 'N's were inserted to replace gaps in the sequence; the maximum number of allowed Ns was increased to 4. Primers were designed to have a minimum length of 18 bases and a maximum length of 22 bases. This is to ensure adequate specificity whilst the primers are short enough to bind easily at the required annealing temperature (Biosoft, n.d.).

The parameters that were controlled include melting temperature; the temperature at which the DNA duplex dissociates to become single-stranded, which is indicative of duplex stability. Melting temperature was kept to be between 52-58°C. Temperatures above 65°C have a tendency for secondary annealing. The GC content tends to be an accurate indicator of the melting temperature (T_m). Higher GC content equates to a greater number of bases held together by 3H bonds. The presence of a GC clump (GC bases within the last 5 bases of the 3' end of the primer) helps to promote specific binding. Care was taken not to have more than 3 'N's in the last 5 bases. Repeats (e.g. ATATATAT) and runs of the same base (e.g. AGCGGGGATGGG) were avoided in order to avoid the creation of primer secondary structures.

Of the original 18 genes identified in the literature search, it was only possible to design primers for 11 of them. The efficacy of each of the designed primers was determined using PCR of a cDNA extraction of *L. tenue* that was then run on a 1% agarose gel electrophoresis stained with Midori green dye. The presence of bands of appropriate size range on the gel, indicated that the primers had successfully amplified the correct region. Some primers, e.g. *LgAP1*, *PvSLL1* and *TSS1* produced only blurred bands of length approximately 100bp. These are indicative of primer-dimers, and suggest that the primers were unable to amplify the correct region of cDNA. Thus, such primers were discounted for the qPCR process. BLAST searches were conducted of the designed forward and reverse primers against the *L. tenue* transcriptome. This step was performed in an attempt to ensure that non-target transcripts were not also being amplified.

Primer design for qPCR reference genes was conducted using the same method as above. Potential genes were selected based on candidates from a study conducted by Huis et al. (2010). From this paper, promising looking reference genes were selected on the basis of their success in amplifying *L. usitatissimum*, particularly in apical tissue. The 4 best candidates, *GAPDH* (the top performer against every metric), *ETIF3H*, *ETIF3E* and *UBI2* were selected and BLAST searched against the *L. usitatissimum* and *L. tenue* transcriptomes. The actual primers from the paper were also investigated. *GAPDH* was selected as the most effective reference gene following PCR amplification of the reference gene candidates in *L. tenue*.

Table 3.2: Candidate genes determined from a literature search, for which qPCR primers were designed.

Candidate Gene	Primers	Reference
PveGLO2	Forward: 5′-NCTGCAGCCCNTCCACCAC-3′ Reverse: 5′-NCCAGACTGCTTGTGNTACTTCT-3′	Nowak et al. (2015)
PvGLO	Primers could not be designed	Nowak et al. (2015)
S-ELF3	Forward:5′-TGAGTTGCATAGACTNATCAAGG-3′ Reverse: 5′-CTCTGCAGAACCTTCCANCTTAT-3′	Yasui et al. (2012)
CYP734A50	Primers could not be designed	Huu et al. (2016)
TSS1	Primers could not be designed	Ushijima et al. (2012)
TkNACE	Primers could not be designed	Labonne et al. (2009)
TkST1	Forward:5′-ACANGCTGGTGTCGATGTGG-3′ Reverse:5′-AGCCTGTNNAANCTGCTCCNCCA-3′	Labonne et al. (2009)
TsRETRO	Primers could not be designed	Labonne et al. (2009)
PvSLL1	Primers could not be designed	Labonne et al. (2009)
LgMYB21	Forward:5′-ACGTNAGGNGAGGNAANATCACT-3′ Reverse:5′-TCCGTCCTTCCNGGNAGATGCT-3′	Ushijima et al. (2012)
LgAP1	Forward: 5′-NTGAAGCCGGTGAATACGT-3′ Reverse:5′-CGCAGGNAGGAAGGANACNGTC-3′	Ushijima et al. (2012)
LgSKS1	Forward: 5′-TGGAACCTNACNGCCAGCGC-3′ Reverse:5′-GAACTCCGCCCAAGAAAAGG-3′	Ushijima et al. (2012)
CCM^{T}	Primers could not be designed	Li et al. (2016)
CYP^{T}	Forward:5′-TGAAGATTTCGCGGATGAGG-3′ Reverse:5′-GAACTCCGCCCAAGAAAAGG-3′	Li et al. (2016)
GLO^{T}	Forward:5′-GATGAAGTTGAGGCGCTTGT-3′ Reverse:5′-GCAGATTGAACTCAGGCACC-3′	Li et al. (2016)
KFB^{T}	Forward:5′-CTTCNCACTCGTCNCGCTCCTGG-3′ Reverse:5′-GCAGATTGAACTCAGGCACC-3′	Li et al. (2016)
PUM^{T}	Forward:5′-TGAGCCTGATGTGGATGGGT-3′ Reverse:5′-GGAGCTTGTGGAGATGATGC-3′	Li et al. (2016)
TPP1	Forward:5′-GGGCCAAGNTCAGGAGAAAGC-3′ Reverse:5′-ATGCCGGAGNCNCGGTCGTTC-3′	Ushijima et al. (2012)

Reference Gene	Primers	Reference
ETIF3H	Forward: 5′-CCATCAAGATCAAGCCAGGG-3′ Reverse: 5′-CGGTCATAGTCACACTGGGT-3′	Huis et al. (2010)
ETIF3E	Forward: 5′-GGGAAAGTTGGCTGCAGAG-3′ Reverse: 5′-ATGAAGAGACTCCAATGCA-3′	Huis et al. (2010)
UBI2	Forward: 5′-TTCGTGAAAACCCTAACCGG-3′ Reverse: 5′-AGGTGNAGNGTNGACTCCTTCT-3′	Huis et al. (2010)
GAPDH	Forward: 5′-ACNACCAACTGCCTTGCTCC-3′ Reverse: 5′-ACGGTGGTCATNAGNCCCTC-3′	Huis et al. (2010)

3.2.2 Post-analysis homology searches at the amino acid level

As has been addressed in the introduction and aims sections, it is appreciated with hindsight that the differential expression analysis comparisons used sequences obtained from nucleotide BLAST searches and resultantly may not have been the closest match to *Linum*. Therefore these sequences may not have been those that they were assumed to be. In an effort to retrospectively address this problem, searches against *L. tenue* were conducted at the amino acid level for four of the most promising candidate genes: CYP^T, PUM^T, GLO^T and *S-ELF3*. Such homology searches revealed many potential hits and have produced convoluted phylogenetic trees (see Figure 3.4 as an example). For CYP^T there were 60 unique sequences with an e score of 5.09E-39 or lower, and over 900 hits to short regions of the same contig in different locations or in alternative reading frames. GLO^T and PUM^T had fewer top matches, however, none definitively showed that an orthologue had been found (Table 3.3 gives the top hits for PUM^T.). This raises concerns both that the wrong genes were being amplified in this study and/or that the primers were in fact locating DNA fragments from a series of closely related genes, rather than the hypothesised heterostyly-specific ones. The phylogenetic tree for CYP^T presented in Figure 3.4 shows just how distant CYP^T is from its most closely matched contigs in *L. tenue*. It is important to note that a warning displayed when computing gene trees that STOP codons were found in the translated regions also calls into doubt the validity of these selections.

The homology searches did suggest that the correct contig was identified at the primer design stage for CYP^T and PUM^T. In CYP^T, the same contig, 86432, is pulled out as the top hit by BLAST when searching amino acid sequences as when searching nucleotide sequences, thus implying that the closest match was indeed extracted during the initial nucleotide analyses. For PUM^T again, the top amino acid hit (31347) matched that of the original investigation confirming that in all likelihood the closest match was selected. On the other hand, amino acid BLAST searches of GLO^T found that contig 107032, the most highly matched hit from both RNA-Seq analysis and nucleotide searches was not the closest match in an amino acid investigation. Instead, the two most similar regions were contig 107033 (e value: 2.79E-45; bit score: 153)and contig 107031 (e value: 7.98E-43; bit score: 151). Observation of the phylogenetic tree presented in Figure 3.3 shows that contigs 107031 and 107033 are further removed from the other *Linum* sequence. The full blast search is recorded in Table D.1 in the appendix and demonstrates the sheer number of closely related sequences and consequent potential gene families. This finding suggests that the thrum-specific *GLO*^{T} was not correctly identified and furthermore, that closely related gene families may have been amplified by the primers in addition to *GLO*^{T}. This is discussed in more detail in the discussion, Section 3.4. Despite attempting with several sequences downloaded from the NCBI website, no amino acid sequence for S-ELF was successfully compared against the *L. tenue* transcriptome. Thus, no conclusions about the identity of the sequence could be drawn. The amino acid blast hits for *CYP*^{T} were notably higher than for *PUM*^{T} or *GLO*^{T}, meaning that the results obtained for *CYP*^{T} can be viewed with the greatest certainty.

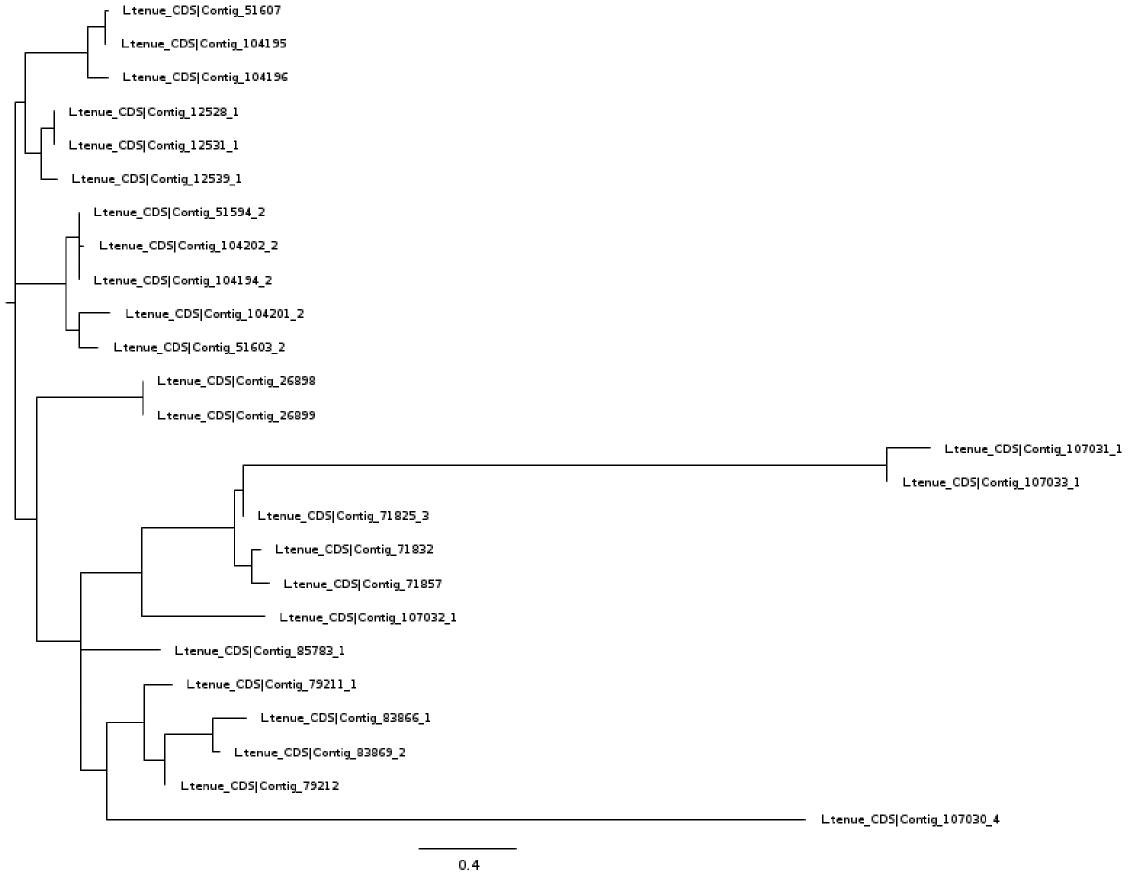

Ltenue_CDS|Contig_51607
Ltenue_CDS|Contig_104195
Ltenue_CDS|Contig_104196
Ltenue_CDS|Contig_12528_1
Ltenue_CDS|Contig_12531_1
Ltenue_CDS|Contig_12539_1
Ltenue_CDS|Contig_51594_2
Ltenue_CDS|Contig_104202_2
Ltenue_CDS|Contig_104194_2
Ltenue_CDS|Contig_104201_2
Ltenue_CDS|Contig_51603_2
Ltenue_CDS|Contig_26898
Ltenue_CDS|Contig_26899
Ltenue_CDS|Contig_107031_1
Ltenue_CDS|Contig_107033_1
Ltenue_CDS|Contig_71825_3
Ltenue_CDS|Contig_71832
Ltenue_CDS|Contig_71857
Ltenue_CDS|Contig_107032_1
Ltenue_CDS|Contig_85783_1
Ltenue_CDS|Contig_79211_1
Ltenue_CDS|Contig_83866_1
Ltenue_CDS|Contig_83869_2
Ltenue_CDS|Contig_79212
Ltenue_CDS|Contig_107030_4
0.4

Figure 3.3: Phylogenetic tree showing a line up of the closest blastx hits for *GLO*^{T} against the *L. tenue* transcriptome. The tree was created using MEGA 7 and visualised using Figtree v. 1.4.4 (Rambaut, 2009; http://tree.bio.ed.ac.uk/software/figtree/).

Table 3.3: Top amino acid blast hits for PUM^T.The numbers in brackets refer to the reading frame.

	% Iden- tical Matches	Length	Mis- matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 31347 (2)	45.556	90	49	0	1361	1630	93	182	3.03e-30	93.2
Contig 121559 (3)	38.710	93	57	0	1352	1630	75	167	1.25e-23	83.6
Contig 121559 (3)	40.351	57	34	0	1720	1890	165	221	1.25e-23	51.2
Contig 31349 (2)	55.914	93	41	0	3346	3624	99	191	5.91e-22	100
Contig 121562 (3)	41.667	60	35	0	1711	1890	68	127	4.35e-19	53.5
Contig 31348 (3)	58.696	46	19	0	3463	3600	200	245	1.01e-07	60.5
Contig 42086	27.174	92	64	1	1355	1630	783	871	1.36e-07	46.6
Contig 129748	33.333	87	57	1	3340	3600	120	205	1.52e-07	59.7
Contig 61874	26.582	79	58	0	1355	1591	487	565	4.17e-07	43.5
Contig 31349 (6)	48.276	87	45	0	3605	3345	38	124	1.50e-06	55.1
Contig 42072	25.000	92	66	1	1355	1630	664	752	8.42e-06	38.1

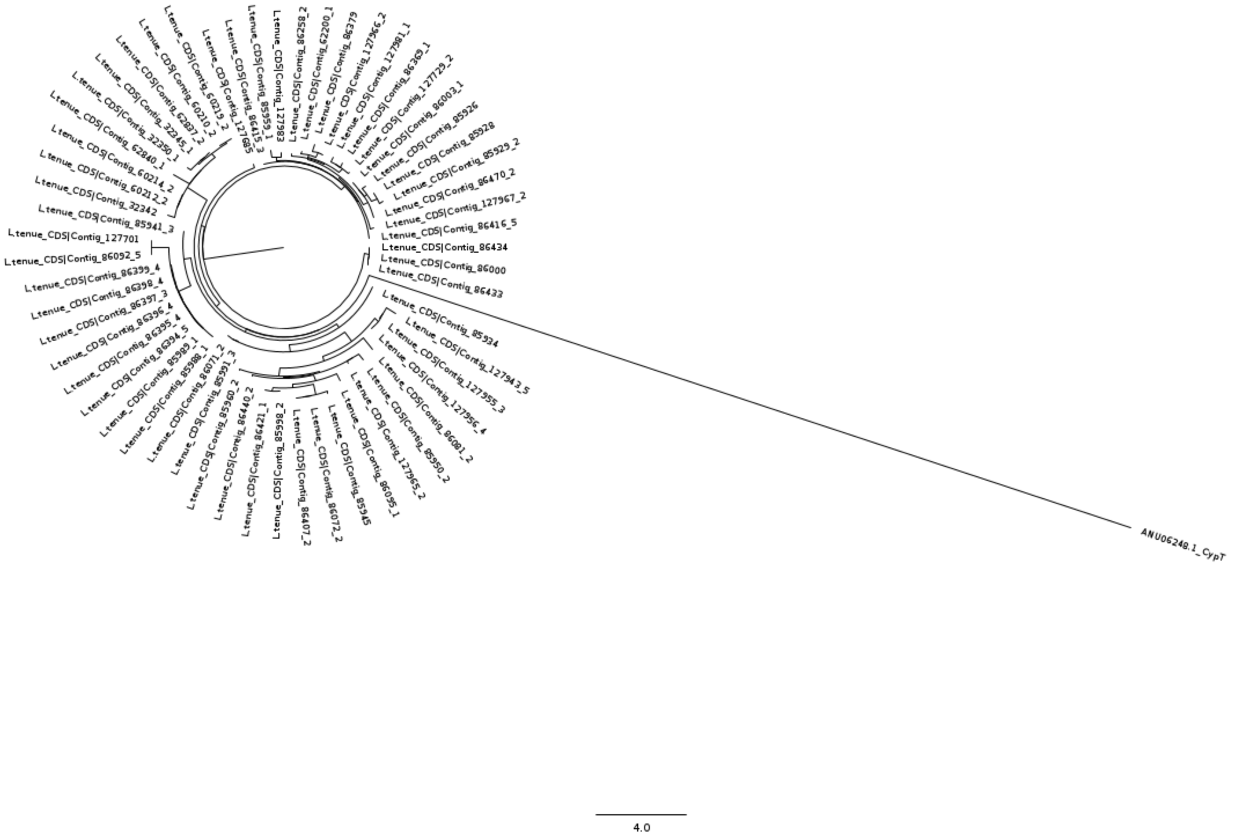

L.tenue_CDS|Contig_85941_3
L.tenue_CDS|Contig_32342
L.tenue_CDS|Contig_60212_2
L.tenue_CDS|Contig_60214_2
L.tenue_CDS|Contig_62840_1
L.tenue_CDS|Contig_32350_1
L.tenue_CDS|Contig_32245_1
L.tenue_CDS|Contig_62837_2
L.tenue_CDS|Contig_60210_2
L.tenue_CDS|Contig_60219_2
L.tenue_CDS|Contig_127685
L.tenue_CDS|Contig_86415_3
L.tenue_CDS|Contig_85959_1
L.tenue_CDS|Contig_127983
L.tenue_CDS|Contig_86258_2
L.tenue_CDS|Contig_62200_1
L.tenue_CDS|Contig_86379
L.tenue_CDS|Contig_127966_2
L.tenue_CDS|Contig_127981_1
L.tenue_CDS|Contig_86369_1
L.tenue_CDS|Contig_127729_2
L.tenue_CDS|Contig_86003_1
L.tenue_CDS|Contig_85926
L.tenue_CDS|Contig_85928
L.tenue_CDS|Contig_85929_2
L.tenue_CDS|Contig_86470_2
L.tenue_CDS|Contig_127967_2
L.tenue_CDS|Contig_86416_5
L.tenue_CDS|Contig_86434
L.tenue_CDS|Contig_86000
L.tenue_CDS|Contig_86433
L.tenue_CDS|Contig_85934
L.tenue_CDS|Contig_127943_5
L.tenue_CDS|Contig_127955_3
L.tenue_CDS|Contig_127956_4
L.tenue_CDS|Contig_86081_2
L.tenue_CDS|Contig_85950_2
L.tenue_CDS|Contig_127965_2
L.tenue_CDS|Contig_86095_1
L.tenue_CDS|Contig_85945
L.tenue_CDS|Contig_86072_2
L.tenue_CDS|Contig_86407_2
L.tenue_CDS|Contig_85998_2
L.tenue_CDS|Contig_86421_1
L.tenue_CDS|Contig_86440_2
L.tenue_CDS|Contig_85960_2
L.tenue_CDS|Contig_85991_3
L.tenue_CDS|Contig_86071_2
L.tenue_CDS|Contig_85988_1
L.tenue_CDS|Contig_85989_1
L.tenue_CDS|Contig_86394_5
L.tenue_CDS|Contig_86395_4
L.tenue_CDS|Contig_86396_4
L.tenue_CDS|Contig_86397_3
L.tenue_CDS|Contig_86398_4
L.tenue_CDS|Contig_86399_4
L.tenue_CDS|Contig_86092_5
L.tenue_CDS|Contig_127701
ANU06248.1_CypT
4.0

Figure 3.4: Phylogenetic tree showing a line up of the closest blastx hits for CYP^T against the *L. tenue* transcriptome. The tree was created using MEGA 7 and visualised using Figtree v. 1.4.4 (Rambaut, 2009; http://tree.bio.ed.ac.uk/software/figtree/).

3.2.3 Post-analysis determination of primer binding efficiency for loci in different species

It is important to determine whether the designed primers were in fact binding effectively to loci in different species. This was especially pertinent given that the primers used for this project were designed from a consensus of only two *Linum* species, *L. usitatissimum* and *L. tenue*, before the ten wild species were sequenced.

The definitive method to determine whether the primers were effectively binding to loci in all species during the qPCR process would be to produce a line-up of all species tested and to show the exact primer matches and mismatches at each of the loci. Unfortunately, it was not possible to complete this as part of this Master's project as no annotated transcriptomes of any of the wild species exist (published or unpublished) to the author's knowledge. Complete annotation of so many transcriptomes could take several years of work, and thus is recommended as a future direction for research but was outside the scope of this analysis. In an attempt to create such an alignment for this thesis, preliminary work was conducted to *de novo* assemble the RNA-Seq data from the 10 wild species (see Section 2.2.3 for further details on the Trinity assembly). However, the RNA-Seq data was found to be too sparse to produce any useful alignment.

```
#A script to prepare primers for blast searching against contigs.
#Script puts in forward primer and reverse-complemented reverse primer,
#separated by 20 Ns

from sys import argv

script, forward_primer, reverse_primer, outfile = argv

textfile = open(outfile, ’w’)
for char in reverse_primer:
char.replace("A", "T")
char.replace("T", "A")
char.replace("C", "G")
char.replace("G", "C")

new_reverse = reverse_primer[::-1]

textfile.write(">" + outfile[:10] + "\n")
print textfile.write(forward_primer + 20 *"N" + new_reverse)

textfile.close()
```

Figure 3.5: A script written to prepare primers for BLAST searching against *L.tenue* and *L.usitatissimum*

.

In order to confirm whether the primers were effective for the two initial species, a short script was written to prepare primers for BLAST searching against an *L. tenue* BLAST database. The script creates a sequence comprised of the forward primer and the reverse-complemented reverse primer, separated by 20 Ns. Primer BLAST queries were created using this method for the following candidate genes: *CYP^T*, *PUM^T*, *GLO^T*, *S-ELF3*, *PveGLO2*, *TPP1*, *LgAP1*, *LgMYB21*,*LgSKS1* and *TkST1*. The restriction to *L. tenue* was caused by doubts over the quality of the *L. usitatissimum* transcriptome and amino acid sequence (see the beginning of Section 3.2.1). This naturally reduces the results to efficacy in one species, however, it is thought that the results are indicative of other species. Unfortunately, searching these combined sequences against the *L.tenue* transcriptome did not reveal any matches. When a database was created solely from the best hits for each gene, and a second BLAST search conducted guided by these initial results, the primer for *GLO*^{T} did align to *L. tenue* contig 107032 (Table 3.4). However, it is highly likely that the primers were not effectively binding to loci during the PCR amplification process; evidenced by the high E value and low bit scores shown in Table 3.4. This is very disappointing; if the primers were not binding to *L. tenue* sequences then it becomes even less likely that they would amplify loci from wild species. Such results are, however, understandable, given that the primers were designed from a consensus of *L. usitatissimum* and *L. tenue* transcriptomes before the wild *Linum* species were sequenced. Ineffective primer binding means that no conclusions can be drawn about relative expression in this study. The results will still be presented and the shortcomings of the study discussed.

Table 3.4: Top amino acid BLAST hits for the GLO^T primer, when blasted against a database of close search results. The low e values indicate the low quality of the matches.

	% Identical Matches	Length	Mismatches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 107032	100.000	6	0	0	19	2	1178	1195	0.035	17.8
Contig 107032	100.000	7	0	0	1	21	1196	1176	0.043	17.6
Contig 107032	100.000	6	0	0	18	1	1179	1196	0.080	16.6
Contig 107032	100.000	6	0	0	2	19	1195	1178	0.15	15.7
Contig 107032	100.000	6	0	0	20	3	1177	1194	0.29	14.8
Contig 107032	100.000	6	0	0	3	20	1194	1177	0.39	14.4

3.2.4 RNA extraction and cDNA synthesis

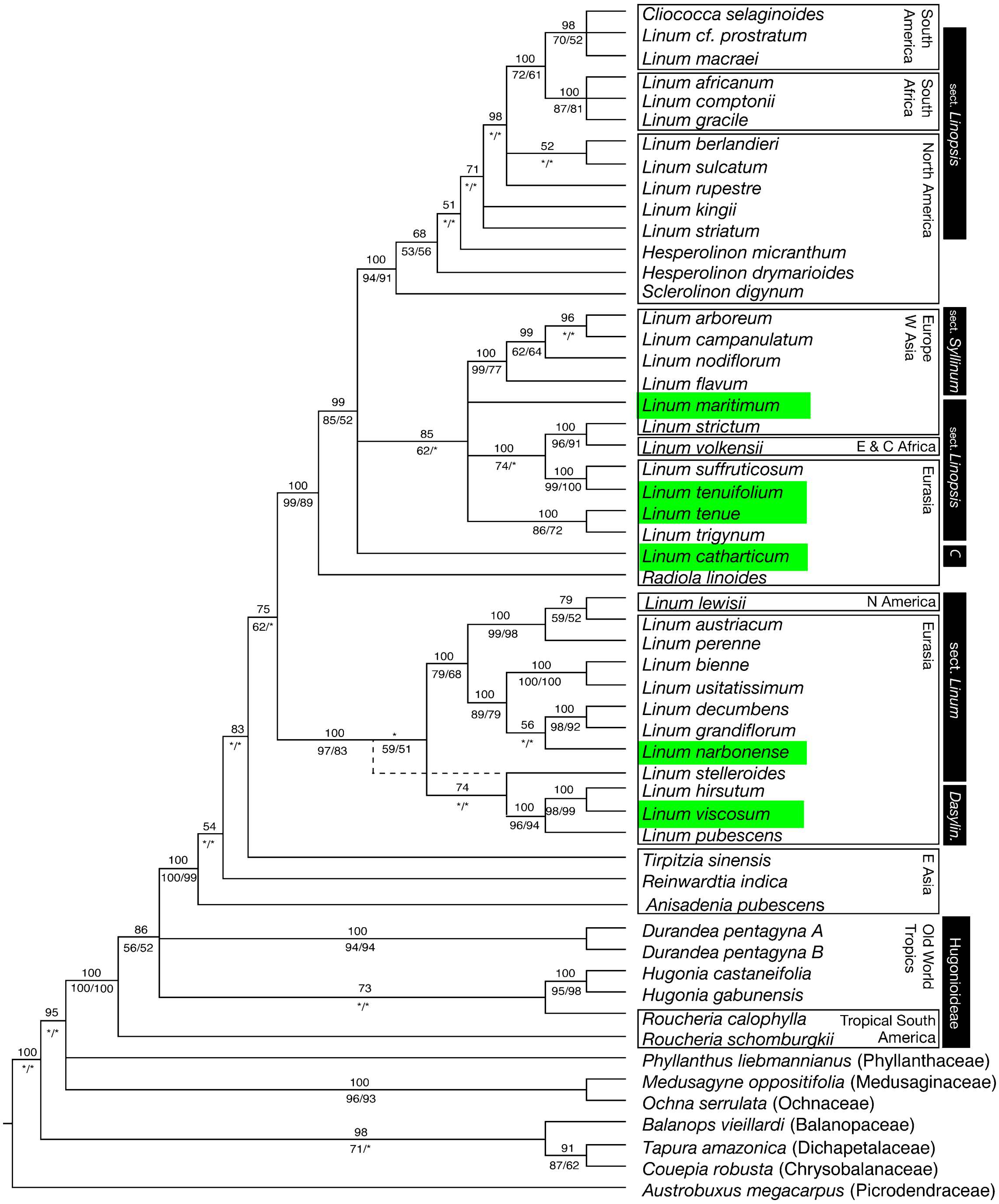


Figure 3.6: Species selected for RNA-Seq analysis, highlighted in green on the most recent published *Linum* Phylogeny (adapted from Ruiz Martín et al. (2018))(with the exception of *L. setaceum*, which is not depicted as part of this phylogeny.)

RNA was extracted from the fresh floral tissue of *Linum tenue* grown in glasshouses at Durham University (Durham, UK) under a 16h day and 8h night cycle. Tissue was sampled at three separate growth stages, detailed in Table 3.6. *L. tenue* tissue was sampled fresh from the glasshouse and collected in 1.5 ml TubeOne® microcentrifuge tubes (STARLAB, Milton Keynes, UK). These tubes were immediately flash-frozen in liquid nitrogen. Tissue samples of 9 wild species (see Figure 3.6), collected from field sites in Andalucia and the Pyranees in Summer 2016. These field sites were as listed in 2.2. The tissue collected was sampled fresh into 1.5mL RNAlater solution at ambient temperature and was later stored at -80C. Samples were all collected between 2 and 4 pm to control for expression differences related to the time of day. Fresh and stored samples were extracted using the same protocol. Three buds from the same plant at each growth stage were placed into 1.5 ml TubeOne® microcentrifuge tubes (STARLAB, Milton Keynes, UK). All floral tissues were flash-frozen using liquid nitrogen and subsequently crushed using micropestles. A Disruptor Genie (Scientific Industries, New York, USA) tissue lyser, equipped with two 1 mm metal beads was used to improve sample grinding. The specimens were crushed in 1ml Invitrogen™ TRIzol™, before being spun in the tissue lyser for 3 x 2 minute periods, leaving time . Samples were left for 5 minutes to permit complete dissociation of the nucleopore complex. 200μl of chloroform was added to each sample before samples were vortexed. Samples were centrifuged for 15 minutes at 12000 rpm (13201 g) at a temperature of 4°C to produce 3 liquid phases. The centrifuge model was Allegra™X-22R Centrifuge (Beckman Coulter, High Wycombe, UK) and the rotor model was F2402H. RPM to G conversions were performed using the Gene Infinity calculator (http://www.geneinfinity.org/sp/sp_rotor.html). For each sample, the top, aqueous, phase containing the RNA was removed and placed into a separate 1.5ml tube. To minimise the possibility of contamination, a second chloroform wash was conducted, using 500μl chloroform, and centrifugation at 14000 rpm (17968 g) for 2 minutes. The RNA was precipitated with the addition of 500μl isopropanol and centrifuged for 10 minutes at 12000 rpm (13201 g) to produce a pellet. The pellets were washed in 75% ethanol (stored at -4°C), centrifuged at 7000 rpm (4492 g) for 5 minutes, before being left beneath the flow-hood to dry. The pellets were resuspended in 50μl of RNase-free water (Promega, Southampton, UK). Pellets were placed onto a heat block at 60°C for 10 minutes, in order to ensure resuspension of the pellet. The extracted RNA was subsequently frozen, or was immediately quantified and treated with DNAfree. RNA was stored under ice during the quantification, DNAse treatment and cDNA synthesis steps in order to minimise degradation and maximise yield.

For each species, 3 biological replicates of each growth stage were performed. To ensure adequate RNA concentrations from the extraction at GS1, 3 individual buds from the same plant contributed to one GS1 sample. The RNA quantities from the freezer sample led to 3 individual buds from the same plant from each growth stage contributing to a single extraction stage. Three biological replicates of each *L. tenue* floral morph were performed at each growth stage.

Table 3.5: cDNA Synthesis Program

	Step 1	Step 2	Step 3	Step 4
Temperature	25°C	37°C	85°C	4 °C
Time	10 min	120 min	5s	∞

Table 3.6: Qualitative Description of Floral Morph Growth Stages. GS3 and GS4 were usually combined for qPCR experiments. Table repeated from Section 2.2 for ease of reference.

Growth Stage	Description
GS1	Immature buds.
GS2	Maturing buds: beginning to open, petals visible through the top of the bud.
GS3	Flower closed, but has completed growth. Petal length now exceeds that of the sepal.
GS4	Open flowers.

RNA quantification and quality assurance was performed using both a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA) for general indication and High Sensitivity (HS) RNA Kit on a Qubit v.2.0 spectrophotometer (Life Technologies, Zug, Switzerland). The HS kit was used because of low RNA concentrations. The presence of RNA was additionally confirmed through 1% agarose gel electrophoresis stained with Midori green dye. Refer to Appendix A for tables containing data on the RNA extractions. The consistently low 260/230 ratios achieved with TRIzol extraction may have been caused by salts. However, multiple studies suggest that there is no impact on downstream analysis at concentrations below 100Mm (Quiagen, n.d.). Following quantification, DNAse from the TURBO DNA-*free*™Kit (Thermo Fisher Scientific, Massachusetts, USA) was applied to the RNA to remove any contaminating DNA. The total reaction volume after DNA Free treatment was 50 μl.

cDNA synthesis was initially perfomed using Agilent's Affymetrix cDNA synthesis kit (Agilent, California, USA). However, ultimately the Applied Biosystems cDNA synthesis protocol using "High c" Multiscribe Reverse Transcriptase (Thermo Fisher Scientific, Massachusetts, USA) was used. The change was made both owing to the reduced cost and the increased simplicity of the method. 10 μl of each extract was converted to cDNA. To conduct the cDNA synthesis, 200 μl tubes were labelled according to sample. The reverse transcriptase buffer, random primers and dNTPs were made into a master mix and added to samples that were then transferred to the Prime thermal cycler (Techne, Staffordshire, UK) and the program described in Table 3.5 was run.

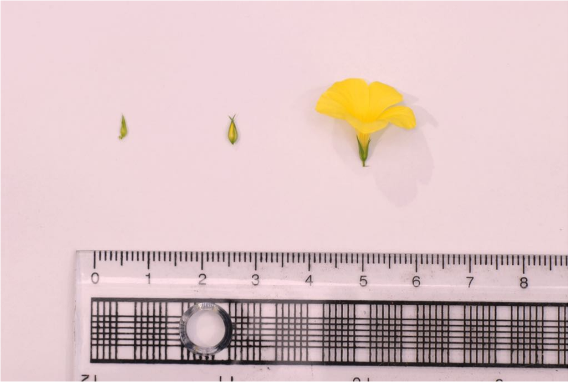


Figure 3.7: The three growth stages sampled for qPCR analysis. GS1: closed bud. GS2, maturing bud and GS4 open flower. Species depicted is *L. tenue*. Photography by Ali Foroozani.

3.2.5 qPCR

SYBR Green dye was used to quantify the amount of PCR product amplified from the cDNA during the qPCR reaction. SYBR Green is a cyanine dye that fluoresces when it binds to double-stranded DNA (dsDNA); consequently, the dsDNA that is amplified during PCR fluoresces upon binding to SYBR. The dsDNA increases with each cycle, and therefore the fluorescence also increases in a predictable way that is determined by the starting quantity of cDNA. The SYBR mastermix was made up as described in Table 3.7.

Table 3.7: Master Mix for cDNA synthesis. To create a total volume of 10 μl in each well. The SYBR mix additionally includes Taq and dNTPs.

Component	**Volume per reaction (μl)**
SYBR mix	7.5
Forward Primer	0.9
Reverse Primer	0.9
Nuclease Free Water	0.7

Quantitative PCR assays were conducted using an Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems-Life Technologies, Carlsbad, Brea, CA,USA) in 96-well plates. For each experiment, 5μl cDNA, diluted 1/25 was placed into each well, along with 10μl SYBR mastermix. The PCR cycling conditions utilised for each assay are indicated in Table 3.8. Enzyme activation at 95°C for 10 minutes, followed by the amplification of gene prodct through 40 successive cycles of 95°C for 15s, then 60°C for 60s. A primer dissociation stage, described in Table 3.9, was included to check for primer specificity and for the formation of multiple products. This melting curve initialised at 95°C and ended at 60°C. The cycling conditions for the qPCR are detailed in Table 3.8. Each standard curve had a negative control containing water in place of cDNA. *GAPDH* was used as the reference gene for each trial (refer to Section 3.2.1 for details of reference selection). *GAPDH* references were included on each sample plate and all sample types for each species were also included on a single sample plate to control for plate variation.

Three different cDNA dilutions were investigated, 1/50, 1/25 and 1/12. Noticeable improvements were observed between the 1/50 and 1/25 dilutions, with the Ct values indicating an average 27 cycles to amplify for *L. tenue* cDNA when diluted 1/50, compared to an average 18 cycles when diluted 1/25. However, there was no observable impact between the 1/25 and 1/12 dilutions. Therefore, in the interest of maximising the number of plates performed with the cDNA, subsequent qpCR experiments were performed at a 1/25 dilution.

Post qPCR analysis was conducted to compare Ct values to relative gene expression. The difference in Ct value between the reference gene and the candidate gene was recorded. The mean across three biological replicates was taken and this was plotted to determine relative candidate gene expression. To determine whether there was a significant difference in candidate gene expression between floral morphs, growth stages and species, a two-way ANOVA was performed. It was assumed that the variance in the dependent data across groups was normally distributed. Floral morph growth stage and species were considered fixed factors.

Table 3.8: qPCR cycling conditions. Data collection was completed at Stage 3, step 2.

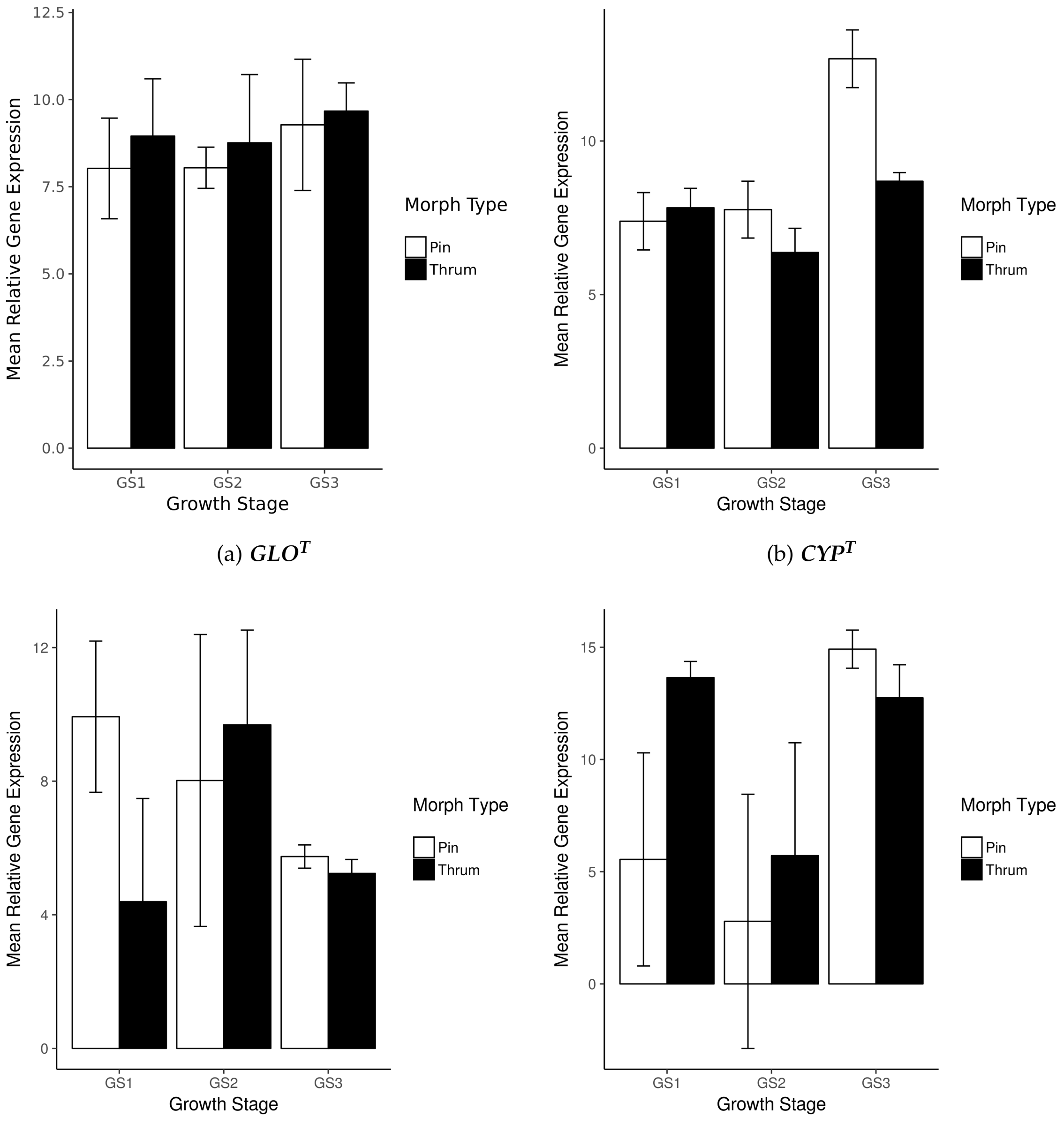
Stage	**Temperature (°C)**	**Time (s)**	**Repeats**
1. Holding	50	120	1
2. Enzyme Activation	95	600	1
3.1 Cycling	95	15	40
3.2 Cycling	60	60	40

Table 3.9: The primer dissociation stage implemented after the qPCR cycling.

Temperature (°C)	**Time (s)**
95	15
60	60
95	0.15
60	15

3.3 Results

3.3.1 Transcripts from Candidate Primers in *L. tenue*



(a) ***GLO***^{T}

(b) ***CYP***^{T}

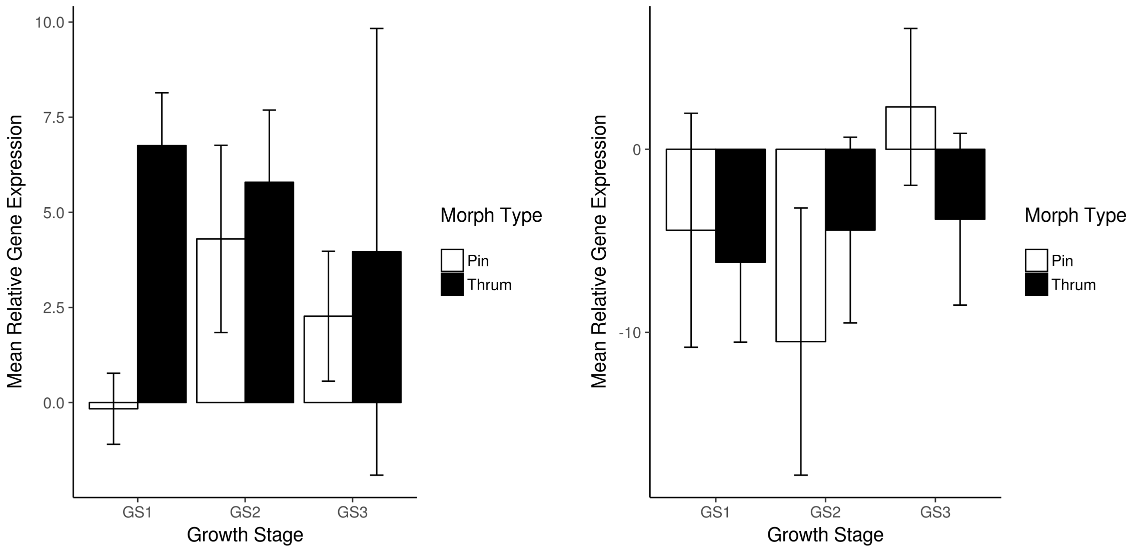
(c) ***PUM***^{T}

(d) ***S-ELF3***

Figure 3.8: Mean relative expression of **a)** *GLO*^{T}, **b)** *CYP*^{T}, **c)** *PUM*^{T} and **d)** *S-ELF3*. *GAPDH* is used as a reference gene. Error bars show standard error of the mean.

Figure 3.8a shows a very slight mean relative increase in *GLO*^{T} expression in short-styled morphs compared to long-styled morphs. However, this interaction is not statistically significant ($F_{2,12}$ = 0.322,p = 0.731). Equally, there is no statistically significant difference in gene expression between the three different developmental stages ($F_{1,12}$ = 0.316, p = 0.585).

Figure 3.8b indicates that there is a significant difference in mean relative *CYP*^{T}



(e) **PveGlo2** (f) ***LgMYB21***

Figure 3.8: Mean relative expression of **e)** *PveGLO2* and **f)** *LgMYB21* in *L. tenue*. *GAPDH* is used as a reference gene. Error bars show standard error of the mean.

expression between the long-styled and short-styled morphs ($F_{1,12}$ = 6.588, p = 0.025) and between the three developmental stages ($F_{2,12}$ = 23.312, p = 0.001). Tukey's Post Hoc test showed a significant increase at GS3 compared with GS1 (Tukey; p = 0.005) and GS2 (Tukey; p = 0.002). Thus, at GS3 there is a greater relative expression of CYP^T in the long-styled morph than the short-styled morph. In Figure 3.8c, there is no significant increase in PUM^T expression between morphs ($F_{1,12}$ = 0.390, p = 0.544), or between developmental stages ($F_{2,12}$ = 0.643, p = 0.543).

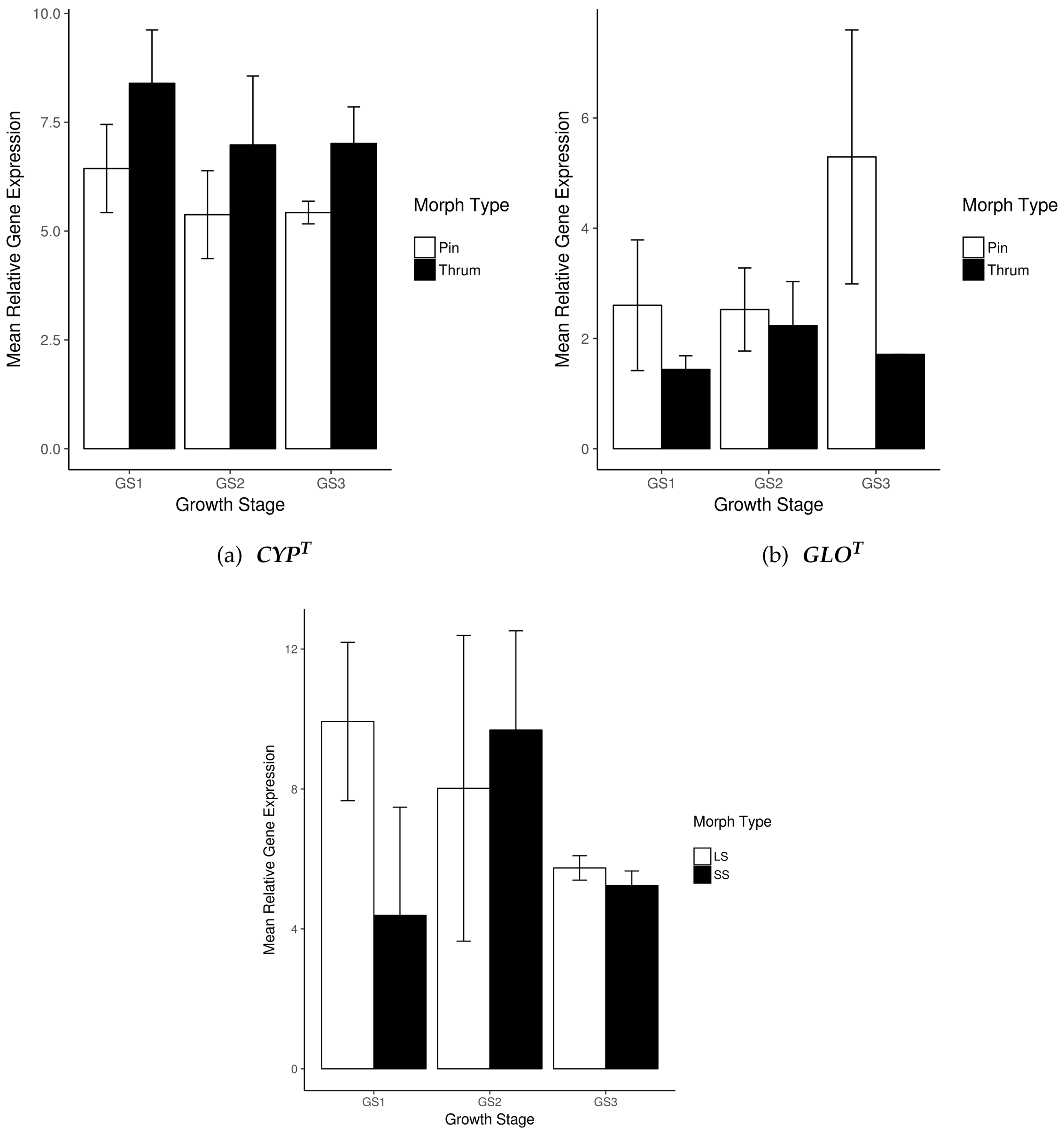
There is no difference in relative *S-ELF3* gene expression between floral morphs ($F_{1,11}$ = 0.401, p = 0.540) or between developmental stages ($F_{2,11}$ = 2.635, p = 0.116) (Figure 3.8d). There is gene expression in all floral morphs. *S-ELF3* expression is approximately constant across the two floral morphs. The lowest overall gene expression, of both morphs, was present at Growth Stage 2.

Figure 3.8e shows slightly greater expression in the short-styled morph at all three growth stages in *PveGLO2*, however neither the effect of growth stage ($F_{2,11}$ = 0.261, p = 0.776), nor the effect of floral morph type ($F_{2,11}$ = 1.068, p = 0.324) is significant. As can be observed by the error bars, which depict the SEM, there is significant variation in obtained relative gene expression value at GS3. In general, expression of both morph types is lower at Growth Stage 3, to the two earlier Growth Stages.

Finally, expression of *LgMYB21* (Figure 3.8f) was extremely variable, even within biological replicates. Many of the wells in the qPCR were undetermined, suggesting that no product was amplified after the 40 cycles of qPCR. There was no difference in

LgMYB21 between the different growth stages ($F_{2,12}$ = 70.249, p = 0.784), however, there was a reported significant difference in *LgMYB21* expression between the pin and thrum morphs ($F_{1,12}$ = 1.190, p = 0.018). However, based on the evidence in Figure 3.8f and the fact that many of the wells took almost 40 cycles to amplify, this significance should be interpreted with caution.

3.3.2 Transcripts from Candidate Primers in *L. narbonense*



(a) ***CYP^T***

(b) ***GLO^T***

(c) ***S-ELF3***

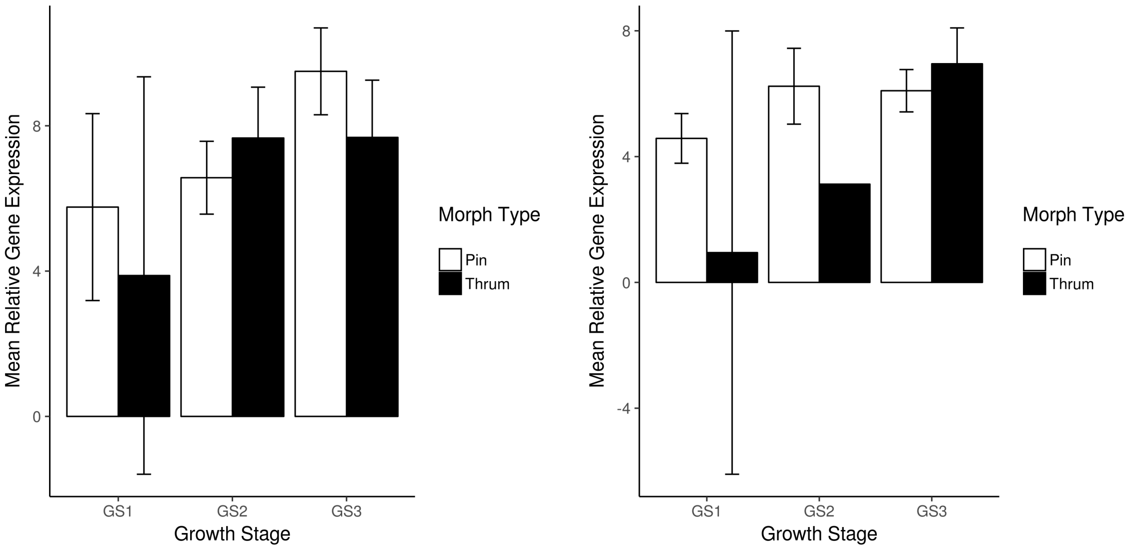
Figure 3.9: Bar charts show mean relative gene expression of **a)** *CYP^T* **b)** *GLO^T* and **c)** *S-ELF3* in *L. narbonense* at all 3 growth stages. Graphs were produced using R Statistics. Error bars show standard error of the mean.

Figure 3.9a shows that there is little difference in relative gene expression between the long and short-styled floral morphs in *L. narbonense*, this is confirmed statistically ($F_{1,12}$ = 3.880 p = 0.072). Although not statistically significant it can be observed that there is slightly greater expression of the short-styled morph at all three developmental stages. There is, however, no difference in *CYP*^{T} expression between the three different growth stages ($F_{2,12}$ = 0.870, p = 0.444). There was no significant difference in *GLO*^{T} expression between the three developmental stages ($F_{2,9}$ = 0.573, p = 0.583) or between the two floral morphs ($F_{1,9}$ = 2.027, p = 0.188) (Figure 3.9b). As with *CYP*^{T}, the greatest expression of this candidate gene was at the open flower stage, GS3. Relative expression of *GLO*^{T} was lower in general than expression of the other two candidate genes. The greatest relative difference in gene expression was at GS3, the open flower stage. However, the large, non-overlapping error bars confirm that there is little statistical power underpinning this conclusion.

Expression of S-ELF3 was more variable, as shown in Figure 3.9c. Gene expression occurred at all growth stages and in both floral morphs. There were no statistically significant differences in S-ELF3 expression between morphs ($F_{1,11}$ = 0.597, p = 0.456) or between growth stages ($F_{2,11}$ = 1.865 p = 0.201). Based on visual evidence from Figure 3.9c, expression was slightly greater in the long-styled morph at GS1, the young bud stage. However, this effect was reversed at GS2; there was little difference in gene expression between the two morphs at GS3.

Some genes were tested in *L. narbonense* as primers failed for these species, leading to the presence of many undetermined samples. As with *L. tenue*, *TSS1* was not expressed in *L. narbonense*; all values were undetermined.

3.3.3 CYP^T and GLO^T expression in *L. viscosum*



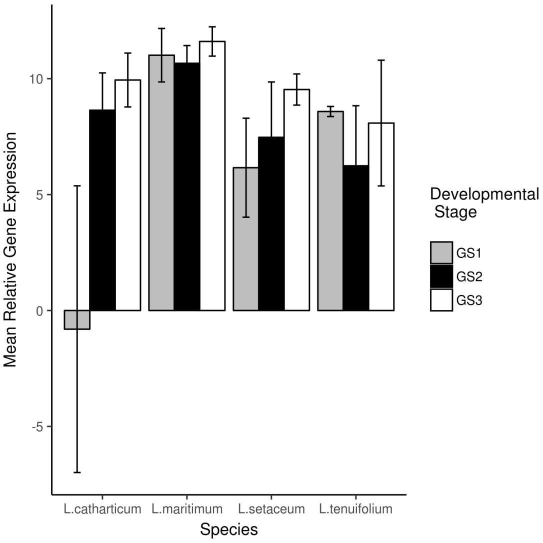
(a) ***CYP^T***

(b) ***GLO^T***

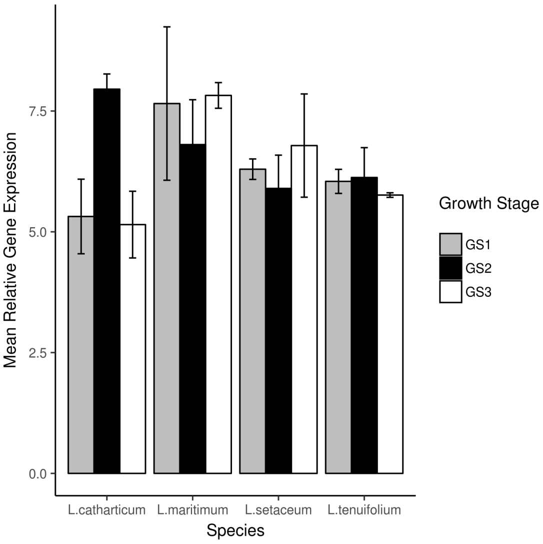
Figure 3.10: Bar charts show mean relative gene expression of **a)** *CYP^T* **b)** *GLO^T* at all 3 growth stages. Graphs were produced using R Statistics. Error bars show standard error of the mean

Growth stage had no significant effect ($F_{2,12}$ = 0.997, p = 0.397) on the relative expression of *CYP^T*. Equally, there was no significant difference in *CYP^T* between the two floral morphs ($F_{1,12}$ = 0.157, p = 0.699). Qualitatively, from Figure 3.10a there appears to be slightly greater relative expression in the long-styled pin morph at two of the three developmental stages (GS1 and GS3) and greater mean relative gene expression in the short-styled thrum morph at GS2. However, as this expression is not statistically significant. There was no statistically significant difference in *GLO^T* expression between growth stages ($F_{2,8}$ = 1.159, p = 0.361) or between the long-styled and short-styled morphs ($F_{2,8}$ = 1.159, p = 0.361) (Figure 3.10b) . Unsuccessful amplification of several qPCR wells reduced the sample size to 14 individuals. The non-amplifying wells were largely short-styled (thrum) morph samples.

3.3.4 *CYP^T* and *GLO^T* expression in homostylous species

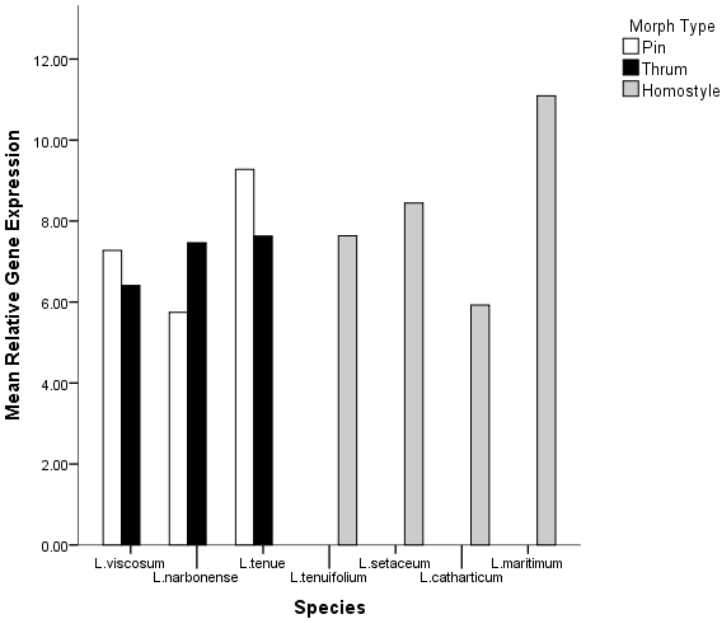


(a) ***CYP^T***

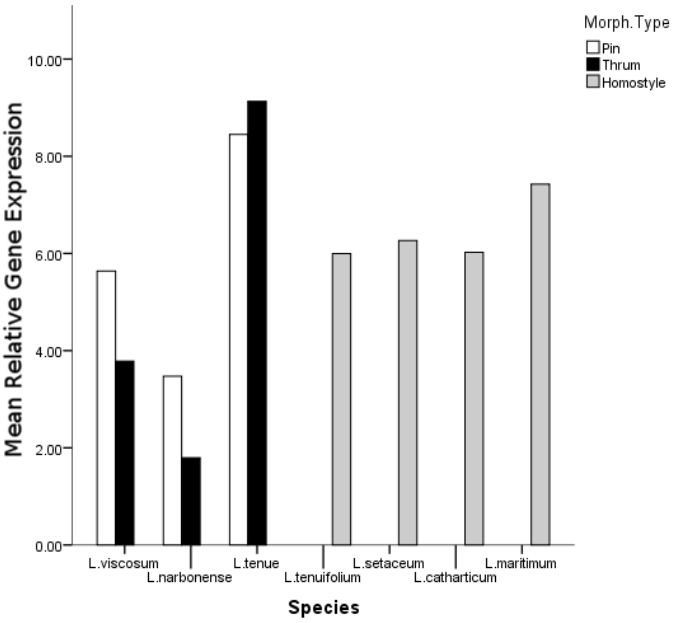


(b) ***GLO^T***

Figure 3.11: **a)** *CYP^T* relative gene expression in four homostylous species. **b)** *GLO^T* relative gene expression in four homostylous species.



(a) CYP^T

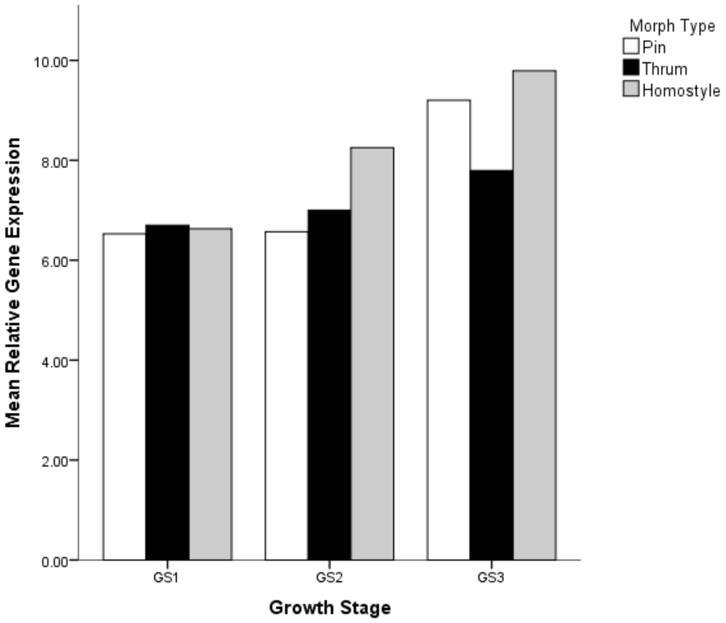


(b) GLO^T

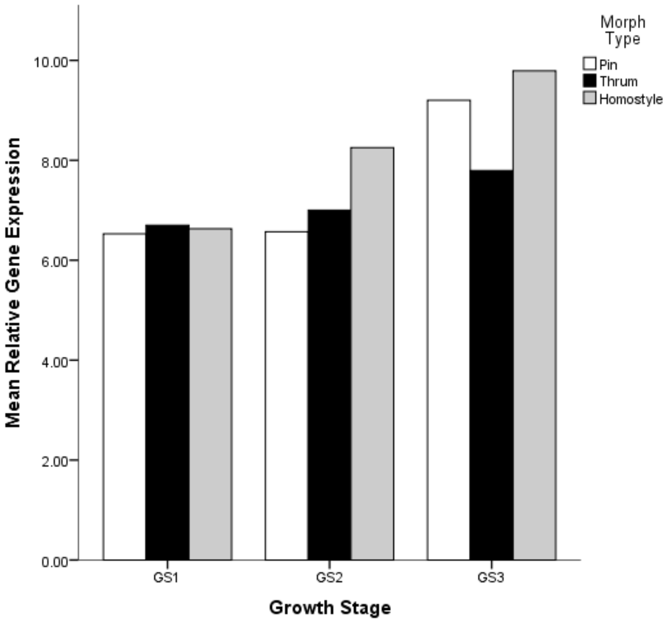
Figure 3.12: Comparison of the relative gene expression of **a) CYP^T** and **b) GLO^T** in all heterostylous and homostylous species at all three growth stages.

Comparing pin and thrum morphs from all tested heterostylous species against the homostylous species, as shown in Figures 3.12 and 3.13 indicates that there is significantly different expression of CYP^T between the three tested growth stages ($F_{2,59}$ = 3.521, p = 0.036) (Figure 3.12a) and between species ($F_{5,59}$ = 2.673, p = 0.030) (Figure 3.13a). However, surprisingly, there is no significant effect of morph type ($F_{1,59}$ = 0.966, p = 0.778). Tukey's post-hoc test with Bonferroni correction shows that CYP^T expression is significantly greater at GS3 than at GS1 (p= 0.031). Tukey's post hoc test additionally shows that expression in *L. maritimum* differs significantly from that in other homostylous species.

GLO^T was highly differentially expressed between species ($F_{5,47}$ = 11.337, p < 0.001), however, showed no significant difference in expression between floral morph type ($F_{1,47}$ = 1.805, p = 0.186) or between growth stages ($F_{2,47}$ = 1.682, p = 0.197). Tukey's post hoc test indicated that there were significant differences between *L. tenue* and *L. narbonense*.



(a) *CYP*^{T}



(b) *GLO*^{T}

Figure 3.13: Comparison of the relative gene expression of **a) *CYP*^{T}** and **b) *GLO*^{T}** in all heterostylous and homostylous species.

As can be seen from Figure 3.11a, *CYP*^{T} is expressed in all tested homostylous species, excepting *L. catharticum* at Growth Stage 1. There was no evidence that *CYP*^{T} expression differed between growth stages ($F_{2,24}$ = 2.199, p = 0.133) or between species ($F_{3,24}$ = 2.076, p = 0.090). *GLO*^{T} is also expressed in all tested homostylous species (Figure 3.11b). There is no significant difference in mean relative gene expression between the tested species ($F_{3,18}$ =1.836, p = 0.177), nor is there any significant difference in expression based on developmental stage ($F_{2,18}$ = 1.523, p = 0.245).

3.4 Discussion

This discussion will focus on the expression of *GLO*^{T} and *CYP*^{T}, as these candidate genes were the most consistently expressed across all tested *Linum* species. Additionally, Li et al. (2016)'s work in *Primula* has shown that *GLO*^{T} controls anther height and *CYP*^{T} mediates style elongation. As the heterostylous phenotype is related to reciprocal changes in the heights of sexual organs, expression of these particular candidate genes in *Linum* make a good starting point for investigation. Potential explanations as to the inefficacy of the other examined candidate genes will also be considered. A more comprehensive study would include testing of the remaining candidates, especially *KFB*^{T}, *CCM*^{T} and *PUM*^{T}, which are still of unknown function in the *Primula* model system (McClure, 2016). Overall, it can be ascertained that there was no evidence of significant differential expression of *GLO*^{T} nor of *CYP*^{T} in the heterostylous species.

Firstly, with regard to *GLO*^{T}, which is a duplication of the floral-B homeotic gene *GLOBOSA* in *Primula* (Nowak et al., 2015; Li et al., 2016), the obtained results varied with the particular *Linum* species tested. All three of the tested heterostylous species exhibited slightly different relative expression patterns of the purported duplicated *GLOBOSA* gene, however, none were statistically significant. Resultantly, these differences are almost certainly associated with the experimental procedure.

There was expression in both the long-styled and the short-styled morphs of all individuals. This was not in accordance with the hemizygous genetic system which had been observed in *Primula*. The expectation, based on Li et al. (2016)'s model was for expression of *S-locus* genes to be restricted to the thrum (short-styled) morph; it was anticipated that the orthologue of the thrum-specific *GLO* from *Primula* would be absent in the long-styled pin morph and present only in the the short-styled thrum morph (Li et al., 2016). However, the likelihood that heterostyly has arisen independently on multiple occasions (Barrett, 2013) points either to an alternative mechanism for heterostyly in *Linum* or to the *GLO*^{T} orthologue not being an integral part of the *S-locus*. The lack of difference in *GLO*^{T} expression across the two morphs possibly suggests that *GLO*^{T} is not as important in the genetic mechanism of heterostyly in *Linum* compared to in *Primula*.

There was no statistically significant evidence of differential expression of *GLO*^{T} in heterostylous *L. tenue*, *L. narbonense* or *L. viscosum*. *GLO*^{T} is expressed to a lesser extent in homostylous individuals than heterostylous individuals (Figure 3.13b). The significance in this case was related to species, rather than to morph type. This finding was in accordance with our expectation as homostylous *Linum* species would be unlikely to have an *S-locus* supergene on account of the fact that they do not have different floral morphs.

In *L.narbonense*, on the other hand, there is evidence that *GLO*^{T} is more highly expressed in the long-styled pin morph, than the short-styled morph. This, again, was contrary to expectation. Some explanation for the greater expression in the long morph may be attributed to potential differences in the control and expression of distyly exhibited by *L.narbonense* and *L.tenue*. *L.narbonense* is likely to be tetraploid, with 2n = 4x = 28 (Bolsheva et al., 2017). These two similar diploid chromosome sets indicate that a genome duplication has recently taken place which could be complicating the pattern of expression of the S locus. The complexity of polyploid genomes often precludes accurate analysis of their genetic make-up. *L.suffruticosum*, for example is polyploid and therefore even specific primers may not amplify the correct gene, or may not pull out the particular gene copy that is responsible for heterostyly.

This work has also suffered issues related to gene families. As discussed in Chapter 2, the *L. tenue* transcriptome used as a reference map may not have differentiated between the various members of the *GLO*^{T} family sufficiently to enable accurate separation of the specific gene duplication copy that may be related to anther-length in heterostyly. This may have gone some way towards preventing gene-specific amplification.

The obtained results might equally arise from the primer design phase. There is a possibility that the devised primers are picking up gene copies with similar functions in addition to the desired, *S-locus*-specific genes. This is plausible as genes may have been copied in order to be recruited for heterostyly, especially if the Lloyd model for the evolutionary origin of the heterostylous polymorphism is to be accepted. Certainly, Li et al. (2016) present evidence for this being the case in *Primula*, especially regarding the thrum-specific *GLO*^{T}. However, it is not obvious from the data presented in this analysis that the *Primula* candidates were involved in the *S-locus*. More experiments, and newly designed primers would be required to definitively rule out a role for *Primula* gene orthologues in the heterostylous system in *Linum* species. It is also very important to note that it was appreciated in hindsight that sequence differences at the loci in different species meant that the primer binding efficiency to those loci in different species is not equal. Thus, amplification efficiency differs and the relative quantitation presented is therefore inaccurate. The new evidence presented in sections 3.2.2 and 3.2.3 calls the identity of *GLO*^{T} and *S-ELF3* into doubt and could not confirm that the primers were binding to the loci at all. The primers were amplifying cDNA in the qPCR experiments so it is not all bad news, however, new primers, designed from a larger number of annotated transcriptomes would be required to make significant progress towards identifying and quantifying orthologue expression.

A further alternative explanation is that these results may be an artefact of the 2016 collection date and storage in RNA-later of the wild species. There may have been some RNA degradation over this time period, which could help to explain the false negative results (e.g. Leonard, 2016). To make a more direct comparison, it would be prudent to attempt to grow wild species such as *L. narbonense* and *L. viscosum* in the greenhouses at Durham University. Maintaining the species under glasshouse conditions would assist in controlling for environmental variation and would also likely reduce the differences in growth stage identification caused by qualitative observation in the field.

There was no expression of *PveGLO2* in the long-styled morph at GS1 in *L.tenue* (Figure 3.8e). However, there was relatively much greater mean expression in the short-styled thrum morph. This is in accordance with expectation based on the hemizygous model in Primula. Interestingly, this was not the case when testing the *GLO*^{T} (Figure 3.8a). The trend was observed in *S-ELF3* (Figure 3.8d), although gene expression was not completely absent in the long-styled morph, only reduced. This is noteworthy as according to the bioinformatic analyses conducted, *S-ELF3*, *PveGLO2* and *GLO*^{T} may be isolating the same contig of the reference *L.tenue* transcriptome, and therefore might be expected to exhibit the same behaviour. Of course, the sensitivities and specificities of qPCR may be a factor in this result. Table 3.2 shows that the primers selected were in fact very different, despite the fact that these candidate genes, isolated from three different plants are all believed to code for the 'A' locus. *PveGLO2* was not amplified in *L. narbonense* or *L. viscosum*; the majority of the wells were recorded as 'undetermined'. Previous comments on methodological error aside, perhaps the successful amplification of the *PveGLO2* primer in *L. tenue* is associated with the fact that the *L. tenue* transcriptome was used (in conjunction with *L. usitatissimum*) for primer design.

This project had noted that both *S-ELF3* and *PveGLO2* may be orthologues, corresponding to *GLO*^{T}, or the *A* in proposed *GPA* linkage group at the *S-locus* based on their isolation of the same contig of the *L. tenue* transcriptome. Such claims are corroborated in the recent literature review of Kappel et al. (2017). However, the evolutionary distance between *Fagopyrum* and *Primula*, the known genetic functions of *ELF* and *GLOBOSA* genes and personal correspondence with Professor Philip Gilmartin regarding unpublished data have proven this assumption to be incorrect. This finding would point towards the existence of the *S-locus* in *Linum*, and across a greater number of species in general. However, the observation may also be an artefact of analysis, related to the incomplete assembly of the *L.tenue* transcriptome assigning two similar gene copies to the same contig. As discussed in further detail in Chapter 2, the annotated transcriptome used for the primer design and for sequence mapping as part of RNA-Seq analysis is still under construction. It may not have sufficient resolution to differentiate between *GLO*^{T} and *GLOBOSA*. An updated transcriptome is awaiting completion (Foroozani, Unpublished Doctoral Thesis), and hopefully the results will soon be able to be re-analysed (Foroozani, personal communication), with the aim of improving the attained mappings. It would be prudent to analyse just one or two sequences, rather than all 44, when attempting to optimise the RNA-Seq analysis pipeline.

CYP^T, which, in *Primula*, encodes the brassinosteroid inhibitor cytochrome P450 also showed little difference in gene expression in *L.tenue*. There was no significant difference in gene expression between floral morph or growth stage (Figure 3.8b). *CYP*^{T} is expressed to a greater extent in the thrum morphs at all growth stages in *L. narbonense*,although there is no statistically significant difference. In *L. viscosum* there is very little difference in *CYP*^{T} expression across morphs (Figure 3.10a). Cytochrome P450 functions as a C-22-α-hydroxylase in brassinolide biosynthesis in *Arabidopsis*. Given the fundamental role of stylar growth in heterostyly, the lack of difference in gene expression across both morphs was all the more surprising. Greater CYP^T expression would have been anticipated in the thrum morph as equal expression of *CYP*^{T} in both floral morphs contradicts the proposed hemizygous model of the inheritence of heterostyly. Trials using *CYP*^{T} are potentially worthy of repetition as the RNA-Seq analysis conducted in Chapter 2 was unsuccessful in mapping any *CYP*^{T} orthologues to the *L. tenue* transcriptome. A second analysis would go further towards determining whether the primers were amplifying the correct product in *Linum* or whether the new transcriptome might enable successful mapping of the *CYP*^{T} contig.

PUM^T again shows no significant difference between floral morphs or growth stages in *L. tenue*. *PUM*^{T} encodes a Pumilio-like RNA-binding protein. This gene has not yet been assigned any specific phenotypic role in *Primula*, although it is present in the *S-locus* and is part of a general functional group. Pumilio-like proteins, in general, are not well functionally characterised in plants (Abbasi et al., 2011). However, in animals and yeast they are involved in diverse variety of roles in post-transcriptional RNA control and metabolism; including RNA decay, transport and processing, and translational repression. Puf protein-mediated deadenylation of mRNA in eukaryotic cytoplasm often accompanies mRNA decay and/or translational repression (Goldstrohm et al., 2007), which can ultimately impact growth and devlopment. Thus, Pumilio-like RNA-binding proteins are likely to have a role in the regulation of such growth and development. Although, to date, this function has not been greatly explored in the plant kingdom, the role of *PUM*^{T} in the *S-locus* will, it would seem, in all likelihood be related to this translational repression. Nevertheless, irrespective of the potential function of *PUM*^{T} in *Primula*, in this work there was no evidence of significantly different relative gene expression of the homologue of *PUM*^{T} between the two floral morphs in *L.tenue*, or indeed between the three developmental stages.

KFB^{T} encodes a protein with similarity to the *Arabidopsis* Kiss-Me-Deadly Kelch repeat F Box protein, which is involved in regulating the activity of cytokinin. The Kfb^{T} primers were not able to amplify any wild-species DNA in either floral morph. It would be advantageous to pay closer attention to primer design in order to further test this gene, which is unique to the thrum-specific 278kb region in *Primula*. Improvements to the primer design process, perhaps even by designing primers specific to each species, might positively impact the attained results. The new RNA-Seq data attained in Chapter 2 may help to inform the primer design process as through undertaking the suggested *de novo* sequencing of the wild species sequences new species-specific primers for all 10 species will be possible.

LgMYB21 expression was not well captured through the qPCR analysis (Figure 3.8f). Despite several attempts, it was difficult to develop good primers to isolate this gene in any of the wild *Linum* species or in greenhouse-grown *L. tenue*. Although none of the wells technically failed to amplify within 40 qPCR cycles, the δCT values were in the high 30s for several reactions. This implies that any amplification was likely to be of e.g. primer dimers or other contaminants. The values that were observed suggest that there was a significant difference in gene expression between floral morphs, the greater expression however, is seen in the pin morph. If this were to be a biological effect, Ushijima et al. (2012) has noted that *LgMYB21* expression was not just restricted to sexual organs; expression was additionally observed in sepal and petal tissues. Ushijima et al. (2012) did note, however, that *LgMYB21* was expressed strongly in the style and much less in the stamen, thus presenting the possibility that *LgMYB21* regulates pistil length. The absence of a thrum-specific polymorphism means that *LgMYB21* is unlikely to be linked to the S locus. It must be noted the unusual and varying phenotypes of heterostyly across a series of species in the genus *Linum* do not entirely discount a role for *LgMYB21* in regulating pistil length. Further evidence from Ushijima et al. (2012) implicates *LgMYB21* as a downstream component of the G portion of the *S-locus* supergene.G is the proposed gene regulating style length. *LgMYB21* may be useful in isolating the G gene and by extension in elucidating the mechanism involved in the regulation of style development.

TSS1 (Ushijima et al., 2012) is an attractive contender for an *S-locus* gene in *Linum*. The fact that it was one of the few candidates identified in a *Linum* species (*L.grandiflorum, (Ushijima et al., 2012)*) by default made it a strong contender for an *S-locus* associated gene in *Linum*. Ushijima et al. (2012) found the gene to be expressed only in styles and only in the thrum morph. However, as *TSS1*-matching genes could not be found using any publicly available search-engines or databases, it is difficult to predict localisation *in silico*. Further examination of *TSS1* would be both useful and insightful. It proved impossible to design primers for *TSS1*, and consequently *TSS1* differential expression was assessed using primers designed for *L. grandiflorum*. Despite the close relations of *L. grandiflorum* to the other wild species, it is unsurprising that the primers were ineffective at amplifying *TSS1* in either *L.tenue* or *L. narbonense*. The primers will have been very specific which may have contributed to their failure to amplify. The failure of any reads to map to the closest *L. tenue* contig (Chapter 2), however, casts doubts on the presence of a closely-related orthologue in any of the tested *Linum* species. Not being able to design primers was unfortunate, especially as *TSS1* was considered to be a strong candidate based on Ushijima et al. (2012)'s study in *L. grandiflorum*. Its style and thrum-specific expression pattern was indicative of similarity to '*S-locus*' gene CYP^T. The primer's failure may be related to sequence differences which made the creation of a consensus sequence using Clustal Omega (Sievers et al., 2011) and GeneDoc (Nicholas & Nicholas, 1997) more challenging. BLAST searches against the published *L.usitatissimum* transcriptome did not produce very strong results. As mentioned in the methods section, flaws in the primer design process are not limited to *TSS1*. Individualised primers designed to match each species would be the best way to address this, although would require the annotation of transcriptomes for each wild species.

The lack of statistically significant differences recorded between the two heterostylous floral morphs may have resulted from methodological errors in the laboratory, despite best efforts to prevent RNA degradation and close attention to the MIQE guidelines (Bustin et al., 2009). There are many sources of variability that can be introduced, from sample collection procedures, through nucleic acid extraction methods to quality control.

A further limitation associated with the employed method of qPCR analysis is that an unnaturally large ΔCT value is achieved for 'undetermined' samples, where the amplification threshold was not reached during the qPCR cycle. Therefore the normalisation technique is ineffective for very small amounts of amplification. Where possible the 'undetermined' samples were filtered out during the post-processing, however, this then reduced the number of biological replicates included in the comparison from three to two or one. One other complication associated with qPCR analysis is that this study was looking to find no expression at all of *S-locus* candidate genes in the pin morph samples. In this qPCR analysis, therefore, these samples should be registering as 'undetermined'. It is possible that those samples that have been attributed to methodological error are in fact evidence of a notable finding. Perhaps it should therefore be more appropriate to classify these samples as 'no-expression' samples.

As with all projects involving wild biological material, experimental analysis can be hampered by differences between flowers even of the same species, and flowers of different species were subject to species-specific stresses. These species, with the exception of *Linum tenue* were sampled during field investigations. It was difficult, for example, to extract RNA from the *L.maritimum* samples. Low RNA yields were obtained and the extractions were performed several times to obtain adequate yields. This could result from the multitude of secondary compounds that are built up to resist external stresses.

One potential issue which hindered adequate amplification of the RNA was vastly different primer melting temperatures; most notably the differences between the reference and candidate primers. As an example, the melting temperature of *LgMYB21* was 63 °C, whereas the melting temperature of *GAPDH* was considerably higher. Furthermore, in some experiments, the *GAPDH* appeared to have an unusual dissociation curve. This is suggestive of some contamination to the reference primer, or alternatively to the samples themselves. Although steps were taken to find the most appropriate reference gene to use for *Linum* species, the qPCR trials could be replicated using a different reference, and then the relative results normalised between these. There was, unfortunately, insufficient time to attempt these experiments during the project.

It appears to be a general trend that, in general, candidate genes were more highly expressed in heterostylous than homostylous species (see Figure 3.11a). The expression differences are likely to be attributable to different morph types. This is in accordance with the accepted literature, which would expect homostylous species not to possess, or to possess pseudogene copies of heterostyly related genes. Homostylous individuals are largely self-compatible (Ganders, 1979) and thus, it would be expected for there to be no expression of potential gene homologues associated with the self-incompatibility portion of the *S-locus* (for example LgSKS1). Interestingly, studies by Fesenko et al.

(2006) show that homostyly in one accession of *F. homotropicum*, a homostylous relative of *F. esculentum* is determined by a single gene. Recent work by de Vos et al. (2018) has contradicted the previously assumed relationship between homostyly and selfing. Homostyly had previously been interpreted as an adaptation to promote autonomous selfing in marginal environments with low pollinator availability (de Vos et al., 2018). Homostylous *Primula* species, *P. halleri* has adopted a mixed mating system owing to varying levels of herkogamy. Thus, there may be a potential locus associated with presence of homostylous individuals.

Beeflies from the genus *Usia* seem to be important pollinators in some Mediterranean distylous species. *L.suffruticosum* has been observed to be almost entirely pollinated by several *Usia* flies. *Usia* are thought to be the main pollinators, not only of *L.tenue*, but to a lesser extent of *L.viscosum* and *L.narbonense*. Additionally, there does seem to be a difference in pollinator between monomorphic species, such as *L.tenuifolium* and heteromorphic species. To some extent, the differing pollinators may have influenced the shape of flower, and by extension the precise phenotype of the mating system in the various species of *Linum*. It is unclear whether the visiting pollinators have chosen the distylous species based on their morphology, or whether this is a case of parallel evolution. Nevertheless, it may provide a potent explanation for the wide variation in mating strategy across the genus *Linum*. *Linum* flowers, whether or not this was pollinator-motivated, do show considerable diversity between species (e.g. Armbruster et al., 2006). This diversity may cause a difference in patterns of gene expression between the tested *Linum* species and may initially hinder our ability to determine the genes responsible for distyly in this genus.

3.4.1 Future Work

As detailed in the discussion section above, methodological errors are likely to have prevailed in the qPCR data, as well as fundamental uncertainties over the nature of the genes being amplified. Conducting more qPCR experiments would help to resolve this. One course of action to confirm the lack of significance could be to redesign the primers and test them in all of the wild species. Furthermore, primers could be redesigned to maximise their specificity to thrum specific genes. This would minimise the risk of primers picking up very similar genes, from the same family, but which are not associated with the development of heterostyly. Specific primers could be designed for *L. tenue*, since a complete, although unpublished transcriptome is available. Although narrowing the number of species for which the primer would be effective; designing specific primers would improve the likelihood of extracting *S-locus* genes from the wild species, should they be orthologues of the confirmed genes in *Primula*. There are also additional wild species present in the -80 °C freezer, having been used for the creation of RNA-Seq libraries which could be used to increase the breadth of the differential expression analysis. Most notably, an additional heterostylous species, *L. suffruticosum* would add a further heterostylous species to the list of samples tested for candidate gene expression. *L. suffruticosum* exhibits 3D reciprocity (Armbruster et al., 2006); the anthers and stigmas are not closely reciprocal in height, but rather in how the stamens and styles bend and twist. This should provide interesting further data as to the differences in the 3D reciprocal species compared to other heterostylous species which show reciprocity solely in the vertical axis. Armbruster et al. (2006) notes that the difference in placement result in dorsal pollen placement by short-styled (thrum) flowers and ventral placement by long-styled (pin) flowers.

With more time available, more replicate samples could have been analysed. A greater number of replicates would go some way towards mitigating the sensitivity of the qPCR. The evidence presented in this thesis is subject to a serious number of methodological errors (as evidenced in the discussion), and, although it yields some interesting insights into the behaviour of distyly in the genus *Linum*, the results observed are as likely to be a factor of the experimental method, or the success of laboratory analysis, than of true biological interest. There were several instances in which the qPCR clearly failed to amplify cDNA in one of the wells (notably all primers tested in *L.catharticum*), however, the biological replicates were unaffected. These failed wells were removed from the analysis of average gene expression, and therefore clearly reduce the robustness of the data. Replication of these qPCR experiments would help to improve the quality of the data in such situations. The *GAPDH* reference gene also failed to amplify in *L. viscosum* meaning that an alternative reference primer *ETIF3E* was used instead. In future works it would be advisable to test the efficacy of all reference primers across the different wild species tested during qPCR, rather than only in *L.tenue*; of course this is a major piece of work and requires the *de novo* sequencing and annotation of transcriptomes of ten different wild species. Furthermore, the methods used to select the qPCR reference genes could be improved or validated using Microsoft Excel methods such as NormFinder (Andersen et al., 2004), BestKeeper (Pfaffl et al., 2004) or the BioConductor package NormqPCR.

From the qPCR experiments, no strong differences related to different developmental stage were observed. In order to further examine a likely possible effect of developmental stage, flowers could be sampled when they are a certain number of days old, rather than relying on a qualitative, visual measure. This would improve fine analysis of the effect of growth stage. In order to use this method, it would be advantageous to grow as many wild species as possible in the glasshouse, so that flowering date could be monitored with accuracy. In the past, wild species have proven difficult to grow under glasshouse conditions in the UK and therefore some effort into perfecting growth conditions for a series of different wild *Linum* species would be required. Additionally, some species are perennials and take in excess of one year to reach reproductive maturity and still others show strong seed dormancy that is difficult to break (Pérez-Barrales, personal communication).

Our laboratory has also produced a 'bi-directional BLAST hits' program, designed by Ali Foroozani in Durham (Foroozani, Unpublished Doctoral Thesis), which is able to identify pairs of genes in two different genomes that are more similar to one another than either is to any other gene in the other genome (Dalquen & Dessimoz, 2013). Leaf tissue could be taken as control data. The early bud stage, GS1, contains a large amount of green leaf-tissue (evident from Figures 2.5 and 3.7). It might therefore be expected that standard gene expression of those genes. Using leaf tissue as a control may help to account for erroneously high read counts. Performing a floral organ-specific qPCR might also help to mitigate these effects.

It may be possible to conduct some form of subtractive hybridisation experiment to compare differential expression between pin and thrum morphs. Subtractive hybridisation would cause identical sequences between the two morphs to be blocked off, thereby reducing the pool of potential genes that must be sequenced as well as faciliating the analysis of differential expression. Suppression Subtractive Hybridisation (SSH) is an approach for identifying and characterising differences between two populations of nucleic acids (Rebrikov et al., 2004). It is a powerful technique for the study of gene expression in specific tissues or cell types at a specific stage. mRNA from the target material is hybridised with first strand cDNA from the subtractor material.

Finally, conducting transformation experiments within a *Linum* model may be worthy of consideration. The purpose of such transformations would be to test the phenotypic effects of mutating various *S-locus* candidate genes. Cultivated flax (*L. usitatissimum*) has been transformed on several previous occasions, including using *Agrobacterium tumefaciens* (e.g. Basiran et al., 1987) and by particle bombardment with plasmid DNA-coated gold particles (Wijayanto & McHughen, 1999). Transformation of so many *Linum* wild species would require significant initial work and financial outlay to set up and would constitute a much longer term project. However, doing so would ultimately lead to a deeper understanding of the genome of various *Linum* species, beyond the genes involved in mating strategy since, unlike *Arabidopsis*, the *Linum* genome is complex, diverse and can be polyploid, e.g. *L.suffruticosum*. Taking the results from the RNA-Seq data analysis and the search for differentially expressed genes using qPCR together, it should possible to define a future course of action.

Chapter 4

Conclusions and Future Work

Heterostyly was first recorded in the genus *Linum* in the days of Darwin (Darwin, 1863). Despite this longstanding pedigree, however, the genes responsible for the development of distyly in *Linum* are not yet known. Based on classical genetic work conducted in *Primula* and more recent molecular genomic studies in *Fagopyrum* (e.g Yasui et al., 2012) and *Turnera* (e.g. Labonne et al., 2009) amongst others, distyly is expected to be controlled by the *S-locus* supergene. It was anticipated that the *S-locus* in *Linum* would be made up of orthologues of at least some of candidate genes already determined in *Primula* and *Fagopyrum*. It was additionally expected that there would no expression of *S-locus* candidate genes in the L-morph individuals, given the conclusions of Li et al. (2016) that a hemizygous model for heterostyly exists in *P. vulgaris*, whereby the *S-locus* haplotype is present only in the short-styled individuals.

To date, there have been relatively few genetic studies of heterostyly, and those that have been conducted have largely been confined to *Primula*. Thus, although this project has failed to find any definitive evidence of the presence of *S-locus* gene expression in any *Linum* species, it has provided new data which may challenge our assumptions of the way in which sexual organ dimorphisms occur in *Linum*. *Linum* deserves more attention, particularly given its fascinating reproductive mechanisms and the vast morphological differences exhibited even by members of the same genus (McDill et al., 2009). The species in which the majority of studies of heterostyly have been conducted are much more well understood than wild *Linum* which goes some way towards explaining why the controlling mechanism behind heterostyly has yet to be found. This project has given an appreciation of the intricacies and pitfalls of a cross-species investigation and has provided the author with a much clearer understanding of the bioinformatic process and of evolutionary distance.

Heterostyly has undoubtedly independently evolved on several occasions (e.g. McDill et al., 2009; Barrett & Shore, 2008). Phenotypic similarities among those plants which exhibit heterostyly are suggestive of convergent evolution in response to comparable selective pressures, although, not necessarily through identical genetic mechanisms and developmental pathways. From evidence presented in this project, it cannot be confirmed that heterostyly acts via the same mechanism in *Linum* species as it does in the well-studied *Primula* model. It is possible, therefore, that a different mechanism for heterostyly prevails in *Linum*. There may be variation even among different *Linum* species. It is certainly the case that the various wild heterostylous *Linum* species exhibit a spectrum of vastly different variants of reciprocity; for example, the 3D heterostyly observed in the open flowers of *L. suffruticosum* (Armbruster et al., 2006).

It was anticipated that genes controlling heterostyly in *Linum* could be found without too much difficulty by designing primers based on the candidate genes found in the 278 kb thrum-specific region in *P. vulgaris* from the recently published study by Li et al. (2016). Unfortunately, it transpired that this was not to be the case. It has subsequently been realised that the fact that the primers were designed from a consensus of *Linum* species, before RNA-Seq analysis of the wild species had been conducted is likely to be the main contributing factor and that consequently the qPCR results presented should not be taken as definitive evidence. A large number of candidate genes were considered following an initial literature review, however, it proved more challenging than initially anticipated to design effective primers for many of these candidates. None of the candidate genes identified from *Turnera* (Labonne et al., 2009; Labonne & Shore, 2011) led to the design of successful primers, and furthermore, none showed evidence of differential expression between floral morphs. Even those genes for which consensus primers could be created were then unsuccessful at amplifying cDNA extracted from the wild samples (described in Chapter 3). This may be indicative of vastly different sequences, however, equally, the challenges of mapping across different species may have precluded the discovery of differential expressions of orthologues of those particular candidates. By and large, the most successful candidate primers were those designed from the Li et al. (2016) paper. Most notably among these were *CYP*^{T} and *GLO*^{T}, the already affirmed genes causing anther height and style length in *Primula* (Li et al., 2016). These genes are clearly very highly conserved across species which in itself provides some evidence that such candidates ought to be involved to some extent in the control of distyly in *Linum*. It was hence very surprising that expression of *TSS1*, a gene identified in a distylous *Linum* species (*L. grandiflorum*, Ushijima et al. (2012)) could not be detected bioinformatically or experimentally. Perhaps *L. grandiflorum* should be included in a list of assessed species; if *TSS1* expression cannot be detected in this species then the results may be attributed to methodological error. It was disappointing, having identified so many candidate genes for the *S-locus*, that only three or four were consistently expressed when tested. However, this finding does suggest that efforts should be concentrated on orthologues of the *Primula S-locus*, and thus focuses future investigation.

There were two overarching sources of differential expression identified in this study. These were: firstly, genes up-regulated in response to stress conditions, which is discussed in detail in Chapter 2 and secondly statistically significant differential expression of candidate genes between species (Figure 3.13a). The latter suggests that the particular *Linum* species is a strong factor impacting the gene expression profiles of these individuals, including those of candidate genes. This correlation was also observed in the PCA plots drawn from the DESeq2 analysis, where samples tended to cluster by species rather than by floral morph or by growth stage. Perhaps this should have been more obvious from the beginning, however, it is possible that the large species effects are masking other signals in the data which may be of biological significance to the control of heterostyly. Even if, as suggested by the post-analysis amino acid BLAST searches, a mixture of genes from the same family were found, rather than the orthologues of *Primula* or *Fagopyrum* genes, the same strong, stress-responsive expression might be expected to occur irrespective of the exact identity of the gene whose transcript was being measured. As previously discussed, it may thus be advisable to test a smaller subset of species at a time, but performing at least 5 biological replicates at each growth stage. The differential expression analyses should be interpreted with caution, given the lack of replication. However, for the purposes of this study the different heterostylous species were considered to be biological replicates and it was hoped that they would give a signal of either presence or absence of *'S-locus'* genes.

As can be observed from the qPCR results, *GLO*^{T}, *CYP*^{T} and *PUM*^{T} are all expressed in all individuals, at all growth stages. It was therefore evident that primers were successfully designed from *Primula S-locus* genes to amplify genetic material within *Linum*; although new analysis in Section 3.2.3 brings into question whether they amplified the intended target. The question remains as to whether or not these genes are associated with heterostyly in the various *Linum* species and further, whether the primers were amplifying the intended regions of cDNA; especially if expression in *Linum* follows the hemizygous model proposed in *Primula* (Li et al., 2015). Subsequent BLAST searches of amino acids have shown that in *GLO*^{T} at least, the most closely related contig may in fact be different to that assumed in this thesis. Furthermore, it is likely that the primers amplified several, very closely genes in addition to the thrum-specific ones intended. There was expression of *GLO*^{T} and *CYP*^{T}, at the very least, in homostylous species.

Expression in homostylous species was unexpected, and had been discussed by some work conducted in *Primula* by the Gilmartin group (Gilmartin, 2017).

CYP^T was expressed in the qPCR analyses, but was not expressed in the RNA-Seq analyses. In fact, no raw read counts of *CYP^T* were observed at any growth stage. The likely cause will have been the poor mapping coverage, although the qPCR results perhaps point to the designed primers being too broad. It might thus be advantageous, though time-consuming and expensive, to design primers on a per-species basis. The added specificity of bespoke primers would ensure the amplification of the correct primer product, and would help to ensure that the raw counts examined from the RNA-Seq analysis were correctly apportioned. The read counts achieved for *CYP734A50*, and the fact that these were not the same as *CYP^T* is further evidence of poor primer specificity. It might have been expected, since these two candidate genes are both believed to be responsible for the style length at the *S-locus* in *Primula vulgaris* and *veris* respectively, that they would correspond to the same portion of the amplified *L. tenue* transcriptome.

GLO^T expression, unlike *CYP^T*, was detected using both RNA-Seq and qPCR. Not only that, but *GLO^T* behaved roughly according to the hypothesis and showed a greater relative expression in the thrum morph than in the pin morph of *L. tenue* but not of *L. narbonense* (Figures 3.8 and 3.9). However, these results were not statistically significant. *GLO^T* expression was detected by qPCR in both floral morphs at all growth stages. The expression in the long-styled pin morph did not accord with the anticipated hemizygous model for the *S-locus*, and it is believed that a lack of primer specificity leading to the amplification of other members of the *GLO* gene family was the cause. The one recorded exception was the amplification of *GLO^T* in *L. narbonense*. However, the large and non-overlapping error bars suggest that this result should be interpreted with caution.

It was expected that floral genes would be most significantly expressed at GS1 and the start of GS2. However, there was no statistical evidence of differential expression of any tested candidate gene during qPCR analysis (see Results Section, Chapter 3) and qualitatively, expression was almost as high, if not higher at the more developed growth stages. Perhaps thrum-specific genes are expressed later than their floral homeotic counterparts. From the bioinformatic analysis of RNA-Seq data, there was evidence of more significant gene expression overall at GS1, which is consistent with the buds undergoing a very developmentally active phase of growth. However, none of the tested candidate genes are differentially expressed (Figure 2.24a). This result could be an artefact of human identification of the different growth stages. The field collections and the library creation for Illumina sequencing were undertaken before the beginning of this project; consequently there are likely to be small differences in identification between the growth-stage classification of individuals used in the RNA-Seq analysis and those which were freshly sampled for RNA extraction and qPCR analysis. Again, it must be reiterated that although progress has been made within the limitations of the initial dataset, any conclusions drawn are tentative.

Ecological data is often impacted by its environment and environmental factors can often influence experimental results. The wild individuals were all collected from field sites in Spain. Resultantly, many genes expressed in response to stresses including heat, herbivory, salinity etc. will have been up-regulated in these plants. This may have skewed the results for gene up-regulation, making more biologically-stressed individuals artifically appear more similar and potentially masking the differential expression effects of floral growth genes. The differential expression data which provides evidence for this is presented in Chapter 2; the top six most up-regulated genes at all three growth stages were stress-response genes, for example HSP90 chaperones. One way to combat these detrimental effects may be to grow all of the wild species in the greenhouses at Durham. In this way, individuals would be subjected to a much smaller range of environmental differences. Ultimately, it may pay to think about expression in a more phylogenetic context, for example, by focusing sampling to sister species with contrasting breeding systems (e.g. *L. tenue* and *L. strictum*). There may, for example, be larger differences in expression in regions involved in floral morphological variation.

The findings of Li et al. (2016) in *Primula* were initially surprising as with a hemizygous mode of gene expression, only one copy of the heterostyly-carrying genes is present in an individual. Most plant and animal genes have two copies; often more in plants. Thus, if one copy is damaged, function can still be maintained from the second copy. If this same mode of inheritance acts in all families in which heterostyly has arisen, it is surprising that heterostyly is as stable as it is, as any genetic damage to key genes could not be repaired. Strong selection on single copies would favour evolutionary conservatism over long periods.

In conclusion, although this study failed to definitively find genes associated for the heterostylous phenotype in wild and cultivated *Linum* species, some promising avenues for future research have been identified. The failure to demonstrate that the selected candidate genes, including GLO^T, CYP^T and PUM^T have any involvement in heterostyly may be because of methodological problems in the bioinformatic pipeline (as discussed in Chapter 2), or similarly as a result of issues concerned with primer design experimental analysis of the candidate genes by qPCR (discussed in Chapter 3). However, again, it may be because the physiologies of different species within the genus *Linum* fundamentally differ. This is surely an area worthy of future attention and more RNA-

Sequencing. Recently, the Slotte laboratory at the University of Stockholm has been awarded a European Research Council grant to investigate heterostyly at genus level. The project is entitled “SuperGenE - Supergene evolution in a classic plant system" and it is anticipated that this deep genomic analysis will provide fascinating insights into the origin and control of heterostyly in *Linum* (Slotte, 2017).

4.1 Future Work

One of the most obvious next steps would be to collect more data in an attempt to replicate the findings and to increase the number of technical replicates. Challenging the current paradigm of the origins of heterostyly would require strong evidence. Generating more RNA-Seq data, especially gathering three or four different samples of each wild species would facilitate bioinformatic analysis using DESeq2. Equally, qpCR plate design could be improved by testing several replicates of each wild-species at each growth stage, rather than maximising the number of wild species tested for each primer. *L. suffruticosum* samples, present in our collection, were not tested owing to time constraints. In an extension to this project, qPCR testing of *L. suffruticosum* would certainly be conducted, as its three-dimensional reciprocity is of significant biological interest.

With more time and experience, significant improvements could be made to the RNA-Seq analysis pipeline. Mapping percentages of 38-50% mean that so much of the read data is being lost at even early stage. Perhaps one issue was attempting to look at too many species at once, and it would have been better to have been less ambitious, focusing on perfecting the mapping of one or more closely related heterostylous pairs. Having more time devoted to fieldwork would also be beneficial; given more time to conduct fieldwork in order to collect the required individuals, it may be very useful to analyse heterostylous and pairs of otherwise evolutionarily similar *Linum* species, for example, *L. tenue* and *L. strictum* or *L. tenuifolium* and *L. suffruticosum*. Gaining insight into the similarities of gene expression between closely related individuals with or without heterostyly, not only may help to narrow down candidate genes for heterostyly. Equally, it may yield some insight into the evolutionary history of heterostyly.

There is potential to conduct functional annotation using GO enrichment. By analysing the GO profile, sets of genes can be interpreted in terms of their functional profile. Thus, it may be possible to more fully determine the types of genes associated with pin or thrum expression.

By using RNA-Seq reads for the bioinformatic analysis, rather than whole genome sequence data, the analysis is limited to differences evident in coding regions alone. Additionally, if there are high degrees of allele-specific expression, potential polymophisms will fail to be observed as the reads will not map (Nowak et al., 2015). In order to progress the search for the *'S-locus'* in the genus *Linum* it would be advantageous to use the RNA-Seq data derived from the 10 wild species for this project to create and annotate transcriptomes. This task, and processing raw sequence data in general, requires a huge volume of work with a scope far exceeding that of a Master's project. In order to draw useful conclusions from this largely untapped dataset, a significant amount of pre-processing will have to be conducted. Transcriptomes will first need to be *de novo* assembled, then annotated and this process is likely to take several years. However, with the availability of annotated transcriptomes, it will be possible to design primers which are more effective across all species, thus facilitating and improving the investigations conducted as a part of this project. With this information, it will also be possible to line up amino acid sequences to ensure that the candidate genes all appear to be functional orthologues. Alternatively, and as addressed in the discussion section of Chapter 3, individual primers for each of the wild species could be designed.

Linum is a very complex genus, within which exist species exhibiting a wide variety of different mating strategies. Combining this information with the fact that heterostyly is known to have evolved independently on at least 28 different occasions, it stands to reason that heterostyly may occur by a different mechanism in some, if not all heteromorphic genera. This, combined with the evolutionary distance between *Linum* and the other heteromorphic species increases the strength of this hypothesis. It would be a grand claim to suggest that this has definitely occurred. However, the possibility must not be discounted, and further work into the origin of heterostyly in *Linum* will hopefully help to elucidate the unusual behaviour of heterostyly candidate genes and more thoroughly confirm or refute the hypotheses presented in this work. Advances in modern analysis techniques are on the point of providing the key to understanding many complicated genetic interactions. The fundamental flaws in experimental design of this work have been recognised, however, notwithstanding, this project has been a useful fact-finding exercise, which can form the basis of future research in this exciting and topical research area.

References

Abbasi, N., Park, Y., & Choi, S. (2011). Pumilio Puf domain RNA-binding proteins in *Arabidopsis*. *Plant Signal Behaviour*, *6*(3), 364-368. doi: 10.4161/psb.6.3.14380

Aii, J., Nagano, M., Penner, G., Campbell, C., & Adachi, T. (1999). Identification of RAPD markers linked to the homostylar (*Ho*) gene in buckwheat. *Breed. Sci.*, *48*, 59-62.

Anders, S., Pyl, P., & Huber, W. (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169. doi: 10.1093/bioinformatics/btu638

Andersen, C., Ledet-Jensen, J., & Ørntoft, T. (2004). Normalization of real-time quantitative RT-PCR data: a model based variance estimation approach to identify genes suited for normalization - applied to bladder- and colon-cancer data-sets. *Cancer Research*, *64*, 5245-5250.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* Retrieved from `http://www.bioinformatics.babraham.ac.uk/projects/fastqc` (Accessed: 2017.03.02)

Armbruster, W., Pérez-Barrales, R., Arroyo, J., Edwards, M., & Vargas, P. (2006). Three-dimensional reciprocity of floral morphs in wild flax (linum suffruticosum): a new twist on heterostyly. *New Phytologist*, *171*, 581-590.

Athanasiou, A., Khosravi, D., Tamari, F., & Shore, J. (2003). Characterization and localization of short-specific polygalacturonase in distylous *Turnera subulata* (Tureraceae). *American Journal of Botany*, *90*, 675-682.

Athanasiou, A., & Shore, J. (1997). Morph-specific proteins in pollen and styles of distylous *Turnera* (Turneraceae). *Genetics*, *146*(2), 669-679.

Babraham Bioinformatics. (2017, December). *FastQC Report summary:* $bad_sequence.txt.Retrievedfrom\ (Accessed : 2018.03.13)$

Babraham Bioinformatics. (2018). *Index of /projects/fastqc/help/3 analysis modules.* Retrieved from `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3`

Analysis Modules/ (Accessed: 2018.03.12)

Badia, J. (n.d.). *El Medi Natural Del Bages - Els prats.* Retrieved from http://ichn.iec.cat/Bages/brolles/Imatges%20grans/Linum%20strictum.htm (Accessed: 2018.02.15)

Barrett, S. (1992). *Evolution and function of heterostyly.* Berlin: Springer-Verlag.

Barrett, S. (2002). The evolution of plant sexual diversity. *Nature Reviews Genetics*, *3*, 274-284.

Barrett, S. (2013). The evolution of plant reproductive systems: how often are transitions irreversible? *Proc. Biol Sci.*, *280*(1765), 20130913. doi: 10.1098/rspb.2013.0913

Barrett, S., & Shore, J. (2008). Self-incompatibility in flowering plants: evolution, diversity and mechanisms. In V. Franklin-Tong (Ed.), (p. 3-32). Berlin: Springer-Verlag.

Basiran, N., Armitage, P., Scott, R., & Draper, J. (1987). Genetic transformation of flax (*Linum usitatissimum* by *Agrobacterium tumefaciens*: regeneration of transformed shoots via a callus phase. *Plant Cell Reports*, *6*(5), 396-399.

Bateson, W., & Gregory, R. (1905). On the Inheritance of Heterostylism in Primula. *Proceedings of the Royal Society of London B*, *76*, 581-585.

Batushansky, A., Kirma, M., Grillich, N., Pham, P., Rentsch, D., Galili, G., … Fait, A. (2015). The transporter GAT1 plays an important role in GABA-mediated carbon-nitrogen interactions in *Arabidopsis*. *Frontiers in Plant Science*, *6*, 785. doi: . http://doi.org/10.3389/fpls.2015.00785

Becker, M. (2015). *Is RNA-seq reads number [sic] related to the quality of RNA preparation?* Retrieved from https://www.researchgate.net/post/Is$_{R}NA -$ $seq_{r}eads_{n}umber_{r}elated_{t}o_{t}he_{q}uality_{o}f_{R}NA_{p}reparation$ (Accessed: 2018.02.21)

Biosoft, P. (n.d.). *PCR Primer Design Guidelines.* Retrieved from http://www.premierbiosoft.com/tech$_{n}otes/PCR_{P}rimer_{D}esign.html$ (Accessed: 2018.02.20)

Biostars. (n.d.). *Biostars: Bioinformatics Explained.* Retrieved from https://www.biostars.org/ (Accessed: 2018.01.29)

Blow, N. (2009). Transcriptomics: The digital generation. *Nature*, *458*, 239-242.

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

Bolsheva, N., Melnikova, N., Kirov, I., Speranskaya, A., Krinitsina, A., Dimitriev, A., … Muravenko, O. (2017). Evolution of blue-flowered species of genus *Linum* based on high-throughput sequencing of ribosomal RNA genes. *BMC Evolutionary Biology*, *17*(Suppl 2), 253.

Bouché, N., & Fromm, H. (2004). GABA in plants; just a metabolite? *Trends in Plant Science*, *9*(3), 110-115. doi: https://doi.org/10.1016/j.tplants.2004.01.006

Bray, N., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525-527. doi: doi:10.1038/nbt.3519

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

Bushnell, B. (2015a). BBMap short-read aligner, and other bioinformatics tools. Retrieved from `sourceforge.net/projects/bbmap` (Accessed: 2018.01.08)

Bushnell, B. (2015b, April). *Question: How does sliding window work in Trimmomatic.* Retrieved from `https://www.biostars.org/p/97848/` (Retrieved 2018.03.10)

Bustin, S., Benes, V., Garson, J., Hellemans, J., Hugget, J., Kubista, M., . . . Wittwer, C. (2009). The MIQE Guidelines: *M*inimum *I*nformation for Publication of *Q*uantitative Real-Time PCR *E*xperiments. *Clinical Chemistry*, *55*(4). doi: 10.1373/clinchem.2008.112797

Charlesworth, D. (1992). Anti-inbreeding Systems. *Trends in Ecology & Evolution*, *7*(12), 428-429.

Charlesworth, D., & Charlesworth, B. (1979). A model for the evolution of distyly. *The American Naturalist*, *114*, 467-498. doi: 10.1373/clinchem.2008.112797

Charlesworth, D., & Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annual Review of Ecological Systems*, *18*, 237-268.

Chen, T.-W., Gan, R.-C., Kang, Y.-K., Chien, K.-Y., Liao, W.-C., Chen, C.-C., . . . Tang, P. (2017). FunctionAnnotator, a versatile and efficient web tool for non-model organism annotation. *Scientific Reports*, *7*(10430). doi: doi:10.1038/s41598-017-10952-4

Chen, T.-W., Gan, R.-C., Wu, T., Huang, P.-J., Lee, C.-Y., Chen, Y.-Y., . . . Tang, P. (2012). FastAnnotator - an efficient transcription annotation web tool. *BMC Genomics*, *13*(7), S9. doi: https://doi.org/10.1186/1471-2164-13-S7-S9

Cloutier, S., Cronk, Q., Cullis, C., Dash, P., Datla, R., Deyholos, M., . . . Zhu, S. (2014). Genomic data of Flax (*Liinum usitatissimum*). *GigaScience Database*. doi: http://dx.doi.org/10.5524/100081

Cocker, J., Webster, M., Li, J., Wright, J., Kaithakottil, G., Swarbeck, D., & Gilmartin, P. (2015). *Oakleaf*: an S locus-linked mutation of *Primula vulgaris* that affects leaf and flower development. *New Phytologist*, *208*(1), 149-161.

Coen, E., & Meyerowitz, E. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature*, *353*, 31-37. doi: 10.1038/353031a0

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-Seq data analysis. *Genome Biology*, *17*, 13. doi: 10.1186/s13059-016-0881-8

Cui, Y., Brugière, N., Jackman, L., Bi, Y., & Rothstein, S. (1999). Structural and transcriptional comparative analysis of the S locus regions in two self-incompatible *Brassica napus* lines. *Plant Cell*, *11*, 2216-2231.

Dalquen, D., & Dessimoz, C. (2013). Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biol Evol*.

Darwin, C. (1862). On the two forms or dimorphic condition in the species of *Primula*, and on their remarkable sexual relations. *Journal of the proceedings of the Linnean Society, Botany.*, 77-96.

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

Darwin, C. (1877). *The Different Forms of Flowers on Plants of the Same Species*. John Murray.

Devlin, Z. (2006). *Information on Fairy Flax.* Retrieved from

de Vos, J., Keller, B., Zhang, L., Nowak, M., & Conti, E. (2018). Mixed mating in homostylous species: Genetic and experimental evidence from an alpine plant with variable herkogamy *Primula halleri*. *International Journal of Plant Science*, *179*(2), 87-99. doi: 10.1086/695527

Dobin, A. (2013, April). *Typical alignment mapping percentage with genome?* Retrieved from `http://seqanswers.com/forums/showthread.php?t=29769` (Retrieved 2018.03.09)

Dobin, A., Davies, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. (2013). STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*, *29*(1), 15-21. doi: 10.1093/bioinformatics/bts635.

Dowrick, V. (1956). Heterostyly and Homostyly in *Primula obconica*. *Heredity*, *10*, 219-236.

Dulberger, R. (1974). Structural Dimorphism of Stigmatic Papillae in Distylous *Linum* Species. *Amerian Journal of Botany*, *61*(3), 238-243.

Dulberger, R. (1992). Evolution and function of heterostyly. In S. Barrett (Ed.), (p. 41-84). Springer.

eCSeq Bioinformatics. (2016, August). *Trimming adapter sequences - is it necessary?*

Retrieved from https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

Elhiti, M., & Stasolla, C. (2009). Structure and function of homeodomain-leucine zipper (HD-Zip) proteins. *Plant Signaling & Behaviour*, *4*(2), 86-88.

Ensembl. (2017, August). *GFF/GTF File Format - Definition and supported options.* Retrieved from https://www.ensembl.org/info/website/upload/gff.html (Accessed: 2017.11.23)

Ernst, A. (1936a). Erblichkeitsforschungen an calycanthemen Primeln. *Theoretical and Applied Genetics*, *8*, 313-324.

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

Ernst, A. (1955). Self-fertility in monomorphic primulas. *Genetica.*

Eveland, A., McCarty, D., & Koch, K. (2008). Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiology*, *146*(1), 32-44.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048. doi: https://doi.org/10.1093/bioinformatics/btw354

Ewing, B., Hillier, L., Wendl, M., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, *8*(3), 175-185.

Fenwick, D. (2005, August). *Aphotoflora - Linum usitatissimum - Flax (Linaceae images).* Retrieved from http://www.aphotoflora.com/d$_linum_usitatissimum_flax.html$ (Accessed: 2018.02.15)

Fesenko, N., Fesenko, I., & Ohnishi, O. (2006). Homostyly of two momorphological different lineages of *Fagopyrum homotropicum* Ohnishi is determined by locus *S4* which is an S-locus related gene in the linkage group #4. *Fagopyrum*, *23*, 11-15.

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-Seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, *14*(2), 130-142. doi: https://doi.org/10.1093/bfgp/elu035

Franklin-Tong, V. (2008). *Self-incompatibility in flowering plants: evolution, diversity, and mechanisms*. Berlin: Springer-Verlag.

Frazee, A. (2015). *High-Resolution Gene Expression Analysis* (Unpublished doctoral dissertation). The Johns Hopkins University.

Frech, C. (2016, December). *% of read unmapped: too short is HUGE.* GitHub. Retrieved from `https://github.com/alexdobin/STAR/issues/169` (Accessed: 2018.01.07)

Ganders, F. (1979). The biology of heterostyly. *New Zealand Journal of Botany*, *17*, 607-635.

Gilmartin, P. (2017, October). *The Primula S locus: gene function and the maintenance and breakdown of heterostyly.* Retrieved from `https://bbsrc.ukri.org/research/grants/grants/AwardDetails.aspx?FundingReference=` (Accessed: 2017.11,13)

Gilmartin, P., & Li, J. (2010). Delineation of the S locus in *Turnera subulata*. Homing in on heterostyly. *Heredity*, *105*, 161-162. doi: doi:10.1038/hdy.2010.69

Godtler, M. (2009, June). *Le Lin à feuilles étroites. Linum tenuifolium L.* Retrieved from `http://www.hautemarne-nature.com/fleurs-p32.html` (Accessed: 2018.02.15)

Goldstrohm, A., Seay, D., Hook, B., & Wickens, M. (2007). PUF protein-mediated deadenylation is catalyzed by Ccr4p. *Journal of Biological Chemistry*, *282*, 109-114. doi: 10.1074/jbc.M609413200

Goodstein, D., Shu, S., Howson, R., Neupane, R., Hayes, R., Fazo, J., ... Rokhsar, D. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186. doi: 10.1093/nar/gkr944

Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644-652. doi: doi: 10.1038/nbt.1883

Gregory, R., De Winton, D., & Bateson, M. (1923). Genetics of *Primula sinensis*. *Journal of Genetics*, *13*, 219-253.

Hansen, K., Brenner, S., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, *38*(12), e131. doi: 10.1093/nar/gkq224

Hillman, G. (1975). The plant remains from Tell Abu Hureyra: A preliminary report. *Proc Prehist Soc*, *41*, 70-73.

Hu, W., Chen, L., Qiu, X., Lu, H., Wei, J., Bai, Y., ... Shen, G. (2016). Morphological, Physiological and Proteomic Analyses Provide Insights into the Improvement of Castor Bean Productivity of a Dwarf Variety in Comparing with a High-Stalk Variety. *Frontiers in Plant Science*, *7*, 1473S. doi: https://doi.org/10.3389/fpls.2016.01473

Huis, R., Hawkins, S., & Neutelings, G. (2010). Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum L.*). *BMC*

Plant Biology, *10*(71).

Huu, C., Kappel, C., Keller, B., Sicard, A., Takebayashi, Y., Breuninger, H., … Lenhard, M. (2016). Presence versus absence of CYP73A50 underlies the style-length dimorphism in primroses. *eLife*, *5*, e17956.

Illumina. (2017). *Study gene expression using RNA sequencing: Introduction to RNA Sequencing.* Retrieved from `https://emea.illumina.com/techniques/sequencing/rna-sequencing.html` (Accessed: 2018.01.26)

Irish, V. (2017). Primer: The ABC model of floral development. *Current Biology*, *27*, R853-R909.

Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., … Gao, G. (2015). An *Arabidopsis* transcriptional regulatory map reveals distinct evolutionary features of novel transcription factors. *Molecular Biology and Evolution*, *32*(7), 1767-1773.

Kadam, B., & Patel, S. (1938). Anthesis in flax. *Journal of the American Society of Agronomy*, *30*, 932-940.

Kajeka. (2014). *Biolayout Discontinued - Graphia Professional.* Retrieved from `https://kajeka.com/biolayout-express-upgrade/` (Accessed: 2018.02.17)

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27-30*, 27-30.

Kappel, C., Huu, N., C, & Lenhard, M. (2017). A short story gets longer: recent insights into the molecular basis of heterostyly. *Journal of Experimental Botany*, *68*(21-22), 5719-5730.

Katz, Y., Wang, E., Airoldi, E., & Burge, C. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009-1015.

Keller, B., Thompson, J., & Conti, E. (2014). Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin’s primroses: Evidence from experimental studies. *Functional Ecology*, *28*, 1413-1425.

Khang, T., & Lau, C. (2015). Getting the most out of RNA-seq data analysis. *PeerJ*, *3*, e1360.

Khosravi, C., D.and Yang, Siu, K., & Shore, J. (2004). High level of α-dioxygenase in short styles of distylous *Turnera* spp. *International Journal of Plant Science*, *165*, 995-1006.

Kim, D., Langmead, B., & Salzberg, S. (2015). HISAT: a fast universal aligner with low memory requirements. *Nature Methods*, *12*, 357-360. doi: doi:10.1038/nmeth.3317

Kumar, S., Stecher, G., & Tamura, K. (2015). MEGA7: Molecular Evolutionary Genetics

Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*(7), 1870-1874.

Kurian, V., & Richards, A. (1997). A new recombinant in the heteromorphy 'S' supergene in *Primula*. *Heredity*, *78*, 383-390.

Kurotani, A., Yamada, Y., Shinozaki, K., Kuroda, Y., & Sakurai, T. (2015). Plant-prAS: A database of physiocochemical and structural properties and novel function regions in plant proteomes. *Plant and Cell Physiology*, *56*(1), e11.

Kvavadze, E., Bar-Yosef, O., Belfer-Cohen, A., Boaretto, E., Jakeli, N., Matskevich, Z., & Meshveliani, T. (2009). 30,000-year-old wild flax fibers. *Science*, *325*, 1359. doi: 10.1126/science.1175404

Labonne, J., Goultiaeva, A., & Shore, J. (2009). High-resolution mapping of the S-locus in Turnera leads to the discovery of three genes tightly associated with the S-alleles. *Molecular Genetic Genomics*, *281*, 673-685.

Labonne, J., & Shore, J. (2011). Positional cloning of the s haplotype determining the floral and incompatibility phenotype of the long-styled morph of distylous *Turnera subulata*. *Molecular Genetics and Genomics*, *285*, 101-111.

Labonne, J., Tamari, F., & Shore, J. (2010). Characterization of X-ray-generated floral mutants carrying deletions at the S-locus of distylous *Turnera subulata*. *Heredity*, *105*, 235-243. doi: doi:10.1038/hdy.2010.39

Lai, Z., Ma, W., Han, B., Liang, L., Zhang, Y., Hong, G., & Xue, Y. (2002). An F-box gene linked to the Self-Incompatibility (S) locus of *Antirrhinum* is expressed specifically in pollen and tapetum. *Plant Molec*, *50*, 29-42.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, *10*(R25). doi: 10.1186/gb-2009-10-3-r25

Le Petit Herboriste. (1998). *Lin a Feuilles de Salsola 3 (Linum suffruticosum).* Retrieved from `http://www.lepetitherboriste.net/photos/linafeuillesdesalsola3.html` (Accessed: 2018.02.15)

Leonard, D. (2016). *Molecular pathology in clinical practice* (2nd ed.; D. Leonard, Ed.). Springer.

Les, D. (2017). *Aquatic dicotyledons of North America: Ecology, life history and systematics* (1st ed.). Boca Raton, Florida: CRC Press.

Lewis, D. (1943). The physiology of incompatibility in plants II. *Linum grandiflorum*. *Annals of Botany- London*, *7*, 115-124.

Lewis, D. (1954). Comparative incompatibility in Angiosperms and Fungi. *Advances in*

Genetics, *6*, 235-285.

Lewis, D., & Jones, D. (1992). Evolution and function of heterostyly. In C. H. Barrett S (Ed.), (p. 129-150). Berlin, Heidelberg: Springer.

Li, H., Handsaker, B., Wysoker, A., Fennel, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, J., Cocker, J., Wright, J., Webster, M., McMullan, M., Dyer, S., ... Gilmartin, P. (2016). Genetic architecture and Evolution of the S locus supergene in *Primula vulgaris*. *Nature Plants*, *2*. doi: 10.1038/NPLANTS2016.188

Li, J., Dudas, B., Webster, M., Cook, H., Davies, B., & Gilmartin, P. (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 of the United States of America*, *107*, 5664-5668.

Li, J., Soroka, J., & Buchner, J. (2012). The Hsp90 chaperone machinery: Conformationaly dynamics and regulation by co-chaperones. *Biochimica et Biophysica Acta*, *1823*(3), 624-635.

Li, J., Webster, M., Dudas, B., Cook, H., Manfield, I., Davies, B., & Gilmartin, P. (2008). The S locus-linked Primula homeotic mutant sepaloid shows characteristics of B-function mutant but does not result from mutation in a B-function gene. *The Plant Journal*, *56*, 1-12.

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

Lin, B. (2011). Resilience in agriculture through crop diversification: Adaptive management for environmental change. *BioScience*, *61*(3), 183-193. doi: https://doi.org/10.1525/bio.2011.61.3.4

Liu, Y., Zhou, J., & White, K. (2014). RNA-seq differential expression studes: more sequence or more replication. *Bioinformatics*, *30*(3), 301-304. doi: https://doi.org/10.1093/bioinformatics/btt688

Lloyd, D., & Webb, C. (1992a). Evolution and function of heterostyly. In (p. 151-178). Berlin: Springer-Verlag. doi: http://dx.doi.org/10.1007/978-3-642-86656-2_6

Lloyd, D., & Webb, C. (1992b). Evolution and function of heterostyly. In (p. 179-207). Berlin: Springer-Verlag. doi: http://dx.doi.org/10.1007/978-3-642-86656-2_7

Love, M., Anders, S., Kim, V., & Huber, W. (2015). RNA-Seq workflow: gene-lebel exploratory analysis and differential expression. *F1000 Research*, *4*, 1070. doi: 10.12688/f1000research.7035.1

Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*. doi: 10.1101/002832

Manfield, I., Pavlov, V., Li, J., Cook, H., Hummel, F., & Gilmartin, P. (2005). Molecular characterization of DNA sequences from the *Primula vulgaris* S-locus. *Journal of Experimental Botany*(56), 1177-1188.

Marchler-Bauer, A., Lu, S., Anderson, J., Chitsaz, F., Derbyshire, M., DeWeese-Scott, C., ... Bryant, S. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, *39*, D225-9. doi: 10.1093/nar/gkq1189.

Mardis, E. (2011). A decades's perspective on DNA sequencing technology. *Nature*, *470*, 198-203. doi: 10.1038/nature09796

Mather, K. (1950). The genetical architecture of heterostyly in *Primula sinensis*. *Evolution*, *4*(4), 340-352.

McClure, B. (2016). The genetic basis of heterostyly. *Nature Plants*, *2*. doi: 10.1038/NPLANTS.2016.184

McCubbin, A., Lee, C., & Hetrick, A. (2006). Identification of genes showing differential expression between morphs in developing flowers of *Primula vulgaris*. *Sexual Plant Reproduction*, *19*, 63-72.

McDill, J., Repplinger, M., Simpson, B., B., & Kadereit, J. (2009). The phylogeny of linum and linaceae subfamily linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. *Systematic Botany*, *34*(2), 386-405.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. (2010). The genome analysis toolkit: a MapReduce framework for analysing next-generation DNA sequencing data. *Genome Research*, *20*, 1297-1303.

Miljuš-Đukić, J., Nincović, S., Radović, S., Maksimović, V., Brkljačić, J., & Nešković, M. (2004). Detection of proteins possibly involved in Self-Incompatibility response in distylous buckwheat. *Biologia Plantarum*, *48*(2), 293-296.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621-628. doi: DOI:10.1038/NMETH.1226

Mrugala, F. (n.d.). *Lin - Linum campanulatum (fam linacees) (Europe meridionale) (03) (photo f. mrugala).jpg.* Retrieved from `http://www.mrugala.net/Nature/Plantes/Photos/index.php?page=60` (Accessed: 2018.02.15)

Nicholas, K., & Nicholas, H. (1997). *GeneDoc:a tool for editing and annotating multiple sequence alignments.* Retrieved from `http://www.psc.edu/biomed/genedoc`

(1997) (Accessed: 2017.10.25)

Nicholls, M. (1985). The evolutionary breakdown of distyly in *Linum tenuifolium* (Linaceae). *Plant Systematics and Evolution*, *150*, 291-301.

Notredame, C., Higgins, D., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*, 205-217. doi: doi:10.1006/jmbi.2000.4042

Nowak, M., Russo, G., Schlapbach, R., Huu, C., Lenhard, M., & Conti, E. (2015). The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly,. *Genome Biology*.

Oshlack, A., Robinson, M., & Young, M. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 220. doi: 10.1186/gb-2010-11-12-220

Patro, R., Duggal, G., Love, M., Irizarry, R., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*, 417-419. doi: doi:10.1038/nmeth.4197

PFAF. (2012). *Linum bienne - Mill.* Retrieved from https://www.pfaf.org/USER/Plant.aspx?LatinName=Linum+bienne (Accessed: 2018.03.12)

Pfaffl, M., Tichopad, A., Prgomet, C., & Neuvians, T. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper - Excel-based tool using pair-wise correlations. *Biotechnology Letters*, *26*(6), 509-515.

QGIS Development Team. (n.d.). QGIS Geographic Information System [Computer software manual]. Retrieved from http://qgis.osgeo.org (Accessed: 2018.03.01)

Quiagen. (n.d.). *Effects of low A260/A230 ratios in RNA preparations on downstream applications.* Retrieved from https://www.qiagen.com/gb/resources/faq?id=c59936fb-4f1e-4191-9c16-ff083cb24574la (Accessed: 2018.01.23)

Ramalho, C. (2012). *Linum setaceum Brot.* Retrieved from http://flora-on.pt/1Linum+setaceum (Accessed: 2018.02.15)

Rebrikov, D., Desai, S., Siebert, P., & Lukyanov, S. (2004). Suppression subtractive hybridization. *Methods in Molecular Biology*, *258*, 107-134.

Renner, S. (2014). The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *American Journal of Botany*, *101*, 1588-1596. doi: http://dx.doi.org/10.3732/ajb.1400196

Riechmann, J., & Meyerowitz, E. (1997). MADS domain proteins in plant development.

Biological Chemistry, *378*, 1079-1101.

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. (2011). Interactive Genomics Viewer. *Nature Biotechnology*, *29*, 24-26.

Robles, J., Qureshi, S., Stephen, S., Wilson, S., Burden, C., & Taylor, J. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, *13*, 484.

Rogers, C. (1979). Distyly and pollen dimorphism in *Linum suffruticosum* (Linaceae). *Plant Systematics and Evolution*, *131*(1), 127-132.

Rogers, C. (1982). The systematics of Linum sect. Linopsis (Linaceae). *Plant Systematics and Evolution*, *140*(2), 225-234.

Ruiz Martín, J., Santos-Gally, R., Escudero, M., Midgley, J., Perez-Barrales, R., & Arroyo, J. (2018). Style polymorphism in *Linum* (Linaceae): a case of Mediterranean parallel evolution? *Plant Biology*, *20*, 111.

Ryan, P., Ó'Maoiléidigh, D., Drost, H., Kwaśniewska, K., Gabel, A., Grosse, I., ... Wellmer, F. (2015). Patterns of gene expression during *Arabidopsis* flower development from the time of initiation to maturation. *BMC Genomics*, *16*(1), 488.

Salguero Quiles, P. (2004). *Linum narbonense.* Retrieved from `http://enciclopedia.us.es/index.php/Linum_narbonense` (Accessed: 2018.02.15)

Sasaki, Y., & Nagano, Y. (2004). Plant Acetyl-CoA Carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Bioscience, Biotechnology, and Biochemistry*.

Schurch, N., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., ... Barton, G. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*, 1-13.

SeqAnswers. (2007). *Tech summary: Illumina's Solexa sequencing technology.* Retrieved from `seqanswers.com/forums/showthread.php?t=21` (Accessed: 2018.01.26)

Shore, J., Arbo, M., & Fernandez, A. (2006). Breeding system variation, genetics and evolution in the Turneraceae. *New Phytologist*, *171*, 539-551.

Sievers, F., Wilm, A., Dineen, D., Gibson, T., Karplus, K., Li, W., ... Higgins, D. (2011). Fast, scalable generagene of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. doi: 10.1038/msb.2011.75

Sims, D., Sudbery, I., Ilott, N., Heger, A., & Ponting, C. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121-132. doi: doi: 10.1038/nrg3642.

Simão, F., Waterhouse, R., Ioannidis, P., Kriventseva, E., & Zdobnov, E. (2015). BUSCO:

assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. doi: doi: 10.1093/bioinformatics/btv351

Slotte, T. (2017). *Slotte lab: Projects supergene evolution in a classic plant system - genomic studies of distyly in Linum.* Retrieved from `https://tanjaslottelab.se/projects-3/` (Accessed: 2018.01.29)

Smith, M. (2015). *Assembly and annotation of ssequence surrounding the S locus in Primula vulgaris* (Unpublished doctoral dissertation). Durham University.

Swenson, N., & Jones, F. A. (2017). Community transcriptomics, genomics and the problem of species co-occurrence. *Journal of Ecology*, *105*, 563-568.

Tamari, F., Athanasiou, A., & Shore, J. (2001). Pollen tube growth and inhibition in distylous and homostylous *Turnera* and *Piriqueta* (Turneraceae). *Canadian Journal of Botany*, *79*(5), 578-591.

Tamari, F., & Shore, J. (2004). Distribution of style and pollen polygalacturonases among distylous and homostylous *Turnera* and *Piriqueta* spp. (turneraceae). *Heredity*, *92*, 380-385. doi: doi:10.1038/sj.hdy.6800390

Tarazona, S., Garcıa-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, *21*, 2213-2223.

Tatusov, R., Koonin, E., & Lipman, D. (1997). A genomic perspective on protein families. *Science*, *278*(5338), 631-637.

Theocharidis, A., van Dongen, S., Enright, A., & Freeman, T. (2009). Network visualization and analysis of gene expression data using Biolayout Express (3D). *Nature Protocols*, *4*, 1535-1550.

Thompson, M., & Jiggins, C. (2014). Supergenes and their role in evolution. *Heredity*, *113*, 1-8.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562-578.

Ushijima, K., Ikeda, K., Nakano, R., Matsubara, M., Tsuda, Y., & Kubo, Y. (2015). Genetic control of floral morph and petal pigmentation in *Linum grandiflorum* Desf., a heterostylous flax. *The Horticulture Journal*, *Advance Online Publication*.

Ushijima, K., Nakano, R., Bando, M., Shigezane, Y., Ikeda, K., Namba, Y., . . . Kubo, Y. (2012). Isolation of the floral morph-related genes in heterostylous flax (linum grandiflorum): the genetic polymorphism and the transcriptional and post-transcriptional regulations of the S locus. *The Plant Journal*, *69*, 317-331.

Venglat, P., Xiang, D., Qiu, S., Stone, S., Tibiche, C., Cram, D., . . . Datla, R. (2011). Gene expression analysis of flax seed development. *BMC Plant Biology*, *11*(74).

Viatour, L. (2006). *Linum bienne en Belgique à Hamois.* Retrieved from `https://commons.wikimedia.org/wiki/File:Linum`$_bienne_Luc_Viatour.JPG$ (Accessed: 2018.02.15)

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63. doi: 10.1038/nrg2484

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., … Deyholos, M. (2012). The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant Journal*, *72*(3), 461-73. doi: doi: 10.1111/j.1365-313X.2012.05093.x

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, J. L., & Meyerowitz, E. (2006). Genome-wide analysis of gene expression during early *Arabidopsis* flower development. *PloS Genetics*, *2*(7), e117. doi: https://doi.org/10.1371/journal.pgen.0020117

Wijayanto, T., & McHughen, A. (1999). Genetic transformation of *Linum* by particle bombardment. *In Vitro Cellular and Developmental Biology*, *35*(6), 456-465.

Wolfe, L. (2001). Associations among multiple floral polymorphisms in *Linum pubescens* (Linaceae), a heterostylous plant. *International Journal of Plant Science*, *162*, 335-342.

Yang, I., & Kim, S. (2015). Analysis of whole transcriptome sequencing data: Workflow and software. *Genomics and Informatics*, *13*(4), 119-125. doi: http://dx.doi.org/10.5808/GI.2015.13.4.11

Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S., … Mori, M. (2016). Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. *DNA Research*, *23*, 215-224.

Yasui, Y., Mori, M., Aii, J., Abe, T., Matsumoto, D., Sato, S., … Ota, T. (2012). S-LOCUS EARLY FLOWERING 3 is exclusively present in the genomes of short-styled buckwheat plants that exhibit heterostylous Self-Incompatibility. *PLoS ONE*, *7*(2), e31264.

Yasui, Y., Wang, Y., Ohnishi, O., & Campbell, C. (2004). Amplified fragment length polymorphism linkage analysis of common buckwheat (*Fagopyrum esculentum*) and its wild self-pollinated relative *Fagopyrum homotropicum*. *Genome*, *47*, 345-351.

Yjiangnan. (2017, May). *Are published RNA seq data analyses often wrong in calculating p-values and fdr?* Retrieved from `ttps://support.bioconductor.org/p/95949/` (Retrieved: 2018.07.03)

Zamore, P., Williamson, J., & Lehmann, R. (1997). The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins.

RNA, *3*(12), 1421-1433.

Zhao, W., He, X., Hoadley, K., Parker, J., Hayes, D., & Perou, C. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, *15*(1), 419. doi: 10.1186/1471-2164-15-419

Zhernakova, D., de Klerk, E., Westra, H.-J., Mastrokolias, A., Amini, S., Ariyurek, Y., ... Franke, L. (2013). DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genetics*, *9*(6), e100359.

Zohary, D., Hopf, M., & Weiss, E. (2012). *Domestication of plants in the old world* (Fourth Edition ed.). Great Clarendon Street Oxford, OX2 6DP: Oxford University Press.

Appendix A

DESeq2 Input

Listing A.1: Example DESEq2 input code. autogobble

```

library("DESeq2")
directory <- "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/HTSeq-Counts/Bbduk/Heterostyle/"
sampleFiles<-grep(‘GS’,list.files(directory),value=TRUE)

sampleCondition <- c(‘Thrum’, ‘Thrum’, ‘Thrum’, ‘Pin’, ‘Pin’,‘Thrum’, ‘Thrum’, ‘Thrum’, ‘Thrum’,\
 ‘Pin’,‘Pin’, ‘Pin’, ‘Pin’, ‘Thrum’, ‘Thrum’, ‘Thrum’, ‘Pin’, ‘Pin’, ‘Pin’, ‘Pin’,‘Pin’, ‘Thrum’,‘Thrum’)

sampleStage <- c(‘GS1’,‘GS2’,‘GS4’,‘GS1’,‘GS2’,‘GS1’,‘GS2’,‘GS3’, ‘GS4’, ‘GS1’, ‘GS2’,\
‘GS3’,‘GS4’,‘GS1’,‘GS2’,‘GS4’,‘GS1’,‘GS2’,‘GS4’, ‘GS1’,‘GS4’,‘GS1’,‘GS4’)
sampleSpecies <- c(‘L.campanulatum’,‘L.campanulatum’,‘L.campanulatum’,‘L.campanulatum’,‘L.campanulatum’,\
‘L.narbonense’,‘L.narbonense’,‘L.narbonense’,‘L.narbonense’,‘L.narbonense’,‘L.narbonense’,‘L.narbonense’,\
‘L.narbonense’,‘L.suffruticosum’,‘L.suffruticosum’,‘L.suffruticosum’,‘L.suffruticosum’,‘L.suffruticosum’,\
‘L.suffruticosum’,‘L.viscosum’,‘L.viscosum’,‘L.viscosum’,‘L.viscosum’)
sampleTable <- data.frame(sampleName = sampleFiles, fileName = sampleFiles, species=sampleSpecies, growthstage=sampleStage, condition=sampleCondition)

```

```
ddsHTSeq <- DESeqDataSetFromHTSeqCount(sampleTable=sampleTable, directory=directory, design=~condition + species + growthstage)
colData(ddsHTSeq)$condition<-factor(colData(ddsHTSeq)$condition, levels=c(‘Thrum’, ‘Pin’))
colData(ddsHTSeq)$condition<-factor(colData(ddsHTSeq)$growthstage, levels=c(‘GS1’,‘GS2’,‘GS3’,‘GS4’))
colData(ddsHTSeq)$species<-factor(colData(ddsHTSeq)$species, levels=c(‘L.campanulatum’,‘L.narbonense’,‘L.suffruticosum’,‘L.viscosum’))
dds <- DESeq(ddsHTSeq)
dds <- dds[rowSums(counts(dds))>5,]
res <- results(dds)
res <- res[order(res$padj),]
head(res)
dim(res)
summary(res)
counts <- assay(dds)
res <- as.data.frame(res)


metadata(res)$filterThreshold

#Fiddling with the Padj
noshrinkage <- DESeq(dds, betaPrior=FALSE)
noshrinkageres <- results(noshrinkage)
head(noshrinkageres)
noshrinkagecounts <- assay(noshrinkage)

#Extract candidate genes from counts file. NB S-ELF3 and PveGLO2 \\
extract the same contig

SELF3 <- counts[grepl("Contig_107032",rownames(counts))]
CYP734A50 <- counts[grepl("Contig_58866", rownames(counts))]
LgAP1 <- counts [grepl("Contig_128966", rownames(counts))]
LgMYB21 <- counts[grepl("Contig_65838", rownames(counts))]
LgSKS1 <- counts[grepl("Contig_21824", rownames(counts))]
TsRETRO <- counts[grepl("Contig_101376", rownames(counts))]
TSS1 <- counts[grepl("TR6663", rownames(counts))]
```

```r
TkNACE <- counts[grepl("Contig_94982", rownames(counts))]
TkST1 <- counts[grepl("Contig_67632", rownames(counts))]
TPP1 <- counts[grepl("Contig_112511", rownames(counts))]


PumT <- counts[grepl("Contig_31347", rownames(counts))]
CcmT <- counts[grepl("Contig_76029", rownames(counts))]
CypT <- counts[grepl("Contig_86432", rownames(counts))]
GloT <- counts[grepl("Contig_107032", rownames(counts))]
KfbT <- counts[grepl("Contig_14963", rownames(counts))]


Expressed_Table <- data.frame(Sequence = sampleFiles, SELF3_PveGLO2, CYP734A50, LgMYB21, LgSKS1, TSS1, TkST1, TkNACE, PumT, CcmT, GloT) #TsRETRO empty #CypT empty

resultsNames(dds)

#VolcanoPlot
volcanotab = data.frame(logFC = res$log2FoldChange, negLogPval = -log10(res$pvalue))
head(volcanotab)
par(mar = c(5, 4, 4, 4))
plot(volcanotab, pch = 16, cex = 0.6, xlab = expression(log[2]~fold~change), ylab = expression(-log[10]~pvalue))
lfc = 2
pval = 0.01
signGenes = (abs(volcanotab$logFC) > lfc & volcanotab$negLogPval > -log10(pval))
points(volcanotab[signGenes, ], pch = 16, cex = 0.8, col = "red")
abline(h = -log10(pval), col = "green3", lty = 2)
abline(v = c(-lfc, lfc), col = "blue", lty = 2)
mtext(paste("pval =", pval), side = 4, at = -log10(pval), cex = 0.8, line = 0.5, las = 1)
mtext(c(paste("-", lfc, "fold"), paste("+", lfc, "fold")), side = 3, at = c(-lfc, lfc), cex = 0.8, line = 0.5)
dev.copy(png,"/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/BasicParametersGS3+4_Homostyle_vs_Heterostyle_SpeciesControlled_VolcanoPlotTHESIS.png" )
dev.off()

# Make a basic volcano plot
with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot", xlim=c(-2.5,2)))

# Add coloured points: red if padj<0.05, orange of log2FC>1, green if both)
with(subset(res, padj<.05 ), points(log2FoldChange, -log10(pvalue), pch=20, col="red"))
with(subset(res, abs(log2FoldChange)>1), points(log2FoldChange, -log10(pvalue), pch=20, col="orange"))
with(subset(res, padj<.05 & abs(log2FoldChange)>1), points(log2FoldChange, -log10(pvalue), pch=20, col="green"))

# Label points with the textxy function from the calibrate plot
library(calibrate)
with(subset(res, padj<.05 & abs(log2FoldChange)>1), textxy(log2FoldChange, -log10(pvalue), labs=Gene, cex=.8))


#Number of significant genes at level 1%
sum(res$padj < 0.01, na.rm=TRUE)

sigDownReg <- res[!is.na(res$padj), ]
sigDownReg <- sigDownReg[sigDownReg$padj < 0.01, ]
sigDownReg <- sigDownReg[order(sigDownReg$log2FoldChange),]
sigDownReg
write.csv(sigDownReg, file = "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/GS2sigDownReg-deseq.csv")


sigUpReg <- res[!is.na(res$padj),]
sigUpReg <- sigUpReg[sigUpReg$padj < 0.01,]
sigUpReg <- sigUpReg[order(sigUpReg$log2FoldChange, decreasing=TRUE),]
sigUpReg
write.csv(sigUpReg, file = "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/GS3_4sigUpReg-deseq.csv")

```

```
#Histogram of unadjusted p-values
hist(res$pvalue, breaks=50)
dev.copy(png, "GS1_PvT_Histogram")
dev.off
#MA plot of DE genes
plotMA(res, alpha=0.01)
dev.copy(png, "GS1_PvT_MAPlot")
dev.off

#MAKE AN MA PLOT TO SHOW DIFFERENCES IN GENE EXPRESSION
plotMA(dds, ylim=c(-15,15))
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/GS4_BasicParameters_SpeciesControlled_PinvsThrum.png")
dev.off()

#Plot counts - examine counts of reads for a single gene across groups
plotCounts(dds, gene=which.min(res$padj))
plotCounts(dds, gene=which.max(res$padj))


#Looking at individual candidates
plotCounts(dds, gene="L.tenue_CDS|Contig_107032")
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/GloT_GS1_Test_Plot.png")
dev.off()
plotCounts(dds, gene="L.tenue_CDS|Contig_58866")
plotCounts(dds, gene="L.tenue_CDS|Contig_65838")
plotCounts(dds, gene="L.tenue_CDS|Contig_67632")

d <- plotCounts(dds, gene=which.min(res$padj), returnData=TRUE)
library("ggplot2")
ggplot(d, aes(x=condition, y=count)) +
geom_point(position=position_jitter(w=0.1, h=0)) +
scale_y_log10(breaks=c(25, 100, 400))
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Pipelines/Sixth_Analysis_26.4.17/GS1_Heterostyle_Homostyle_Results/GS1_Heterostyle_HomostylePlotCountsMinpvalue.png")
dev.off()
BasicParametersGS1_Homostyle_vs_Heterostyle_SpeciesControlled_
```

```
   VolcanoPlotTHESIS.png

#Exporting results to CSV files

write.csv(as.data.frame(resOrdered), file ="/media/ellie/TOSHIBA
   _EXT/Pipelines/Sixth_Analysis_26.4.17/GS1_Heterostyle_
   Homostyle_Results/GS1_Heterostyle_Homostylefloral_morph_
   results.csv")
resSig_order <- subset(resOrdered, padj < 0.1)
resSig_order

#Extracting transformed values

rld <- rlog(dds, blind=FALSE)
vsd <- varianceStabilizingTransformation(dds, blind=FALSE)
vsd.fast <- vst(dds, blind=FALSE)
head(assay(rld),3)

#this gives log2(n+1)
ntd <- normTransform(dds)
library("vsn")
notAllZero <- (rowSums(counts(dds))>0)

meanSdPlot(assay(ntd)[notAllZero,])
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/
   BasicParameters_PvT_GS1MeanSdPlot_ntd.png")
dev.off()
meanSdPlot(assay(rld[notAllZero,]))
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/
   BasicParameters_PvT_GS1MeanSdPlot_rld.png")
dev.off()
meanSdPlot(assay(vsd[notAllZero,]))
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/
   BasicParameters_PvT_GS1MeanSdPlot_vsd.png")
dev.off()

#Principal component plot of the samples

plotPCA(rld,intgroup=‘species’)
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Aug_2017_Sequence_Run/
    Heterostyle_PCA_rld_intgroupspecies_OutlierRM.png")
dev.copy(png, "~/Documents/University/MSc/Masters_Thesis/Own_
    Template_Thesis/Heterostyle_PCA_rld_intgroupspecies_Bbduk.png
    ")
dev.off()
plotPCA(vsd, intgroup=‘species’)
plotPCA(rld, intgroup=‘condition’)
dev.copy(png,  "~/Documents/University/MSc/Masters_Thesis/Own_
    Template_Thesis/Heterostyle_PCA_rld_intgroupcondition_Bbduk.
    png")
dev.off()

plotPCA(rld, intgroup=‘growthstage’)
dev.copy(png,  "~/Documents/University/MSc/Masters_Thesis/Own_
    Template_Thesis/Heterostyle_PCA_rld_intgroupgrowthstage_Bbduk
    .png")

#Heatmap of count matrix

library("pheatmap")
select <- order(rowMeans(counts(dds,normalized=TRUE)),
decreasing=TRUE)[1:20]
df <- as.data.frame(colData(dds)[,c("species","condition","
    growthstage")])
pheatmap(assay(ntd)[select,], cluster_rows=FALSE, show_rownames=
    FALSE,
cluster_cols=FALSE, annotation_col=df)
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Pipelines/Sixth_Analysis
    _26.4.17/HTSeq/TwoFactor_heterostyle_GS1+4/CountMatrixHeatMap
    _ntd.png")
dev.off()

pheatmap(assay(rld)[select,], cluster_rows=TRUE, show_rownames=
    TRUE,
cluster_cols=TRUE, annotation_col=df)
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Pipelines/Sixth_Analysis
    _26.4.17/HTSeq/TwoFactor_heterostyle_GS1+4/CountMatrixHeatMap
    _rld.png")
dev.off()

```

```
pheatmap(assay(vsd)[select,], cluster_rows=FALSE, show_rownames=FALSE,
cluster_cols=FALSE, annotation_col=df)
dev.copy(png, "/media/ellie/TOSHIBA_EXT/Pipelines/Sixth_Analysis_26.4.17/HTSeq/TwoFactor_heterostyle_GS1+4/CountMatrixHeatMap_vsd.png")
dev.off()


#Heatmap of sample-to-sample distances
sampleDists <- dist(t(assay(rld)))
library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(rld$growth_stage, rld$floral_morph, sep="-")
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9,"Blues")) )(255)
pheatmap(sampleDistMatrix,
clustering_distance_rows = sampleDists,
clustering_distance_cols = sampleDists,
col = colors)
dev.copy(png, "~/Documents/University/MSc/Masters_Thesis/Own_Template_Thesis/Heterostyle_Heatmap_rld_Bbduk.png")
dev.off()
```

Appendix B

Supplementary Code

Listing B.1: Code for the AccessionGrab.py Python script used to extract the Accession number from the top BLAST hit and write it to a separate text file.

```
””” A program to extract fasta files from blast search
data into a separate text document”””

””” Ellie Desmond 18/10/16 ”””

#Open a .txt file
import csv
from sys import argv

script, infile, outfile = argv

data = list(csv.reader(open(infile, ’rb’), delimiter=’\t’))

text_file = open(outfile, ’w’)

for line in data:
text_file.write(line[1] + ’\n’)

text_file.close()
```

Listing B.2: Code for the fastaGrab.py python script (Foroozani, A., personal communication) used to extract wanted sequences from a specified fasta file and write them to a new file. Used to extract the sequences at desired contigs in the annotated *L. tenue* transcriptome

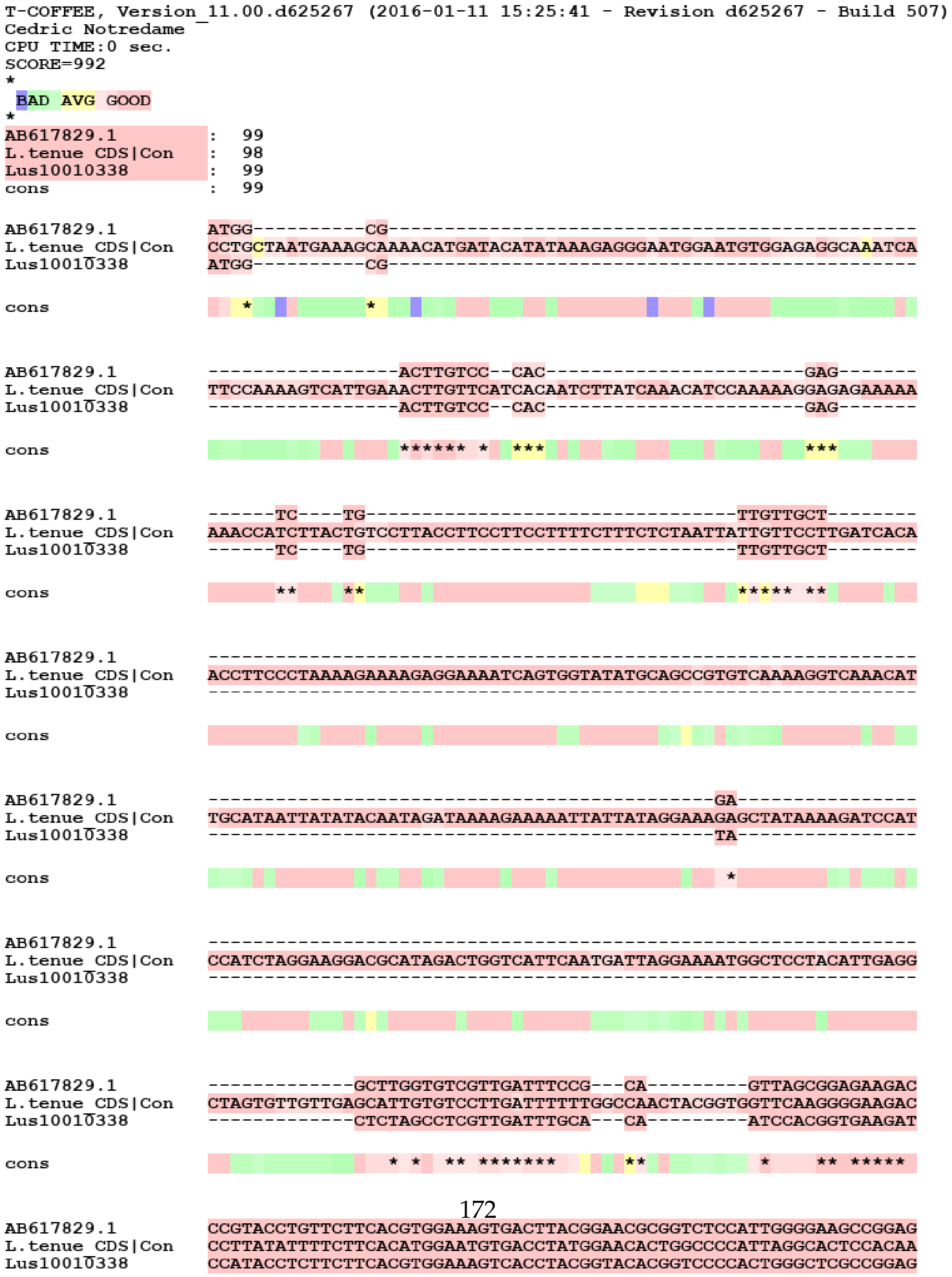
```
################
```

```
#fastaGrab.py
#Ali Foroozani (Feb 2016)
#

'''
This script extracts wanted sequences from a FASTA file.

Usage:
python fastaGrab.py -s <sequences_file.fa> \\
-w <wanted_sequences.txt> \\
-o <output_file.fa>
'''



#Import modules for the environment and set variables
########################################################

from optparse import OptionParser
from Bio import SeqIO

parser=OptionParser()
parser.add_option("-s", "--sequences", dest="s")
parser.add_option("-w", "--wanted", dest="w")
parser.add_option("-o", "--output", dest="o")
(options, args) = parser.parse_args()



fasta_file = options.s
wanted_file = options.w
result_file = options.o



#Parse the wanted.txt file for headers
########################################

wanted = set()
```

```
with open(wanted_file) as f:
for line in f:
line = line.strip()
if line !="":
wanted.add(line)

#Extract wanted seqs from .fa and write to new file
####################################################

fasta_sequences = SeqIO.parse(open(fasta_file), 'fasta')
with open(result_file, "w") as f:
for seq in fasta_sequences:
if seq.id in wanted:
SeqIO.write(seq, f, "fasta")
```

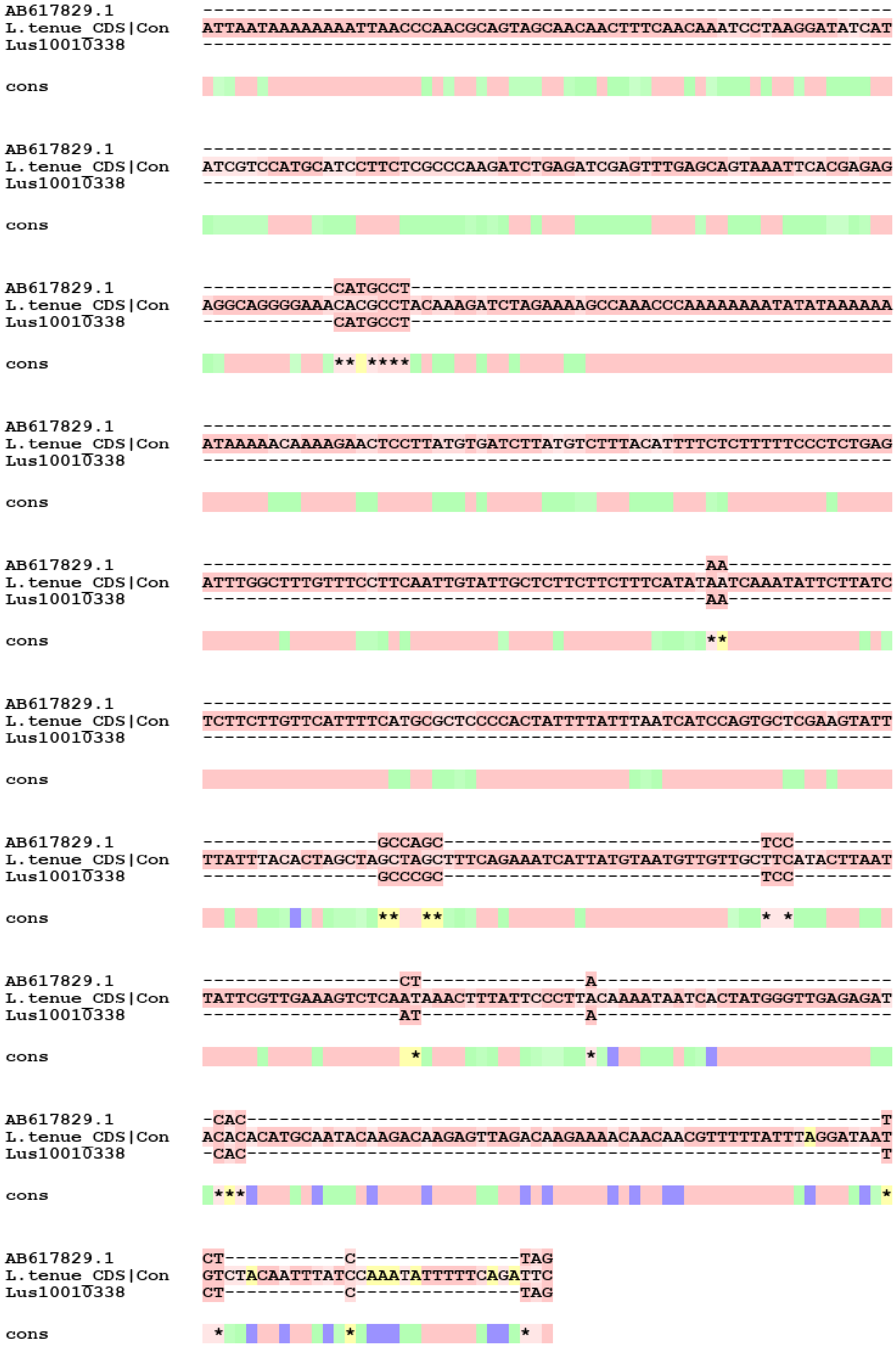
Appendix C

T-Coffee Example Output



```
T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)
Cedric Notredame
CPU TIME:0 sec.
SCORE=992
*
 BAD AVG GOOD
*
AB617829.1        :   99
L.tenue_CDS|Con   :   98
Lus10010338       :   99
cons              :   99

AB617829.1        ATGG----------CG-----------------------------------------------
L.tenue_CDS|Con   CCTGCTAATGAAAGCAAAACATGATACATATAAAGAGGGAATGGAATGTGGAGAGGCAAATCA
Lus10010338       ATGG----------CG-----------------------------------------------

cons                 *          *


AB617829.1        -----------------ACTTGTCC--CAC-----------------------GAG-------
L.tenue_CDS|Con   TTCCAAAAGTCATTGAAACTTGTTCATCACAATCTTATCAAACATCCAAAAAGGAGAGAAAAA
Lus10010338       -----------------ACTTGTCC--CAC-----------------------GAG-------

cons                               ****** *  ***                       ***


AB617829.1        ------TC----TG---------------------------------TTGTTGCT--------
L.tenue_CDS|Con   AAACCATCTTACTGTCCTTACCTTCCTTCCTTTTCTTTCTCTAATTATTGTTCCTTGATCACA
Lus10010338       ------TC----TG---------------------------------TTGTTGCT--------

cons                    **    **                                 ***** **


AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   ACCTTCCCTAAAAGAAAAGAGGAAAATCAGTGGTATATGCAGCCGTGTCAAAAGGTCAAACAT
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        ---------------------------------------------GA----------------
L.tenue_CDS|Con   TGCATAATTATATACAATAGATAAAAGAAAAATTATTATAGGAAAGAGCTATAAAAGATCCAT
Lus10010338       ---------------------------------------------TA----------------

cons                                                              *


AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   CCATCTAGGAAGGACGCATAGACTGGTCATTCAATGATTAGGAAAATGGCTCCTACATTGAGG
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        -------------GCTTGGTGTCGTTGATTTCCG---CA---------GTTAGCGGAGAAGAC
L.tenue_CDS|Con   CTAGTGTTGTTGAGCATTGTGTCCTTGATTTTTTGGCCAACTACGGTGGTTCAAGGGGAAGAC
Lus10010338       -------------CTCTAGCCTCGTTGATTTGCA---CA---------ATCCACGGTGAAGAT

cons                              * *  ** *******      **          *    ** *****


AB617829.1        CCGTACCTGTTCTTCACGTGGAAAGTGACTTACGGAACGCGGTCTCCATTGGGGAAGCCGGAG
L.tenue_CDS|Con   CCTTATATTTTCTTCACATGGAATGTGACCTATGGAACACTGGCCCCATTAGGCACTCCACAA
Lus10010338       CCATACCTCTTCTTCACGTGGAAAGTCACCTACGGTACACGGTCCCCACTGGGCTCGCCGGAG
```

```
AB617829.1          CCGTACCTGTTCTTCACGTGGAAAGTGACTTACGGAACGCGGTCTCCATTGGGGAAGCCGGAG
L.tenue_CDS|Con     CCTTATATTTTCTTCACATGGAATGTGACCTATGGAACACTGGCCCCATTAGGCACTCCACAA
Lus10010338         CCATACCTCTTCTTCACGTGGAAAGTCACCTACGGTACACGGTCCCCACTGGGCTCGCCGGAG

cons                ** **  * ******** ***** ** ** ** ** ** * * * *** * **    **  *


AB617829.1          CAGGTGATCCTCATCAATGACGAGTTCCCTGGCCCGGCGCTGAACACGACCACCAACAACAAC
L.tenue_CDS|Con     CAAGTAATATTGATTAACGGCGAGTTTCCAGGCCCCGTCATCAATTCCACATCCAACAACAAC
Lus10010338         CAGGTGATCCTCATCAACGACGAGTTCCCCGGCCCGGCCCTGAACACAACGACCAACAACAAC

cons                ** ** **  * ** ** * ****** ** ***** *   * **  * **  ***********


AB617829.1          GTGGTGGTCAATGTGTTCAACAACCTGGATGAACCGTTCCTCGTCACGTGGAGCGGCATCCAG
L.tenue_CDS|Con     GTGGTTGTGAATGTGTTCAACAACCTGGATGAGCCATTCCTCATCACTTGGAGTGGGGTCCAG
Lus10010338         GTCGTCATCAACGTTTTCAACAACCTCGACGAGCCGTTTCTCATCACGTGGAGCGGGATCCAG

cons                ** **  * ** ** *********** ** ** ** ** *** **** ***** **  *****


AB617829.1          CAGAGGAAGAACTCGTGGCAAGACGGCATGCCCGGTACCCAGTGCCCCATCCCGCCCGGCACC
L.tenue_CDS|Con     CAGAGGAAGAACTCGTGGCAAGATGGGGTGCTTGGGACCAACTGCCCAATCCCACCAGGCTCA
Lus10010338         CAGAGGAAGAACTCGTGGCAGGATGGCATGCCAGGTACTCAATGCCCCATTCCACCTGGCACC

cons                ******************** ** **  ***  ** **  * ***** ** ** ** *** *


AB617829.1          AACTACACCTACCATTTCCAGGTCAAGGACCAGATCGGCAGCTTCATGTACTTTCCTTCCACC
L.tenue_CDS|Con     AACTACACCTACCACTTCCAGGTGAAGGACCAGATCGGAAGCTTCATGTACTACCCGTCCACA
Lus10010338         AATTACACTTATCATTTCCAGGTCAAGGACCAGATCGGCAGCTTCTTGTACTACCCTTCCACC

cons                ** ***** ** ** ******** ************** ****** ******  ** *****


AB617829.1          GCAATGCACAAATCGGCGGGAGGGTTCGGTGGGATCCACATCAACAGCCGTCTCCTGATTCCT
L.tenue_CDS|Con     GCCATGCACAAGGCCGCAGGTGGATTCGGAGGCCTGCACATCAATAGCCGCCTCCTAATCCCA
Lus10010338         GCGATGCACAAGTCAGCAGGCGGCTTCGGTGGGATCCACATCAACAGCCGTCTACTGATTCCT

cons                ** ********  * ** ** ** ***** **  * ******** ***** ** ** ** **


AB617829.1          GTCCCTTATGCCGACCCTGAAGCTGACTACACTGTCATCATCAATGATTGGTTCAGCAAGACC
L.tenue_CDS|Con     GTCCCTTACCCAGATCCCGAGGATGACTACACGGTCATTATCAACGATTGGTACACCAAGACC
Lus10010338         GTCCCGTATGCTGATCCTGAAGCTGACTACACTGTTATCATCAATGATTGGTATAGCAAGACC

cons                ***** **  * ** ** ** * ********* ** ** ***** *******  * *******


AB617829.1          CACTCCGCCCTCCGAACCATGCTCGACAGCGGCCGAACCTTGGCTAGACCCGAAGGCGTCTTG
L.tenue_CDS|Con     CACTCGGCACTCCGCAACATGCTGGACAGCGGCCGCACCTTGGGCAGACCCGAAGGCGTCCTC
Lus10010338         CATTCCGCTCTCCGAACCATGCTCGATAGCGGCCGAACCTTGGCTAGACCGGAAGGGGTCTTG

cons                ** ** ** ***** * ****** ** ******** *******  ***** ***** *** *
```



```
AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   ATTAATAAAAAAAATTAACCCAACGCAGTAGCAACAACTTTCAACAAATCCTAAGGATATCAT
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   ATCGTCCATGCATCCTTCTCGCCCAAGATCTGAGATCGAGTTTGAGCAGTAAATTCACGAGAG
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        ------------CATGCCT--------------------------------------------
L.tenue_CDS|Con   AGGCAGGGGAAACACGCCTACAAAGATCTAGAAAAGCCAAACCCAAAAAAAATATATAAAAAA
Lus10010338       ------------CATGCCT--------------------------------------------

cons                          ** ****


AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   ATAAAAACAAAAGAACTCCTTATGTGATCTTATGTCTTTACATTTTCTCTTTTTCCCTCTGAG
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        ----------------------------------------------AA---------------
L.tenue_CDS|Con   ATTTGGCTTTGTTTCCTTCAATTGTATTGCTCTTCTTCTTTCATATAATCAAATATTCTTATC
Lus10010338       ----------------------------------------------AA---------------

cons                                                              **


AB617829.1        ---------------------------------------------------------------
L.tenue_CDS|Con   TCTTCTTGTTCATTTTCATGCGCTCCCCACTATTTTATTTAATCATCCAGTGCTCGAAGTATT
Lus10010338       ---------------------------------------------------------------

cons


AB617829.1        ----------------GCCAGC-----------------------------TCC---------
L.tenue_CDS|Con   TTATTTACACTAGCTAGCTAGCTTTCAGAAATCATTATGTAATGTTGTTGCTTCATACTTAAT
Lus10010338       ----------------GCCCGC-----------------------------TCC---------

cons                              **  **                             * *


AB617829.1        ------------------CT---------------A---------------------------
L.tenue_CDS|Con   TATTCGTTGAAAGTCTCAATAAACTTTATTCCCTTACAAAATAATCACTATGGGTTGAGAGAT
Lus10010338       ------------------AT---------------A---------------------------

cons                                 *               *


AB617829.1        -CAC----------------------------------------------------------T
L.tenue_CDS|Con   ACACACATGCAATACAAGACAAGAGTTAGACAAGAAAACAACAACGTTTTTATTTAGGATAAT
Lus10010338       -CAC----------------------------------------------------------T

cons               ***                                                          *


AB617829.1        CT-----------C---------------TAG
L.tenue_CDS|Con   GTCTACAATTTATCCAAATATTTTTCAGATTC
Lus10010338       CT-----------C---------------TAG

cons               *           *               *
```

Figure C.1: An alternative multiple alignment of LgMYB21 using T-COFFEE (Notredame et al., 2000). The output is colour-coded based on the quality of alignment. Closely aligned bases are coloured in red, whilst weakly aligned bases are coloured in green.

Appendix D

Post-Analysis Homology Searches

Table D.1: Top amino acid blast hits for GLO^T. The numbers after the contigs refer to the reading frame.

	% Identical Matches	Length	Mis-matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 107031 1	71.134	97	28	0	142	432	1	97	2.79e-45	153
Contig 107033 1	71.134	97	28	0	142	432	1	97	7.98e-43	151
Contig 71832	39.456	147	84	1	1	426	76	222	7.54e-36	128
Contig 71857	49.057	106	54	0	1	318	75	180	1.18e-33	124
Contig 71825 3	49.057	106	54	0	1	318	76	181	1.60e-32	123
Contig 51607	41.667	180	100	3	1	528	57	235	4.41e-32	121
Contig 51603 2	41.143	175	98	3	1	513	38	211	2.07e-31	118
Contig 107032 1	76.000	75	18	0	1	225	151	225	2.93e-31	122
Contig 104195	41.667	180	100	3	1	528	57	235	7.31e-31	120
Contig838661	36.275	204	115	4	1	579	209	408	2.07e-30	120
Contig 12531 1	39.247	186	100	3	1	522	237	421	2.36e-30	120
Contig 12539 1	39.247	186	100	3	1	522	220	404	5.78e-30	118
Contig 104196	42.857	168	91	3	1	492	57	223	1.51e-29	117
Contig 83869 2	33.484	221	115	5	1	579	197	413	1.81e-29	117
Contig 12528 1	41.765	170	92	2	1	492	194	362	1.99e-29	117
Contig 104194 2	45.205	146	75	3	1	426	38	182	8.00e-29	112
Contig 79212	38.587	184	95	5	1	513	29	207	9.88e-29	110
Contig 26899	37.778	180	96	3	1	507	149	323	1.64e-28	114
Contig 26898	37.778	180	96	3	1	507	148	322	1.73e-28	114
Contig 26896 2	35.955	178	98	3	7	510	1	172	2.49e-28	108
Contig 104202 2	45.833	144	73	3	1	420	38	180	3.59e-28	111
Contig 104201 2	46.429	140	70	3	1	408	38	176	4.64e-28	110
Contig 95123 1	38.312	154	77	3	1	432	33	178	6.55e-28	107

Table D.1

	% Identical Matches	Length	Mismatches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 107030 4	86.207	58	8	0	1	174	258	315	1.02e-27	109
Contig 51594 2	44.444	144	75	2	1	420	38	180	1.27e-27	112
Contig 79211 1	41.935	155	72	5	1	426	168	317	3.15e-27	110
Contig 85785	35.165	182	109	3	1	531	137	314	4.73e-27	108
Contig 85783 1	35.465	172	102	3	1	501	292	459	8.21e-27	110
Contig 73472 1	35.638	188	87	5	1	531	33	197	9.47e-27	105
Contig 51602 2	45.139	144	74	3	1	420	38	180	9.78e-27	109
Contig 51598 2	45.139	144	74	3	1	420	38	180	1.19e-26	108
Contig 13142	40.397	151	74	3	1	420	146	291	1.64e-26	106
Contig 125924 1	35.928	167	97	4	1	492	54	213	2.23e-26	104
Contig 38890	39.873	158	80	3	1	447	213	364	2.87e-26	107
Contig 38878	37.968	187	98	4	1	534	213	390	3.03e-26	107
Contig 38877	39.873	158	80	3	1	447	213	364	4.11e-26	107
Contig 26888 1	38.372	172	98	3	1	495	11	181	7.40e-26	107
Contig 103010 1	34.742	213	98	8	1	552	100	300	1.03e-25	106
Contig 103009 1	34.742	213	98	8	1	552	100	300	1.33e-25	106
Contig 85782 1	34.555	191	112	5	1	558	92	274	1.80e-25	103
Contig 85793 1	34.066	182	111	3	1	531	92	269	1.81e-25	104
Contig 85787	34.066	182	111	3	1	531	137	314	4.34e-25	104
Contig 85791	35.503	169	100	3	1	492	137	301	5.65e-25	104
Contig 85784 1	35.503	169	100	3	1	492	92	256	7.41e-25	103
Contig 73469 1	36.000	175	92	3	1	492	33	198	1.03e-24	103
Contig 85790	38.776	147	81	3	1	426	137	279	3.11e-24	100

Table D.2

	% Iden- tical Matches	Length	Mis- matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 26885	36.416	173	101	2	1	495	303	474	4.24e-24	102
Contig 125921 2	37.952	166	95	3	1	492	58	217	4.29e-24	99.0
Contig 85786	38.776	147	81	3	1	426	137	279	5.62e-24	100
Contig 85789	38.776	147	81	3	1	426	137	279	5.85e-24	100
Contig 124377 1	36.986	146	86	2	1	426	25	168	9.57e-24	98.2
Contig 12557 1	61.644	73	27	1	1	219	188	259	2.66e-23	98.2
Contig 12549	61.644	73	27	1	1	219	192	263	2.72e-23	98.2
Contig 103008 1	37.500	160	78	4	1	420	100	257	3.03e-23	99.8
Contig 34350	57.333	75	31	1	1	225	57	130	3.72e-23	94.4
Contig 124375 2	40.268	149	78	4	1	426	157	301	1.61e-22	96.3
Contig 79214 1	64.516	62	22	0	1	186	61	122	1.71e-22	91.3
Contig 83876	64.516	62	22	0	1	186	60	121	3.17e-22	90.5
Contig 79216 2	64.516	62	22	0	1	186	147	208	1.63e-21	90.9
Contig 79217 1	64.516	62	22	0	1	186	179	240	3.10e-21	90.9
Contig 83880	57.534	73	29	1	1	219	177	247	3.42e-21	90.9
Contig 95121 1	37.500	128	59	3	13	360	1	119	4.41e-21	87.4
Contig 85788 1	55.405	74	31	1	1	222	92	163	7.34e-21	89.7
Contig 101140 2	56.000	75	32	1	1	225	58	131	1.04e-20	92.8
Contig 49670 2	34.969	163	98	3	28	492	2	164	1.93e-20	87.4
Contig 101144 2	56.757	74	31	1	1	222	58	130	5.28e-20	89.0
Contig 103012 1	62.712	59	22	0	1	177	100	158	6.75e-20	90.5
Contig 107032 4	77.966	59	13	0	179	3	319	377	7.16e-20	90.1

	% Identical Matches	Length	Mismatches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 130032 2	61.017	59	23	0	1	177	103	161	5.42e-19	84.3
Contig 101142 1	52.113	71	33	1	1	213	54	123	8.65e-19	85.1
Contig 26891 2	43.396	106	54	2	7	324	1	100	1.50e-18	86.7
Contig 49661 2	33.333	165	100	1	28	492	2	166	4.46e-18	84.3
Contig 26894 2	43.396	106	54	2	7	324	1	100	6.70e-18	84.3
Contig 31242	59.016	61	25	0	1	183	172	232	6.26e-17	80.9
Contig 79218	70.213	47	14	0	1	141	55	101	9.79e-17	75.5
Contig 26900 4	60.714	56	22	0	10	177	1	56	1.10e-16	75.9
Contig 34355	71.429	49	14	0	1	147	81	129	1.65e-16	75.9
Contig 34356 2	71.429	49	14	0	1	147	99	147	2.27e-16	75.9
Contig 107030 2	73.684	57	15	0	173	3	1	57	4.36e-16	78.2
Contig 31246 2	54.386	57	26	0	1	171	53	109	2.02e-15	74.3
Contig 9865 1	60.377	53	21	0	1	159	82	134	3.52e-15	74.3
Contig 11683 1	54.386	57	26	0	1	171	74	130	4.37e-15	74.7
Contig 11679 2	54.386	57	26	0	1	171	90	146	7.70e-15	74.7
Contig 31244	54.386	57	26	0	1	171	91	147	1.48e-14	74.3
Contig 107031 3	45.161	93	51	0	420	142	152	244	1.63e-14	73.2
Contig 9867	60.377	53	21	0	1	159	179	231	3.26e-14	74.3
Contig 9863	60.377	53	21	0	1	159	179	231	3.29e-14	74.3
Contig 9869	60.377	53	21	0	1	159	179	231	3.37e-14	74.3
Contig 9868	60.377	53	21	0	1	159	179	231	3.54e-14	73.9
Contig 11678 1	54.386	57	26	0	1	171	74	130	3.63e-14	73.6
Contig 11685 1	54.386	57	26	0	1	171	74	130	3.69e-14	73.6
Contig 9862 3	60.377	53	21	0	1	159	179	231	3.86e-14	73.9
Contig 107033 3	45.161	93	51	0	420	142	308	400	5.01e-14	73.2

Table D.2

	% Iden- tical Matches	Length	Mis- matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 34347 1	73.333	45	12	0	1	135	161	205	2.05e-13	69.3
Contig 34134 1	54.717	53	24	0	1	159	65	117	3.54e-13	70.1
Contig 34130 1	42.667	75	42	1	1	225	66	139	3.75e-13	70.1
Contig 34135 2	54.717	53	24	0	1	159	102	154	3.99e-13	70.5
Contig 34131 1	54.717	53	24	0	1	159	130	182	4.16e-13	70.5
Contig 34136 1	54.717	53	24	0	1	159	66	118	4.35e-13	69.7
Contig 34128 1	54.717	53	24	0	1	159	66	118	5.41e-13	70.5
Contig 34138	54.717	53	24	0	1	159	122	174	1.08e-12	68.2
Contig 34132	54.717	53	24	0	1	159	122	174	1.60e-12	68.6
Contig 34137	54.717	53	24	0	1	159	122	174	1.64e-12	68.6
Contig 12549 4	58.333	72	29	1	218	3	73	143	6.01e-11	63.9
Contig 12531 5	56.944	72	30	1	218	3	265	335	1.77e-10	63.2
Contig 107032 2	33.613	119	41	2	190	432	261	379	1.98e-10	63.2
Contig 34351	30.556	144	94	3	118	537	1	142	2.29e-10	61.6
Contig 107037	33.613	119	41	2	190	432	17	135	2.84e-10	62.0
Contig 34350 3	60.345	58	23	0	176	3	64	121	3.02e-10	60.5
Contig 12539 4	56.944	72	30	1	218	3	185	255	3.04e-10	62.4
Contig 12557 4	58.333	72	29	1	218	3	65	135	3.19e-10	62.0
Contig 79219	29.054	148	89	3	106	513	1	144	4.81e-10	60.1
Contig 12528 5	56.944	72	30	1	218	3	306	376	7.26e-10	61.2
Contig 51603 3	62.069	58	22	0	176	3	176	233	8.74e-10	60.5
Contig 51602 3	62.069	58	22	0	176	3	401	458	1.33e-09	60.5

Table D.2

	% Iden- tical Matches	Length	Mis- matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 104194 4	62.069	58	22	0	176	3	213	270	1.37e-09	60.1
Contig 104201 4	62.069	58	22	0	176	3	234	291	1.41e-09	60.1
Contig 104202 4	62.069	58	22	0	176	3	243	300	1.47e-09	60.1
Contig 130031 1	45.902	61	32	1	4	186	204	263	1.51e-09	60.1
Contig 51598 4	62.069	58	22	0	176	3	389	446	1.62e-09	60.1
Contig 94463 1	45.902	61	32	1	4	186	204	263	1.92e-09	60.1
Contig 104195 5	60.345	58	23	0	176	3	369	426	1.96e-09	60.1
Contig 51594 4	62.069	58	22	0	176	3	491	548	1.98e-09	60.1
Contig 51601	30.147	136	90	2	118	513	1	135	2.31e-09	59.3
Contig 49666 2	54.000	50	23	0	28	177	2	51	2.48e-09	57.4
Contig 85786 3	66.667	51	17	0	158	6	207	257	2.66e-09	59.7
Contig 85790 5	66.667	51	17	0	158	6	162	212	2.76e-09	59.3
Contig 104196 5	60.345	58	23	0	176	3	383	440	3.09e-09	59.3
Contig 85788 4	49.412	85	38	2	248	6	47	130	3.19e-09	58.2
Contig 94461 1	45.902	61	32	1	4	186	204	263	3.48e-09	59.3
Contig 85789 3	66.667	51	17	0	158	6	202	252	3.52e-09	59.3
Contig 107031 2	50.685	73	36	0	143	361	1	73	3.86e-09	58.2
Contig 51607 5	58.621	58	24	0	176	3	222	279	3.94e-09	58.9
Contig 85785 3	66.667	51	17	0	158	6	167	217	4.01e-09	58.9
Contig 51600	30.147	136	90	2	118	513	1	135	4.21e-09	58.2
Contig 79215 2	71.429	35	10	0	1	105	187	221	7.27e-09	57.4
Contig 13142 4	54.545	66	26	1	200	3	91	152	8.52e-09	57.8
Contig 85783 3	64.706	51	18	0	158	6	245	295	9.46e-09	58.2

Table D.2

	% Iden-tical Matches	Length	Mis-matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 130032 5	50.000	64	32	0	194	3	33	96	9.90e-09	56.6
Contig 85791 4	64.706	51	18	0	158	6	312	362	1.13e-08	57.8
Contig 83880 5	60.784	51	20	0	155	3	20	70	1.65e-08	56.6
Contig 26900 1	61.224	49	19	0	155	9	62	110	1.71e-08	54.3
Contig 124377 4	55.769	52	23	0	158	3	184	235	1.79e-08	56.6
Contig 85787 3	64.706	51	18	0	158	6	268	318	1.80e-08	57.0
Contig 85784 5	64.706	51	18	0	158	6	314	364	1.93e-08	57.0
Contig 26898 4	54.545	66	26	1	200	3	333	394	1.97e-08	57.0
Contig 83869 5	60.784	51	20	0	155	3	463	513	2.08e-08	57.0
Contig 26899 4	54.545	66	26	1	200	3	347	408	2.14e-08	57.0
Contig 85793 5	64.706	51	18	0	158	6	242	292	2.32e-08	56.6
Contig 85782 4	64.706	51	18	0	158	6	168	218	2.74e-08	56.2
Contig 26888 4	60.784	51	20	0	155	3	434	484	5.24e-08	55.8
Contig 73468 4	71.875	32	9	0	1	96	110	141	5.43e-08	53.5
Contig 26885 5	60.784	51	20	0	155	3	459	509	5.87e-08	55.8
Contig 107032	48.276	58	30	0	3	176	152	209	6.64e-08	55.5
Contig 107040	54.762	42	19	0	307	432	46	87	7.64e-08	54.7
Contig 107032 3	47.458	59	31	0	177	1	320	378	1.30e-07	54.7
Contig 79218 3	57.447	47	20	0	143	3	1	47	1.35e-07	51.6
Contig 12537 2	69.444	36	11	0	1	108	294	329	1.83e-07	53.9
Contig 12536 3	69.444	36	11	0	1	108	296	331	2.61e-07	53.5
Contig 124375 4	53.704	54	25	0	167	6	144	197	2.73e-07	53.5
Contig 26894 4	58.000	50	21	0	155	6	397	446	4.27e-07	53.1
Contig 104198	32.673	101	63	2	118	408	1	100	4.62e-07	52.8

Table D.2

	% Iden- tical Matches	Length	Mis- matches	Gaps	Start alignment in query	End alignment in query	Start alignment in subject	End alignment in subject	E value	Bit Score
Contig 26891 4	58.000	50	21	0	155	6	546	595	5.56e-07	52.8
Contig 26896 3	58.000	50	21	0	155	6	137	186	5.56e-07	51.6
Contig 31242 3	55.769	52	23	0	158	3	91	142	6.06e-07	52.4
Contig 34354	32.673	101	63	2	118	408	1	100	7.06e-07	52.4
Contig 107031 4	61.290	93	36	0	419	141	152	244	8.53e-07	51.6
Contig 103009 4	56.863	51	22	0	155	3	325	375	1.31e-06	51.6
Contig 103010 3	56.863	51	22	0	155	3	355	405	1.34e-06	51.6
Contig 103012 3	56.863	51	22	0	155	3	385	435	1.35e-06	51.6
Contig 34356 3	55.102	49	22	0	149	3	1	49	1.40e-06	49.7
Contig 103008 3	56.863	51	22	0	155	3	391	441	1.40e-06	51.6
Contig 107033 4	61.290	93	36	0	419	141	308	400	1.58e-06	51.2
Contig 34347 3	60.000	45	18	0	137	3	1	45	2.31e-06	50.1
Contig 73472 4	54.902	51	23	0	155	3	139	189	2.94e-06	50.1
Contig 79211 4	42.623	61	35	0	183	1	204	264	3.17e-06	50.4
Contig 95123 5	54.902	51	23	0	155	3	122	172	3.85e-06	49.3
Contig 73469 4	54.902	51	23	0	155	3	376	426	4.63e-06	50.1
Contig 51607 4	44.068	59	33	0	177	1	222	280	5.85e-06	49.7
Contig 107030 3	49.020	51	26	0	21	173	265	315	6.74e-06	49.3
Contig 128898 4	33.898	59	39	0	4	180	61	119	7.35e-06	48.9
Contig 38884	29.197	137	89	3	127	522	1	134	9.53e-06	47.4
Contig 104199	32.673	101	63	2	118	408	1	100	4.83e-07	52.4
Contig 34355 3	55.102	49	22	0	149	3	1	49	5.16e-07	50.4
Contig 79212 3	58.824	51	21	0	155	3	163	213	1.28e-08	57.0
Contig 107033 2	50.685	73	36	0	143	361	1	73	1.51e-08	57.4

Appendix E

RNA Quantification Tables

E.1 Heterostylous Species

Table E.1: *L. tenue* thrum quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	Nanodrop			Qubit
	ng/μl	260/280	260/230	ng/μl
GS1	36.6	1.94	0.31	25.8
GS2	95.8	1.77	0.61	62.6
GS3	84.7	1.83	0.30	56.0
GS1	20.7	1.75	0.15	6.52
GS2	80.1	1.78	0.22	50.6
GS3	196.5	1.66	0.43	152
GS1	70.5	1.87	1.14	30.0
GS2	280.3	1.35	0.92	40.6
GS3	87.1	1.90	0.31	45.1

Table E.2: *L. tenue* pin quantification using Nanodrop and Qubit.Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	205.6	1.96	0.46	168
GS2	85.0	1.97	0.58	80.4
GS3	30.1	1.72	0.07	33.0
GS1	66.9	1.80	0.19	46.2
GS2	26.3	1.78	0.16	11.9
GS3	49.4	1.73	0.14	37.4
GS1	92.4	1.80	0.22	84.4
GS2	88.3	1.83	0.25	85.2
GS3	93.7	1.88	0.34	56.3

Table E.3: *L. narbonense* thrum quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	823.1	1.94	1.16	230
GS2	40.6	1.46	0.48	31.7
GS3	37.0	1.57	0.58	23.5
GS1	253.3	2.04	1.22	178
GS2	159.0	1.94	1.32	103
GS3	30.7	2.06	0.83	62.4
GS1	135.6	1.89	1.13	99.3
GS2	182.8	1.84	1.19	127
GS3	55.4	1.94	0.61	47.2

Table E.4: *L. narbonense* pin quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	268.2	1.87	1.63	188
GS2	262.7	1.78	0.87	146
GS3	25.2	1.79	0.45	30.2
GS1	250.5	1.99	1.96	200
GS2	184.8	1.76	1.12	154
GS3	235.4	1.85	0.70	211
GS1	207.1	2.02	1.55	138
GS2	187.3	1.84	0.89	97.4
GS3	272.9	1.87	0.97	253

Table E.5: *L. viscosum* thrum quantification using Nanodrop. RNA extraction and cDNA synthesis performed by Conor Hughes (3rd year project student). Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	46.0	1.83	0.28	32.5
GS2	114.9	1.82	0.39	100
GS3	35.0	1.76	1.61	15.9
GS1	27.0	1.32	0.78	45.4
GS2	52.8	1.33	0.47	36.0
GS3	93.2	1.70	0.16	85.4
GS1	54.6	1.25	1.68	20.7
GS2	60.0	1.44	0.13	53.8
GS3	25.3	1.31	0.52	30.4

Table E.6: *L. viscosum* pin quantification using Nanodrop and Qubit. RNA extraction and cDNA synthesis performed by Conor Hughes (3rd year project student). Values given to 3 significant figures.

Growth Stage	Nanodrop			Qubit
	ng/μl	260/280	260/230	ng/μl
GS1	97.2	1.53	0.14	107
GS2	23.6	1.31	0.52	18.9
GS3	32.4	1.78	0.27	27.4
GS1	27.9	1.74	0.43	32.0
GS2	18.9	1.50	0.17	10.0
GS3	19.3	1.59	0.09	22.1
GS1	16.2	1.78	0.34	33.9
GS2	15.0	1.78	0.34	7.45
GS3	47.8	1.57	0.45	62.3

E.2 Homostylous Species

Table E.7: *L.setaceum* quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	Nanodrop			Qubit
	ng/μl	260/280	260/230	ng/μl
GS1	160.1	1.90	0.33	50.4
GS2	95.7	1.90	0.26	47.8
GS3	42.5	2.21	0.23	42.8
GS1	149.3	1.71	0.38	27.4
GS2	224.3	1.84	0.61	121
GS3	87.5	1.91	0.46	27.2
GS1	210.6	1.83	0.64	106
GS2	225.5	1.97	0.65	201
GS3	310.6	2.10	1.03	151

Table E.8: *L.catharticum* quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	146.9	1.72	0.42	186
GS2	46.5	1.91	0.84	92.6
GS3	42.7	1.86	0.10	68.2
GS1	143.1	1.91	0.70	93.6
GS2	63.8	1.85	0.38	84.4
GS3	86.1	1.88	0.49	142
GS1	29.1	1.78	0.24	94.8
GS2	43.5	1.77	0.19	126
GS3	34.5	1.94	0.66	140

Table E.9: *L.tenuifolium* quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	**Nanodrop**			**Qubit**
	ng/μl	260/280	260/230	ng/μl
GS1	153.6	1.92	0.42	92.6
GS2	83.0	1.95	0.56	58.2
GS3	36.1	1.74	0.09	42.0
GS1	122.3	1.88	0.49	124
GS2	26.2	1.90	0.33	16.1
GS3	57.5	1.86	0.42	46.4
GS1	98.0	1.81	1.25	80.8
GS2	87.0	1.99	0.55	66.1
GS3	75.0	1.92	1.03	56.3

Table E.10: *L.maritimum* quantification using Nanodrop and Qubit. Values given to 3 significant figures.

Growth Stage	Nanodrop			Qubit
	ng/μl	260/280	260/230	ng/μl
GS1	78.2	1.92	0.42	92.6
GS2	20.2	1.95	0.07	58.2
GS3	130.0	1.74	0.34	42.0
GS1	300.9	1.88	0.14	124
GS2	60.6	1.90	0.20	48.0
GS3	290.1	1.86	0.57	164
GS1	192.0	1.81	0.38	80.8
GS2	33.2	1.99	0.26	20.6
GS3	81.6	1.92	1.03	85.0