INVESTIGATIONS INTO PLANT GENOME EVOLUTION USING MASSIVE PARALLEL SEQUENCING

by

Saemundur Sveinsson

M.Sc., The University of Iceland, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2014

Abstract

The advancements of various massively parallel sequencing (MPS) methods in the last five years have enabled researchers to tackle biological problems that until recently seemed intractable. One of the most widely used MPS methods comes from Illumina®, which combines short and accurate sequencing reads with high throughput. Data generated using Illumina sequencing is used in every chapter of this thesis to characterize patterns of genome evolution using phylogenetic approaches in various plant genera. The thesis is focused on three main aspects of plant genome evolution: transposable elements (chapter 2), polyploidy (chapters 3 and 4) and plastid genomes (chapters 5 and 6). In every chapter phylogenetic hypotheses are generated from sequences assembled from Illumina reads, which I use to frame my research questions. In chapter 2 I investigated intra- and interspecific patterns of transposable element (TE) abundance in *Theobroma cacao* and related species. I found that reference based mapping of short sequencing reads works well to characterize TEs within the same species but is not reliable for interspecific comparison. In chapter 3 I used Illumina sequenced transcriptomes of 11 flax species, to investigate the presence of paleopolyploidy event within the genus. I discovered a previously unknown paleopolyploidy event, occurring 23 – 42 million years ago. In chapter 4 I used low coverage Illumina whole genome sequencing to test a hypothesis regarding the allopolyploid origin of a North-American *Lathyrus* species, *L. venosus.* I conclude that *L. venosus* is not of hybrid origin, since no incongruencies were detected between nuclear and plastid phylogenies. In chapter 5 I pinpointed the evolutionary origin of highly repetitive plastid genomes that are known to exist within the clover genus (*Trifolium)*. I discovered that the repetitive plastomes are restricted to a single clade within *Trifolium,* which I estimated to be 12.4 – 13.8 million years old. In chapter 6 I investigated the pattern of gene rearrangements in the

IRLC clade of legumes. While plastomes are highly rearranged in this group, I characterized certain highly conserved gene blocks that have not been rearranged internally, and argue that these blocks may represent the fundamental gene regulatory organization of the plastid.

Preface

All chapters received input from coauthors.

A version of chapter 2 has previously been published as:

Sveinsson S, Gill N, Kane NC, Cronk Q. **2013**. Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (*Theobroma cacao L.)* and related species. *BMC Genomics* **14**: 502.

I was the lead investigator, carried out the data analysis and wrote the manuscript. Q. Cronk provided guidance with planning of the project and contributed significantly to writing and editing. N. Kane and N. Gill also helped with writing and editing.

A version of chapter 3 has been published as:

Sveinsson S, McDill J, Wong GKS, Li J, Li X, Deyholos MK, Cronk QCB. 2014. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics. *Annals of Botany* **113**: 753-761.

I was the lead investigator, carried out the data analysis and wrote the manuscript. Sequence data was obtained from the OneKP project (http://onekp.com). M. Deyholos assisted in generating Figure 4.1. G. Wong, J. Li and X. Li work for the OneKP project. Q. Cronk, M. Deyholos, J. McDill and G.Wong contributed to the writing and editing of the manuscript.

A version of chapter 5 has been submitted for publication. I was the lead investigator, carried out data analysis and wrote the manuscript. Q. Cronk assisted with writing and editing.

Chapters 4 and 6 were read and edited by Q. Cronk.

Table of Contents

Abstract ..... **ii**

Preface ..... **iv**

Table of Contents ..... **v**

List of Tables ..... **xii**

List of Figures ..... **xiv**

List of Abbreviations ..... **xxi**

Acknowledgements ..... **xxiii**

Dedication ..... **xxv**

Chapter 1: Introduction **1**

1.1 Theme of the thesis 1

1.2 Illumina® sequencing: the MPS method of choice in this thesis research 2

1.3 Phylogenetic analysis using data generated by MPS 3

1.4 Computational pipelines constructed for this thesis 4

1.4.1 T2Phy 4

1.4.2 Plast2Phy 6

1.5 Overview of the thesis 6

Chapter 2: Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (*Theobroma cacao* L.) and related species **9**

2.1 Introduction 9

2.2 Material and methods 11

2.2.1 Plant material and Illumina sequencing 11

2.2.2 Mapping of reads, coverage estimates and SNP calling 11
2.2.3 Identification of, and mapping to UCOS contigs 14
2.2.4 Phylogenetic analysis using the UCOS contigs 15
2.2.5 Graph based clustering of the Illumina reads 16
2.2.6 Statistical analysis 17
2.3 Results 18
2.3.1 Sequence coverage estimates and phylogenetic analysis using the UCOS contigs .. 18
2.3.2 Variation in TE abundance using short read mapping 19
2.3.3 Variation of TE copy number using *de novo* approaches 20
2.3.4 Intraspecific variation of TE abundance in *T. cacao* using short read mapping and PCA 21
2.3.5 Sequence conservation of transposable elements in *T. cacao* 21
2.4 Discussion 23
2.4.1 Different levels of nucleotide conservation in class I and class II TEs 23
2.4.2 Inter- and intraspecific differences in TE abundance in *H. balaensis*, *T. grandiflorum* and *T. cacao* 24
2.4.3 Mapping vs. *de novo* approaches to studying TEs from short reads 25
Chapter 3: Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics 35
3.1 Introduction 35
3.2 Material and methods 36
3.2.1 Illumina sequencing and *de novo* assembly 36
3.2.2 Identification of orthologues and phylogenetic analyses of *Linum* 37

3.2.3 Whole-genome duplication inference from age distributions of paralogues 40
3.2.4 Inference of a whole-genome duplication event from phylogenetic analysis of paralogues 41
3.3 Results 43
3.3.1 Illumina sequencing and *de novo* assembly 43
3.3.2 Phylogenetic analysis of *Linum* 43
3.3.3 Paralogue age distributions 44
3.3.4 Phylogenetic analysis of paralogues 45
3.4 Discussion 47
3.4.1 Consistency of date estimation 47
3.4.2 Relationship between the two polyploidy events in the evolution of cultivated flax (*Linum usitatissimum*) 47
3.4.3 Polyploidy and chromosome number 49
Chapter 4: Evidence for the origin of veiny pea (*Lathyrus venosus* Muhl. ex Willd., Fabaceae) by autopolyploidy 56
4.1 Introduction 56
4.2 Material and methods 59
4.2.1 Source of plant material and mapping data 59
4.2.2 Construction of Illumina sequencing libraries 60
4.2.3 Plastome assembly 60
4.2.4 Assembly of ribosomal subunits 61
4.2.5 Interspecific plastome and 45S rDNA sequence variation 62
4.2.6 Analysis of plastome rearrangements 63

4.2.7 Phylogenetic analysis 63
4.2.8 Examination of nucleotide additivity in the 45S rDNA sequence of *L. venosus* 65
4.2.9 Extraction of exomic nuclear data 66
4.2.10 Read depth analysis of the exomic region in whole genome shotgun sequencing data 67
4.2.11 Mining low coverage whole genome shotgun data of exomic SNPs 68
4.2.12 Estimation of genetic distances from exomic SNPs 68
4.2.13 Phylogenetic inference based on exomic SNPs 69
4.2.14 Dating the divergence of *L. ochroleucus* and *L. venosus* 70
4.3 Results 70
4.3.1 Illumina sequencing, de novo assembly and depth of coverage 70
4.3.2 Interspecific variation of plastomes and rDNA in *Lathyrus* 71
4.3.3 Plastome rearrangements within *Lathyrus* 72
4.3.4 Phylogenetic relationships among North-American *Lathyrus* 73
4.3.5 Lack of nucleotide additivity in *L. venosus* 75
4.3.6 Frequency distribution of read depths in the exomic region of *L. odoratus* 75
4.3.7 Exomic divergence between *L. venosus* and other species 76
4.3.8 Phylogenetic inference using exomic SNPs 76
4.3.9 The divergence time of *Lathyrus ochroleucus* and *L. venosus* 77
4.4 Discussion 77
4.4.1 Polyploid origin of *Lathyrus venosus* 77
4.4.2 Plastome rearrangements within *Lathyrus* 78

4.4.3 Phylogenetic potential of low depth whole genome sequencing for North-American *Lathyrus* 79

Chapter 5: Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*) 99

5.1 Introduction 99

5.2 Material and methods 100

5.2.1.1 Plant material and Illumina sequencing 100

5.2.2 Plastid genome assemblies and annotation 101

5.2.3 Identification and analysis of repeated DNA in the plastid genomes 102

5.2.4 Phylogenetic analysis 103

5.3 Results 104

5.3.1 Plastome assembly and structural variability 104

5.3.2 Phylogenetic distribution of the refractory species 105

5.3.3 The nature of the duplicated regions 106

5.3.4 Instances of gene loss 106

5.4 Discussion 107

5.4.1 The Phylogenetic distribution of plastome types 107

5.4.2 Potential causes of genome instability and functional significance of the repeat regions 108

Chapter 6: Delimitation of conserved gene clusters in the scrambled plastomes of the IRLC legumes (Fabaceae: Trifolieae and Fabeae) 115

6.1 Introduction 115

6.2 Material and methods 116

6.2.1 Source of plant material 116
6.2.2 Illumina sequencing 117
6.2.3 Plastid genome assemblies and annotation 117
6.2.4 Phylogenetic analysis 118
6.2.5 Identification of locally collinear blocks (LCBs) in plastid genomes and determination of gene clusters (GCs) 119
6.3 Results 120
6.3.1 Phylogenetic distribution of scrambled plastomes within the IRLC 120
6.3.2 Conserved gene clusters among the IRLC legume plastomes 121
6.4 Discussion 122
6.4.1 The rearrangements of plastomes in IRLC legumes 122
6.4.2 Do conserved blocks in otherwise rearranged plastomes represent operons? 123
Chapter 7: Conclusions 134
7.1 Summary 134
7.2 Conclusions 137
7.3 Future directions 138
Bibliography 140
Appendices 171
Appendix A - Supplementary material for chapter 2 171
A.1 Relative copy-number of transposable elements using reference based mapping to conserved regions of the class I LTR elements. Relative copy-numbers of the TE super-families in the three species represented with bar plots. Relative copy-number was calculated by dividing the total coverage of each super-family, within a sample, by the

sample's mean UCOS coverage. The mapping was preformed with relaxed settings in the short read aligner and the reads were mapped to conserved regions of class I LTR elements. 171

Appendix B - Supplementary material from chapter 4 172

B.1 Source of plant material and voucher information 172

B.2 Phylogenetic relationships among the *Lathyrus* species based on the small ribosomal subunit (5S). The tree was inferred using maximum likelihood and node support values calculated from 100 bootstrap replicates. 173

B.3 Phylogenetic relationships among the *Lathyrus* species based on a *BEAST species tree reconstruction of the large ribosomal subunit (45S) and protein coding regions from the plastome. Posterior probability scores are indicated on the nodes of the tree. 174

B.4 Densitree generated using SNAPP A DensiTree visualization of the SNAPP analysis. 175

Appendix C - T2Phy 176

C.1 T2Phy: from transcriptomes to phylogeny 176

List of Tables

Table 2.1 Sequence summary statistics. Illumina sequence summary statistics and observed average coverage of the UCOS contigs for *Theobroma cacao*, *T. grandiflorum* and *Herrania balaensis* based on Burrows-Wheeler Aligner (BWA) alignments. 33

Table 2.2 LTR retrotransposon frequencies in the three species estimated with two different methods. Comparison of estimated LTR retrotransposon frequencies as percentages of the genome, calculated with reference based mapping (upper half) and graph based clustering (lower half). Within *T. cacao* there is little discrepancy between the methods. Heterologous mapping between species produces different results suggesting that graph-based clustering may be more appropriate for inter-species comparisons (see Discussion). 34

Table 3.1 Number of reads acquired per species, in addition to tissue type info. 54

Table 3.2 Summary of the patterns observed in paralogue phylogenies. 55

Table 4.1 Native *Lathyrus* species in North America. The genus is very diverse in the west. *Lathyrus venosus* is one of a small number of eastern species. The taxonomic treatment in this table follows that of S. Broich (pers. comm.)..... 95

Table 4.2 Illumina sequencing and *de novo* assembly summary..... 96

Table 4.3 Composition and size of plastid gene blocks shown in Figure 4.3..... 97

Table 4.4 Coverage of reference transcriptome from whole genome sequencing at raw read depth of 0.144X..... 98

Table 5.1 Summary Illumina sequencing accessions and plastome assembly information. An asterisk marks plastome sequences newly reported in this thesis. Only a partial assembly of *T. pratense* was possible with our data (see text for explanation)..... 114

Table 6.1 Summary Illumina sequencing accessions, plastome assembly – and voucher information. An asterisk marks plastome sequences newly reported in this thesis. **RP: Roger Parsons Sweet Peas. ***DLP: Desert Legume Project. [1]See table B.1 in appendix B. 131

Table 6.2 Gene cluster identified in the locally collinear blocks from the MAUVE alignment. The table lists the genes in each gene cluster in addition to the boundary and length of the gene cluster, in the *Cicer arietinum* plastome. 133

List of Figures

Figure 2.1 Phylogeny of *Herrania balaensis*, *Theobroma grandiflorum* and nine of the *T. cacao* varieties. The phylogenetic tree was constructed using partial sequence data of 97 ultra conserved orthologus sequences (UCOS). *Theobroma cacao* cv. Scavina-6 was excluded from the phylogenetic analysis due to low sequencing coverage. Nodes marked with asterisk have high bootstrap support (>90%). 26

Figure 2.2 Relative copy-number of transposable elements using reference based mapping. Relative copy-numbers of the TE super-families in the three species represented with bar plots. Relative copy-number was calculated by dividing the total coverage of each super-family, within a sample, by the sample's mean UCOS coverage. The much lower recovery of transposable elements in the other species is apparently due to mapping failure as the graph based clustering indicates that TE copy numbers are comparable in all species. Error bars represent standard deviation and correspond to intraspecific variation. 27

Figure 2.3 Graph based clustering analysis of repetitive elements in the three species. Graph layouts of the four largest clusters of repetitive elements detected in the graph based clustering analysis. *Herrania balaensis* is shown on the left, *T. grandiflorum* in the middle and *T. cacao* cv. Criollo on the right. Clusters are ordered by size, with largest at the top and fourth largest at the bottom. Below each graph layout is the class of the repetitive element, the genome percentage of each cluster and number of paired reads belonging to it in parentheses. Coloured regions in the some graphs represent conserved domains identified by RepeatExplorer. A total of 11,243,224 reads were used in the graph based clustering. 29

Figure 2.4 PCA of the transposable element composition in the *Theobroma cacao* genotypes. A biplot from a principal component analysis (PCA) using the standardized abundance of each

TE super-family as explanatory variables. Percentage of the explained variance is shown in parentheses in the legend of the x- and y-axis. 30

Figure 2.5 Nucleotide variability of transposable elements in *Theobroma cacao*. Box plot showing the nucleotide diversity across the super-families in *T. cacao*. This shows that DNA transposons have less variation at the superfamily level (see Discussion). Analyses were performed on standardized data sets (Methods) and values are presented transformed to a log10 scale. 31

Figure 2.6 Nucleotide diversity of LTR/Copia and LTR/Gypsy elements in *Theobroma cacao*. (A) Schematic diagram of the structure of the two most common LTR retrotransposons super-families in the *T. cacao* genome. (B) Partitioning of nucleotide variation is shown as percentage values next to each of the retrotransposon components. The white arrows with black background represents the long terminal repeat (LTR), black line regions in between open reading frames (ORFs) and LTRs and grey boxes represent the following open reading frames: Reverse transcriptase (RT), integrase (IT), capsid protein (GAG), aspartic proteinase (AP) and Rnase H (RH). 32

Figure 3.1 STAR phylogeny of the 11 *Linum* species constructed from 413 gene trees. Branch lengths were estimated with GARLI and all nodes on the tree have 100 % bootstrap support. The tree shows the two major clades of Linum species studied here and their dominant flower colour. The tree was rooted with *Bischofia javanica* (Phyllanthaceae; taxon not shown here). 50

Figure 3.2 Cladogram, estimated with the STAR method, of the 11 *Linum* species (left), with their corresponding duplicate age distributions and SiZer plots (right). The SiZer plots are placed underneath each paralogue age distribution (the x-axis being Ks values). Different

bandwidths used in the Gaussian smoothing of the Ks values are plotted on the y-axis of the SiZer plots and an optimal binning of the Ks values is plotted on the x-axes. These plots are composed of four colours: blue represents a significant increase in Ks value density, red represents a significant decrease in density, purple represents regions where there is no significant increase or decease in density, and grey areas represent insufficient data. Peaks in duplicate age distributions generated by paleopolyploidy events are characterized by the SiZer plots as blue areas flanked by red and purple areas, generally located in the middle of the y-axis. The position on the x-axis depends on the age of the duplication event. The blue areas around Ks 0.68 in all the SiZer plots of the blue-flowered Linum species represent statistical evidence supporting the occurrence of a polyploidy event. 51

Figure 3.3 A phylogeny of orthologous groups that is consistent with a polyploidy event occurring on the branch leading to the blue-flowered *Linum* (black dot). The species relationship within each of the clades is consistent with the species phylogeny (Figure 3.1). The tree was rooted on its midpoint for visualization purposes and bootstrap support is shown near the nodes of the phylogeny. Y indicates the yellow-flowered clade, b indicates the blue-flowered 1 clades. Based on a BLASTX search on Phytozome (Goodstein et al., 2012) using the L. usitatissimum contigs, this gene appears to be a Co-chaperone-like protein in the GrpE family. L. usitatissimum-I corresponds to the gene Lus10029654 and L. usitatissimum-II matches Lus1002803 in the genome assembly of cultivated flax (*L. usitatissimum*). The best hit of both paralogues in the *Arabidopsis thaliana* genome assembly is AT5G17710. 52

Figure 4.1 Maps of the distribution of *Lathyrus ochroleucus* and *L. venosus*. (A) locations of representative herbarium specimens of *L. venosus*, representing the main distributional area of the species. *Lathyrus venosus* is a predominantly eastern species that does not occur west

of the rocky mountains. The dashed line marks the boundary of the mountain west (the eastern edge of the western cordilleras). (B) Map showing selected herbarium records of *L. ochroleucus*, indicating the main distributional area of the species. *L. ochroleucus* has a transcontinental-northern distribution, and in western Canada it occurs on both sides of the Rocky Mountains. 82

Figure 4.2 Photographs of some of the studied *Lathyrus* species: a) *Lathyrus palustris,* b) *L. venosus*, c) *L. ochroleucus* and d) *L. pubescens*..... 83

Figure 4.3 Circular representations of sequence variability among *Lathyrus* species within the large ribosomal subunit (A) and the plastome (B). Number of variable sites were binned into 100 bp (A) or 1,000 bp (B) blocks and variability visualized on a heatmap. The number of variable sites that each colour represents is shown in the small circular legend at the centre of each figure. (A) The ribosomal genes (18S, 5.8S and 26S) are in yellow, the intragenic – and external transcribed spacers (ITS and ETS) are in light green and the intergenic region is shown in black. (B) Protein coding genes are shown in dark green, tRNA are in yellow, rRNAs are in olive green and intergenic spacers are in black. 85

Figure 4.4 Linear representation of gene order in the *Lathyrus* and *Pisum sativum* plastomes. Each numbered box corresponds to a set of plastid genes in a particular order (see Table 4.3). Major plastotypes (MPt), which were defined based on gene order, are shown underneath each plastome, with the species that sharing each MPt shown in parentheses. The orientation of individual arrowed boxes indicates in what direction (5'-> 3' or 3'<-5') majority of the genes within each block are transcribed in *L. venosus*. Inversions of entire gene blocks between major plastotypes are represented by a horizontal flip of the coloured boxes. Structural similarity among *Lathyrus* MPt was examined by comparing Lathyrus-02 –

Lathyrus-05 to Lathyrus-01 (a-d). Possible inversions and/or translocations events of syntenic gene blocks are shown using coloured rectangles and arrows in the same color. A single arrow indicates a translocation event, where two crossing arrows indicate an inversion or a translocated inversion. 87

Figure 4.5 Phylogenetic trees representing the relationships among the sequenced *Lathyrus* species, based on the transcribed region of the large ribosomal subunit (A) and protein coding regions within the plastome (B). Trees were inferred using maximum likelihood and the support values, that are drawn on the nodes, are based on a 100 bootstrap searches. The long outgroup branch, splitting *P. sativum* and *Lathyrus,* was replaced with a wrinkle to save space. (A) Support values below 50 are omitted and (B) star indicates a 100% bootstrap support (*). 88

Figure 4.6 Pairwise genetic distances calculated from exomic SNPs. (A) Standard p-distance between *Lathyrus venosus* and all other species, error bars are +/- 2*standard deviations. (B) A Principal Coordinate Analysis (PCoA) of the pairwise modified p-distance (see methods). *Pisum sativum* and *L. sativus* were omitted from the analysis, in order to enhance the resolution among genetically similar taxa. 90

Figure 4.7 A DensiTree visualization of the three most frequently observed topologies of the North-American *Lathyrus,* produced by SNAPP (A), and individual topogies with their corresponding frequency in parenthesis (B-D). The color of trees B-D matches each topology in the densi-tree (A). 92

Figure 4.8. A network analysis of the North-American *Lathyrus* species and *L. davidii* using SplitsTree, based on the exomic SNPs. Two outgroup species, *L. pubescens* and *L. sativus*, were also included in the analysis but are not shown here. 93

Figure 5.1 Gene map of two plastid genomes and dotplots illustrating repeated regions. Gene maps (a and b) and dotplots (c and d) of *Trifolium boisseri* (a and c) and *Trifolium meduseum* (b and d). The gene maps shows the postions and translation direction of protein coding genes, tRNAs and rRNAs in the plastid genome. The dotplots illustrate repeated sequences within the plastomes, identified by BLASTN, where aligned sequences are shown as lines. Since the plastomes are aligned to themselves, the entire plastome is represented as a long and uninterupted diagonal line. Shorter lines and dots, which fall above and below the long line, are repeats. 110

Figure 5.2 Phylogenetic relationships among the *Trifolium* species, generated using maximum likelihood from a concatenated matrix of 58 protein coding plastid genes with a combined aligned length of 48,058 characters. The black dot marks the clade with plastomes refractory to assembly and the black vertical lines indicate the plastome size categories. The age of the most recent common ancestor (MRCA) of the refractory clade is estimated to be 12.4 – 13.8 million years old. All nodes have a 100% bootstrap support. 112

Figure 6.1 A phylogram showing relationships among the analyzed IRLC legume species. The phylogeny was generated using maximum likelihood from a concatenated matrix of 70 protein coding plastid genes with a combined aligned length of 62,525 characters. The black triangle marks the loss of the inverted repeat and the red triangle marks the evolutionary origin of highly rearranged plastomes (i.e. scrambled, see results). When bootstrap support was lower then 100%, it written next to the corresponding node. 125

Figure 6.2 A MAUVE alignment showing the boundaries of plastid identified locally collinear blocks (LCBs) and their rearrangements. The species in this figure are (from top to bottom): *Cicer arietinum* (top), *Medicago papillosa, Trifolium boissieri, Trifolium strictum, Trifolium*

glanduilferum, Lens culinaris, Vicia sativa and *Pisum sativum* (bottom). Each row represents a different plastid genome and the LCBs are color-coded. 127

Figure 6.3 A MAUVE alignment showing the boundaries of plastid identified locally collinear blocks (LCBs) and their rearrangements. The species in this figure are (from top to bottom): *Cicer arietinum* (top), *Lathyrus clymenum, Lathyrus tingitanus, Lathyrus sativus, Lathyrus odoratus, Lathyrus inconspicuus, Lathyrus pubescens, Lathyrus davidii, Lathyrus graminifolius, Lathyrus palustris, Lathyrus littoralis, Lathyrus japonicus, Lathyrus ochroleucus* and *Lathyrus venosus* (bottom). Note that the plastomes of *L. palustris* and *L. graminifolius* are collinear as are the plastomes of: *Lathyrus japonicus, L. littoralis, L. ochroleucus, L. venosus.* Each column represents a different plastid genome and the LCBs are color-coded. White areas within each LCB represent a drop in nucleotide sequence similarity..... 129

List of Abbreviations

AIC - Akaike information criterion

BWA - Burrows-Wheeler Aligner

GAII – genome analyzer II

IR – inverted repeat

IRLC – inverted repeat lost clade

Ks - synonymous substitution rates

LCB - locally collinear blocks

LTR – long terminal repeat

ML – maximum likelihood

MPS – massively parallel sequencing

MRCA – most recent common ancestor

NGS – next generation sequencing

NO – number of

ORF – open reading frame

PCA – principal component analysis

RT – reverse transcriptase

SNP – single nucleotide polymorphism

STAR - species tree estimation using average ranks of coalescences

TE – transposable element

TIRS – terminal inverted repeats

UBC – University of British Columbia

UCOS – ultra conserved orthologous elements

WGD – whole genome duplication

WGSS – whole genome shotgun sequencing

Acknowledgements

First of all, I am very grateful to my supervisor Dr. Quentin Cronk, for giving me the opportunity to come to UBC and work in his lab. His guidance has been extremely in valuable during my Ph.D. and has helped me to discover my strengths and weaknesses as a researcher. I am also grateful to Drs. Jeannette Whitton and Keith Adams, for their support and guidance throughout my graduate studies. I thank UBC for funding me for the first four years of my Ph.D. and I would also like to thank Natural Sciences and Engineering Research Council of Canada, the Education Centre of Southern Iceland and the Icelandic Club of British Columbia for funding.

Parts of this thesis relied on data obtained by other researchers, which I am very grateful for having been granted access to. The Illumina sequencing data in chapter 2 was obtained from the cacao sequencing project, and I would like to thank Drs. Hannes Dempewolf and Jan Engels who were instrumental in setting that project up. The project received funding from the World Bank through the Development Marketplace competition. The sequencing data in chapter 3 was obtained from the One Thousand Plants Project (1KP). I would like to acknowledge Dr. Douglas Soltis for making the transcriptome sequence of *Bischofia javanica* available for analysis. I would like to all the people who collected plant material for my thesis research: Curtis Bjork, Dr. Armando Geraldes, Dr. Diana Percy and Xinxin Xue. I am grateful to David Kaplan for greenhouse assistance and I would like to acknowledge Dr. Tiffany Fields United States Department of Agriculture (USDA) and Dr. Matthew B. Johnson at the Desert Legume Project at the University of Arizona, for facilitating access to germplasm resources.

I received extensive assistance in several aspects of data analysis. I would especially like to thank my lab mates, Drs. Armando Geraldes and Charles Hefer, for helping me in getting several computer programs to work and instructing me in the basics of modern biological data

analysis. I am also grateful to Andrew LeBlanc and Dr. Alistair Blachford, at UBC's Zoology Department computing unit, for their assistance. I would like to acknowledge the Western Canada Research Grid (Westgrid) for access to their high-performance computing resources, which were very useful in parts of the data analysis of this thesis. I am grateful to Dr. Daisie Huang for assistance with the program aTRAM and tips regarding some aspects of data analysis. I would like to thank Anastasia Kuzmin, at UBC's Biodiversity NextGen Sequencing Facility, for help with Illumina sequencing library preparations. I am also grateful to my former lab mate, Dr. Julia Nowak, for helping me with analysis of leaf anatomy. I would also like to thank my lab mate, Xinxin Xue, for her help in RNA extractions.

I am grateful to my parents, Sveinn Runólfsson and Oddný Sæmundsóttir, for their support and encouragements throughout my studies. Lastly I will be forever grateful to my wife, Vigdís Finnbogadóttir, for her support and love.

To Vigdís, Óli and Siggi

Chapter 1: Introduction

1.1 Theme of the thesis

The great leaps in the development of massively parallel sequencing (MPS) technologies have enabled researchers to tackle a host of biological questions, which only a few years ago seemed intractable (Nystedt et al., 2013; Zimin et al., 2014**)**. This results from a dramatic reduction of the cost per base in sequencing, which has revolutionized many fields of evolutionary biology (e.g. Lambert et al., 2013; Lemmon and Lemmon, 2013; Wagner et al., 2013). In the past five years, which is the time that I have been conducting my doctoral research, I have witnessed this revolution first-hand. For example, at the beginning of my doctoral research I wanted to determine the evolutionary origin of a North-American *Lathyrus* species (chapter 4). I needed a well-supported phylogeny of the group and I started out by Sanger sequencing a handful of plastid regions. However, I was not able to retrieve a robust phylogeny, mostly because of a lack of sequence variation. Only three years later, I was able to assemble complete plastid genomes for several *Lathyrus* species, from Illumina sequencing data, and retrieve a well-supported phylogeny (Figure 4.5). This transition from the use of individual gene sequences to genome scale data, in asking questions regarding the processes of plant genome evolution, was very influential in the scoping of my Ph.D. research. Furthermore, I found phylogenetics an especially powerful framework to pose my research questions. The overall theme of this thesis is the use of MPS data and phylogenetics to answer evolutionary questions in the field of plant genome evolution, focusing on transposable elements, polyploidy and plastid genome evolution.

1.2 Illumina® sequencing: the MPS method of choice in this thesis research

Illumina® sequencing was the MPS method of choice in this thesis research and is used in chapters 2 - 6 of this dissertation. Illumina sequencing was chosen for two main reasons. First I had access to whole genome shotgun sequencing - and transcriptome data, used in chapter 2 and 3 respectively, that were both acquired using the Illumina platform. Second, the Biodiversity Research Centre at UBC purchased an Illumina HiSeq 2000 system in 2011, which provided me with direct access to an in-house sequencing facility. Furthermore, data acquired on the Illumina platform had previously been used successfully by other researchers to tackle questions similar to those posed in chapters 4 (Bock et al., 2014) and 5 (Sabir et al., 2014).

Briefly, the sequencing chemistry of Illumina is based on multiple additions of fluorescently-labeled dNTPs that contain a reversible terminator bond. A fluorescence signal is emitted each time a dNTP is added to a template nucleotide string and recorded using digital imaging (see Metzker, 2010). The template nucleotide strings are generated by shearing the input DNA, using sonication, and size selecting fragments of desired lengths, usually around 500 bp. Two flavors of Illumina sequencing are used in this thesis, where the only difference is the type of input DNA material that was used to construct the sequencing libraries: (i) whole genome shotgun sequencing (WGSS), used in chapters 2, 4, 5 and 6, and (ii) transcriptome sequencing used in chapters 3 and 4. As the name suggests, the input material in WGSS is total genomic DNA present in plant cells. WGSS reads therefore contain information from the nuclear - and organelle genomes (Kane et al., 2012; Straub et al., 2012). When the average read depth of sequencing is low, this technique is known as genome skimming (Straub et al., 2012) or ultra-barcoding (Kane and Cronk, 2008; Kane et al., 2012). The input material for the transcriptome sequencing in chapters 3 and 4, was an RNA sample that had been converted to cDNA by

reverse transcription. In this thesis, polyA selection was used to enrich for nuclear messenger RNA (mRNA) in the sequencing libraries. The RNA samples used to construct the sequencing libraries originated either from a single extraction (chapter 4) or a pool of RNA from different tissue types of the same species (chapter 3).

1.3 Phylogenetic analysis using data generated by MPS

The increased availability and lowered costs of MPS have revolutionized many aspects of modern phylogenetics (Lemmon and Lemmon, 2013; Soltis et al., 2013) and phylogeography (McCormack et al., 2013). MPS has been especially useful for phylogenetic reconstruction for resolving very deep phylogenetic splits (e.g. Drew et al., 2014; Ruhfel et al., 2014), but also for more shallow relationships (Emerson et al., 2010; Wagner et al., 2013; Bock et al., 2014). In this thesis, Illumina sequencing data was used for phylogenetic analysis in chapters 2-6. Three different methodologies were employed for phylogenetic reconstruction: (i) Maximum Likelihood (Felsenstein, 1973) estimation from a concatenated matrix of protein coding plastid genes and ribosomal subunits using GARLI (Zwickl, 2006) (chapters 4, 5 and 6); (ii) species tree inference using the STAR method (species tree estimation using average ranks of coalescences) (Liu et al., 2009b) implemented in Phybase (Liu and Yu, 2010) (used in chapters 2 and 3); and (iii) species trees inference from exomic SNPs using SNAPP (Bryant et al., 2012) (chapter 4).

The STAR (species tree estimation using average ranks of coalescences) is a method of generating a species tree from a set of gene trees (Liu et al., 2009b) and is partly based on coalescence theory (see Liu et al., 2009a). A species tree topology is constructed using a distance matrix that is generated from ranked coalescence events among the input gene trees (see Liu et al., 2009a; Liu et al., 2009b). It has been used to infer species relationships in various groups of

organisms, such as mammals (Song et al., 2012), insects (Johnson et al., 2013) and pines (DeGiorgio et al., 2014). Since STAR only infers the topology of a species tree, I used GARLI (Zwickl, 2006) to estimate the branch lengths.

SNAPP (SNP and AFLP Package for Phylogenetic analysis) is a phylogenetic method that infers species trees from unlinked SNPs or AFLPs (Bryant et al., 2012). It uses Bayesian methods to implement a full coalescent model to generate a species tree directly from the molecular markers (Bryant et al., 2012). It has proven to be very useful in resolving phylogenetic relationships among recent radiations, such as in lizards (Lambert et al., 2013). It has also been used to study interspecific hybridization in birds (Rheindt et al., 2014) and plants (Hamlin and Arnold, 2014). The downside to this method is that it tends to be quite slow. I ran SNAPP on a data set of 12 taxa with about 23,000 SNPs and it took about three weeks to finish executing 10 million generations, using 8 CPUs on a UBC's computer cluster. Other researchers have reported similar experiences (Yoder et al., 2013). Despite its lengthy computational time, I found SNAPP very useful in my analysis of the evolutionary history of *Lathyrus venosus* (see chapter 4).

1.4 Computational pipelines constructed for this thesis

1.4.1 T2Phy

I originally developed the T2Phy (from Transcriptome to Phylogeny) pipeline to generate a phylogeny from the *de novo* assembled *Linum* transcriptomes investigated in chapter 3 (see Figure 3.1). The pipeline takes *de novo* assembled transcriptomes, or contigs of coding sequences from genome assemblies, as its input and outputs a phylogeny. The pipeline consists of two major stages. First it infers groups of orthologous sequences that it then uses to generate phylogenies. The first stage involves three steps (1) translating the mRNA (cDNA) sequences to

amino acid format using prot4EST (Wasmuth and Blaxter, 2004), (2) all vs. all BLASTP (Altschul et al., 1997) each of the translated sequences and finally (3) orthoMCL (Li et al., 2003) uses the all vs. all BLASTP output to infer orthologous groups. The second stage of the pipeline involves generating phylogenetic trees from the orthologous groups inferred in the first stage. Sequences are aligned using MAFFT (Katoh and Standley, 2013), alignment gaps trimmed using trimAl (Capella-Gutiérrez et al., 2009), appropriate models of nucleotide substitution inferred using jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) and phylogenetic trees for each ortholog groups generated using RAxML (Stamatakis, 2006). In addition to inferring gene trees for each ortholog group, T2Phy also outputs a concatenated matrix of all alignments into a NEXUS file. The NEXUS file is partitioned according to the orthologous group and contains information regarding the appropriate nucleotide substitution model of each partition. It also generates a species tree, from individual gene trees inferred by RAxML, using the STAR method (see section 1.3). More details regarding the pipeline are given in chapter 3, section 3.2.2. In addition to using T2Phy to generate a phylogeny for the *Linum* species in chapter 3, part 2 of the pipeline was used to estimate the relationships among cacao varieties in chapter 2 (Figure 2.1). As a further test of the method, the pipeline was used to construct a phylogeny of IRLC legumes for which transcriptomes are available at the NCBI's short read archive and from the OneKP project (http://onekp.com). The resultant phylogeny (see Figure 2 in appendix C.1) is identical to the most recent published phylogeny (see Figure 3 in LPGW, 2013). The release of T2Phy and a manuscript describing it in more detail are in preparation.

1.4.2 Plast2Phy

The plastomes of *Trifolium, Lens, Vicia, Pisum* and *Lathyrus* are highly rearranged (see chapters 4, 5 and 6), which makes it difficult to infer orthologous regions of non-coding sequences. Therefore I decided to restrict the phylogenetic inference, using plastid sequences, to protein coding genes. I put together a pipeline, written in Python, to automate this process that I named Plast2Phy (Plastid to Phylogeny). It reads in plastid gene sequences in either amino acid or nucleotide format, from both DOGMA (Wyman et al., 2004) and NCBI's Genbank fasta outputs. It works similarly to the second part of the T2Phy pipeline (see above). Plast2Phy starts by aligning individual genes with Mafft v. 7.0.5(-auto flag) (Katoh and Standley, 2013), then it trims gaps in the alignments using trimAl v.1.2 (-automated1 flag) (Capella-Gutiérrez et al., 2009) and finally it generates a concatenated alignment of all genes in a NEXUS format. It also includes options to generate individual plastid gene trees using RAxML (Stamatakis, 2006) and determine the appropriate models of nucleotide substitution using jModelTest (Guindon and Gascuel, 2003; Darriba et al., 2012). Plast2Phy was used to generate phylogenies in chapters 4, 5 and 6 and is available for download at at https://github.com/saemi/plast2phy.

1.5 Overview of the thesis

In chapter 2 I used Illumina WGSS sequencing to investigate transposable element (TE) variation within multiple individuals of cacao (*Theobroma cacao*) and two related species. Transposable elements (TEs) and other repetitive elements are a large and dynamically evolving part of eukaryotic genomes, especially in plants where they can account for a significant proportion of genome size. Their dynamic nature gives them the potential for use in identifying and characterizing crop germplasm. However, their repetitive nature makes them challenging to

study using conventional methods of molecular biology. I used the Illumina reads to analyze TEs using both an alignment/mapping approach and a *de novo* (graph based clustering) (see Novak et al. 2013) approach. I used a standard set of ultra-conserved orthologous sequences (UCOS) to standardized TE data between samples and provide a phylogenetic information on the relatedness among samples.

In chapter 3 I investigated the presence of a paleopolyploidy event within the flax genus (*Linum*). Cultivated flax (*Linum usitatissimum*) is known to have undergone a whole-genome duplication (WGD) around 5–9 million years ago. However it was unclear whether the genus had been shaped by any older *Linum* specific polyploidy events. I analyzed transcriptomes of 11 *Linum* species that were sequenced using the Illumina platform. The short reads were assembled *de novo* and the DupPipe pipeline was used to look for signatures of polyploidy events from the age distribution of paralogues. In addition, I assembled phylogenies of all paralogues within an estimated age window of interest.

In chapter 4 I investigated the evolutionary origin of *Lathyrus venosus*, which is a widespread North-American species. *Lathyrus venosus* is a tetraploid species, where most *Lathyrus* species are diploid. The origin of *L. venosus* is still uncertain, and both auto- and allopolyploid origins have been suggested including a proposed specific allopolyploid parentage of *L. palustris* × *L. ochroleucus* (Gutiérrez et al., 1994). However *Lathyrus* species rarely hybridize and autopolyploid cytotypes are known from some species (Broich, 1989; Khawaja et al., 1995; Khawaja et al., 1997). I sequenced eight individuals from six closely related species of North-American *Lathyrus* species and four outgroup species, each using a partial lane of the Illumina HiSeq-2000 platform (c. 0.1-0.2 Gb of quality sequence for each). The whole plastomes and 45S ribosomal units were reconstructed for each species *de novo* and used in phylogenetic

analysis. This sequencing strategy resulted in a low overall coverage per nuclear genome (0.1x – 0.5x) but this was sufficient read depth to assemble the plastome and the 45S rDNA repeat unit, and to call a large number of nuclear single copy exomic SNPs.

The aim of chapter 5 was to determine the evolutionary origin of unusual plastome structure that is known to occur among some clover species, such as *Trifolium subterraneum* (Cai et al., 2008) and *Trifolium repens* (Sabir et al., 2014). These plastomes are highly unusual, as they have been enlarged with many duplications, some gene losses and the presence of DNA unique to *Trifolium*. In order to investigate whether this is true of all *Trifolium* species or whether it is specific to *T. subterraneum* and *T. repens*, I sequenced and assembled the plastomes of eight additional *Trifolium* species widely sampled from across the genus.

In chapter 6 I investigated the highly rearranged plastomes of *Lens, Vicia, Pisum* and *Lathyrus*, a monophyletic group known comprising the Fabeae and *Trifolium* species. The gene order among these plastomes is known to be very variable (Palmer and Thompson, 1982; Cai et al., 2008; Magee et al., 2010; Sabir et al., 2014), which is due to unusual high frequencies of translocations and/or inversions in their plastid genomes. Most plastid genomes of higher plants are however highly conserved in gene order (Wicke et al., 2011). Plastids are of bacterial origin (Margulis, 1970) and the gene space of their genomes is known to have operon-like features (Sugita and Sugiura, 1996), where some genes are transcribed in co-regulated blocks (Stern et al., 2010). The aim of this study is to analyze these rearrangements in these genera within IRLC, in order to investigate whether they can be used to study the organization of plastid genomes into operons.

Chapter 2: Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (*Theobroma cacao* L.) and related species

2.1 Introduction

Transposable elements (TEs) are a large and dynamically evolving part of plant genomes (Kumar and Bennetzen, 1993; Feschotte and Pritham, 2007). They occupy between 15% - 84% of plant genomes (Kelly and Leitch, 2011) and TE expansion is known to cause a significant increase in genome size in many cases (Sun et al., 2012). Transposable elements are a major force in plant evolution, not only by causing genome expansions but also by altering gene function either through disruption (Martin et al., 2005) or acting as a raw material for new genes and novel functions (Craig et al., 2002; Zhou et al., 2004).

Transposable elements are usually classified into two major classes based on their transposition mechanisms. Class I retrotransposons move about in a 'copy-and-paste' fashion, through a RNA intermediate, which is encoded back into DNA by an endogenous Reverse Transcriptase (RT) enzyme (Boeke and Corces, 1989). The two largest super-families of retrotransposons in plants, the LTR/Copia and LTR/Gypsy, have several other open reading frames, which play a role in the transposition, located between two regions of long terminal repeats (LTR) (Kumar and Bennetzen, 1993). Class II DNA elements move about in genomes through a DNA intermediate. The most extensively studied group of class II elements transpose by a 'cut-and-paste' mechanism and are classified into several super-families based on sequence similarity (Wicker et al., 2007). Cut-and-paste DNA transposons are characterized by a transposase gene and a pair of flanking terminal inverted repeats (TIRS) (Craig et al., 2002).

Transposable elements are known to vary extensively in copy-number and nucleotide sequence among closely related species (Novick et al., 2011; Sun et al., 2012) and even within the same species (Vicient et al., 1999). Plant LTR retrotransposons are well known to have intraspecific variation in copy-number (Pearce et al., 2000; Huang et al., 2008). This, in combination with the easily amplifiable LTR domain, has been used in the development of molecular markers for several crop species (Kumar and Hirochika, 2001; Syed et al., 2005; Schulman et al., 2012). In addition to the extensive presence/absence variability of the LTR elements, sequence heterogeneity is also known to be quite extensive (Flavell et al., 1992). The reverse transcriptase domain is the most extensively studied retrotransposon gene and it is known to show levels of heterogeneity from about 5% to 75% at the amino acid level (Flavell et al., 1992). Heterogeneity and sequence evolution of class II DNA transposons is relatively less studied, but a recent study shows that they can be quite heterogeneous (Novick et al., 2011).

Cacao (*Theobroma cacao* L.) is an economically important tree in the mallow family (Malvaceae) (Wood and Lass, 2001). It is widely grown in tropical regions as the source of cocoa beans for the manufacture of chocolate (Wood and Lass, 2001). Cacao has long been known to be genetically diverse (Motamayor et al., 2008) and traditionally three major lineages of Cacao varieties have been recognized: Trinitario, Criollo, and Forastero (Cheesman, 1944). Recent work based on a variety of markers, including microsatellites and whole chloroplast genome sequences of several cacao varieties, has confirmed that the Criollo and Forastero groups are two distinct genetic lineages while the Trinitario group is of hybrid origin (Motamayor et al., 2003; Kane et al., 2012). Cacao has a relatively small genome, estimated to be around 430 Mb and it has a published genome assembly of about 75% of its estimated genome size (Argout et al., 2010). This small genome size can be partly explained by the relatively low abundance of

transposable elements, compared to other angiosperms. TEs comprise only approximately a quarter of the cacao genome (Argout et al., 2010).

In this study we use low-coverage Illumina whole genome shotgun sequencing to investigate the evolutionary dynamics and comparative analysis of 3,500 TE families in nine *T. cacao* varieties and two related species, *Theobroma grandiflorum* and *Herrania balaensis.*

2.2 Material and methods

2.2.1 Plant material and Illumina sequencing

Total genomic DNA was extracted from leaf tissue from 11 individuals belonging to three species in the Malvaceae: one *Herrania balaensis*, one *Theobroma grandiflorum* and nine *T. cacao.* Each *T. cacao* individual represented a different cultivated variety (see Table 2.1). DNA extraction was performed using the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) according to the manufacturer's protocol. Sequencing libraries were constructed using standard protocols and chemistry for the Illumina platform. Each library was sequenced on a single lane and generated either 60- or 80-bp paired-end sequences (see Table 2.1) on the Illumina GAII platform by Cofactor Genomics of St. Louis, MO (http://www.cofactorgenomics.com/). The raw reads are available on NCBI's Short Read Archive [SRA048198].

2.2.2 Mapping of reads, coverage estimates and SNP calling

The reference sequences of the transposable element (TE) families used in this study were extracted and characterized by the authors of the publication describing the *T. cacao* genome (Argout et al., 2010), who graciously made their data available for this study. Briefly they identified class I retro-transposons using LTR_finder (Xu and Wang, 2007), LTRharvest

(Ellinghaus et al., 2008) and in-house software that looked for signatures of class I retroelements, such as the long terminal repeat (LTR) and reverse transcriptase (RT). Class II elements were discovered using a BLASTX search of the transposase gene against the Repbase database proteins (Jurka et al, 2005). In all they identified 650 class I - and 2860 class II families. For more details see the supplementary methods in (Argout et al., 2010).

In order to estimate copy-number and sequence evolution of the TE families using our sequenced libraries of three species and nine *T. cacao* varieties, we mapped reads from each sequenced library to the TE reference contigs. First, the reads were trimmed for quality, with bases below quality of 20 trimmed from the ends of each read. Quality trimmed reads were treated as single-end sequences and mapped to the TE reference contigs using BWA v0.6.1 (Li and Durbin, 2009) with the program's default settings. The rationale behind treating the paired-end sequences as single-end was that TE copy-number estimation from coverage of the latter was believed to be more accurate, as paired-end information often links the repeat to different single-copy portions of the genome, preventing pairs from mapping near the boundaries of the repeated segment. Coverage estimates for each nucleotide position in the reference contigs were extracted from the sorted BAM file output of BWA using the genomeCoverageBed tool in the bedTools package v2.15.0 (genomeCoverageBed flags: -d -ibam) (Quinlan and Hall, 2010). Relative copy-number of each TE family was estimated by counting the number of reads covering each position of the reference contig and dividing by the length of the contig. Proportional abundance was calculated for each species, by dividing the abundance of each TE super-family by the abundance of all TEs. Information on nucleotide variants detected in the reads, compared to the TE reference contigs, was extracted using samtools v0.1.7a (Li et al., 2009). Nucleotide diversity was estimated for each TE reference contig by counting the number of variable sites, with read-

depth higher than 6 and base qualities higher than 20 (column 6 from samtools pileup –vcf output), and dividing by the length of the contig. To control for the effect of different read depths between different libraries, subsampling was used to ensure equality of total reads. Due to the repetitive nature of TEs, a variable site could represent a single nucleotide polymorphisms in a homologous copy, i.e. a heterozygote, or could stem from sequence divergence between different copies of a transposable element.

To account for differences in sequencing depth and read length of different libraries, reduced equalized data sets were used for some of the analyses presented here (Figures 2.4 and 2.5). The reduced data sets were generated by trimming the read length of all libraries to 60-bp and randomly extracting reads from all but the smallest sequenced library (Scavina-6). The purpose of this step was to make sure that variable read lengths and sequencing depths were not the cause of observed differences in TE coverage and nucleotide diversity. However any observed differences in UCOS coverage could be due to differences in genome sizes among the three species and the *T. cacao* varieties. Furthermore 49 sampling replicates were generated in order to test the effect of data sub-sampling on TE coverage estimates.

The class I LTR retrotransposons reference contigs were annotated using LTRHarvest (Ellinghaus et al., 2008) and LTRDigest (Steinbiss et al., 2009). These programs use similarity searches of conserved regions of LTR elements, such as the long terminal repeat and protein coding genes, to estimate the coordinates of the various features of the elements. That information was then used to estimate the variability of each LTR element feature by combining the feature file output of LTRDigest with the nucleotide variant output of samtools.

In order to test whether we could better account for sequence divergence in the class I TE among the three species, we tried mapping the reads exclusively to conserved regions of the LTR

retrotransposons and with relaxed BWA alignment stringency. Protein coding regions and the LTR were extracted from the reference contigs, based on the annotations from LTRHarvest (Ellinghaus et al., 2008) and LTRDigest (Steinbiss et al., 2009) and the BWA alignment step was performed with more relaxed settings (bwa flags: -l 1024 -i 0 -o 3). These settings allowed for more gaps and BWA used longer seed length for its short read alignments. These relaxed settings and the conserved regions of the LTR elements were only used to generate (data not shown)

2.2.3 Identification of, and mapping to UCOS contigs

A set of 357 Ultra Conserved Orthologous sequences (UCOS, http://compgenomics.ucdavis.edu/compositae_reference.php) was used to estimate the sequencing coverage of individual libraries as well as to estimate phylogenetic relationships between the three species and among the nine *T. cacao* varieties. These sequences represent single copy genes in *Arabidopsis thaliana* and tend to be conserved as single copy genes across Eukaryotes (Kozik et al., 2008). Since these genes are highly conserved and often present in a single copy in the genome, they are useful for estimating the sequencing coverage of each library and estimating copy number of the TE families. The 357 putative UCOS homologs in *T. cacao* were identified using BLASTX with an e-value threshold of 1E-34 (Altschul et al, 1997). The single copy status of the UCOS was verified by removing all contigs that had multiple hits to the *T. cacao* genome with an E-value lower than 1E-06. This left 245 UCOS to which the reads were mapped to using BWA, coverage of each UCOS contig was estimated using bedTools and single nucleotide polymorphisms (SNPs) called using samtools. Finally an average coverage was calculated for each library, by calculating the mean coverage of the 245 UCOS contigs.

Coverage of each TE reference contig was divided by the mean UCOS coverage, in order to estimate a relative copy-number of TEs to single copy nuclear genes.

2.2.4 Phylogenetic analysis using the UCOS contigs

A phylogenetic matrix was constructed by using the 245 UCOS contigs identified above as reference for short read mapping and by calling SNPs using previously described methods. *Theobroma cacao* cv. Scavina-6 was excluded from the phylogenetic analysis due to low sequencing coverage. For the construction of the matrix, only positions that were covered by 6 or more high quality reads in a given sample, with base quality equal or larger to 20 (column 6 in samtools pileup -vcf output) were used. Positions containing any ambiguous nucleotides, i.e. heterozygotes, were converted to Ns as were all other positions that did not meet previously mentioned criteria. Finally Ns were converted to gaps, trimAl v.1.2 (Capella-Gutierrez et al., 2009) used to remove all gaps and to convert the alignments to nexus format (trimAl flags: -nogap -nexus). All alignments shorter than 50 nucleotides were excluded from further analysis, leaving 97 UCOS for further analysis. A matrix with positional information of each of the UCOS contigs was constructed using phyutility v.2.2.4 (Smith and Dunn, 2008) (phyutility flags: -concat), for a combined analysis that includes separate analyses of each contig using a coalescence-based program (see below). Gene trees of individual UCOS alignments longer than 50 nucleotides were estimated with RAxML v7.2.6 (Stamatakis, 2006), using 10 independent runs and the GTRGAMMA sequence substitution model (raxml flags: -m GTRGAMMA -N 10). In order to estimate a single phylogeny of the three species and nine remaining *T. cacao* varieties (Scavina-6 excluded), a STAR (species trees based on average ranks of coalescences) phylogeny (Liu et al., 2009b) was constructed using the phybase R package (v.1.3) (Liu and Yu, 2010).

STAR uses the mean ranks of coalescent occurrences in a set of gene trees to construct a species tree topology (Liu et al., 2009b). In order to estimate branch lengths on the STAR tree, model parameters of the entire matrix were estimated using jModelTest v2.0.2 (Guindon and Gascuel, 2003; Darriba et al., 2012) and GARLI v2.0 (Zwickl, 2006) used to optimize model parameters and to add branch lengths to the STAR tree. Support values for the STAR phylogeny were estimated using a multi locus bootstrap (Seo, 2008) method implemented in the phybase package (Liu and Yu, 2010). One thousand multi locus bootstrap replicates were analyzed using Phyml v3.0 (Guindon et al., 2010), STAR trees estimated for each set of bootstrap replicates and a consensus tree constructed from all the STAR trees using the consense program in the PHYLIP package v3.69 (Felsenstein, 1993).

2.2.5 Graph based clustering of the Illumina reads

The repetitive elements of three species studied here were also investigated in a de novo fashion using RepeatExplorer (Novák et al., 2013), which is a graph based clustering method of characterizing repetitive elements described in (Novák et al., 2010), with the program's default settings. The Criollo-22 individual was chosen as the *T. cacao* representative. Briefly, RepeatExplorer uses information from sequence similarity among the reads and their partial overlap to construct graphs. Graphs are constructed using a Louvain method (Blondel et al., 2008), where sequence reads are represented by vertices, edges are connected with overlapping reads and edge weights correspond to the similarity score among reads. The graph layouts are then examined in order to find separate clusters of reads that are often connected and correspond to distinct families of genomic repeats. These clusters are analyzed in regards to their size, determined by the number of reads comprising each cluster as well as their graph topology which

gives information about their structure and variability. RepeatExplorer also performs a sequence similarity search of each cluster against RepBase (Jurka et al., 2005) in order to identify the type of the repetitive elements present in the cluster. If the predicted number of nodes exceeds the capacity of the available RAM, RepeatExplorer randomly subsamples the reads. RepeatExplorer outputs a comma separated value (csv) file, containing relevant information of the clusters it identified and that consist of 0.01% or more of the reads used in the analysis (default cut-off). The program calculates the genome percentage, which is the number of reads in each cluster divided by the all the reads used in the graph based clustering (11,243,224 reads in total). An in house Python script (available on request) was used parse the csv output file and combine parts of it with the figures of graph layouts. The output of that script is a panel of graph layouts, with each cluster's most abundant class of element, in addition to the genome percentage and number of paired-end reads belonging to the cluster.

2.2.6 Statistical analysis

Similarity of TE composition among sequenced individuals was investigated with a principal component analysis (PCA) using coverage of each TE super-family in the genomes of *H. balaensis*, *T. grandiflorum* and the nine *T. cacao* varieties. The PCA was performed using the prcomp function in R v2.14.1 (R Core Team, 2014), using the abundance of each super-family as explanatory variables with a natural logarithmic transformation and scale = TRUE. An in-house R script was used to run the PCA analysis on all sub-sampled data sets. The reduced data set ensured that differences in sequencing depth and read length did not affect the results.

2.3 Results

2.3.1 Sequence coverage estimates and phylogenetic analysis using the UCOS contigs

The Illumina sequencing yielded 1.7 – 5.9 Gbp of high quality sequence per sample (Table 2.1). Average coverage of the ultra-conserved orthologous sequences (UCOS), estimated with BWA mapping, varied between 1.8 and 9.4X per sample (see Table 2.1). This value represents a relative measure of per single copy locus sequencing depth for our libraries. However it is important to note that this method may slightly underestimate the sequencing coverage of *H. balaensis* and *T. grandiflorum* due to sequence divergence in the UCOS among the three species. Furthermore the results using UCOS are consistent with results using flow cytometry genome size estimates (see Table 1 in Argout et al., 2010). The UCOS data was used to standardize the TE data between samples to provide relative TE abundance data. It was also used to estimate the relatedness of the accessions in order to provide an evolutionary framework for TE variation.

The UCOS contigs were informative for the phylogenetic analysis of *H. balaensis*, *T. grandiflorum* and nine of the *T. cacao* varieties (Figure 2.1). Scavina-6 was excluded from the phylogenetic analysis due to low sequencing depth. The matrix used to construct the phylogeny consisted of 97 UCOS contigs with combined length of 20,438 nucleotides. Individual UCOS alignments varied in length, with the shortest alignment being 54 nucleotides and the longest 1,473 nucleotides. *Herrania balaensis* was set as the outgroup in the analysis which resulted in two main *Theobroma* clades (Figure 2.1). The first clade consists of only *T. grandiflorum*, with a 100% bootstrap support, and the second consisting of all *T. cacao* individuals. This is as expected given that *T. grandiflorum* and *T. cacao* are biologically distinct species. Within *T. cacao* there are two well-supported clades with *T. cacao* cv. Stahel and Amelonado grouping together and another well-supported grouping of *T. cacao* cv. Pentagonum, ICS06, ICS39,

Criollo-22 and B97 (Figure 2.1). B97 is the variety used in the whole genome sequencing project of *T. cacao* (Argout et al., 2010). EET-64 and ICS01 are unresolved on a polytomy.

2.3.2 Variation in TE abundance using short read mapping

The coverage of the single copy UCOS genes was used as a baseline for standardization of TE coverage. Using this method, copy-numbers of TE superfamilies relative to the UCOS coverage in the three species were calculated (Figure 2.2). The intraspecific variation of copy-number in *Theobroma cacao* is represented by the error bars. The LTR super-families are the most numerous elements in the genomes of *H. balaensis*, *T. grandiflorum* and *T. cacao*, as previously shown. The difference in relative copy-number of class I LTR retroelements between the three species is quite striking. Using this method and the estimated genome sizes for the species (Argout et al., 2010), it can be calculated that LTR/Gypsy and LTR/Copia elements make up 9% and 7% of the genome respectively in *T. cacao*. In contrast these make up just 2% and 2% in *T. grandiflorum* and 0.6% and 0.5% in *Herrania* (Table 2.2). *De novo* approaches show the low values in the latter species to be artifacts (see next section). The apparently lower numbers in species distant from the reference are therefore likely due to mapping incompatibility, and this method is therefore not suitable for interspecies studies. The same pattern is also observed even when reads are mapped to the conserved regions of the LTR retrotransposons using less stringent settings in the short read aligner (see A.1 in appendix A). Mapping incompatibility is most likely attributable to retroelement divergence between distant species. This divergence is an important part of genome differentiation between species and potentially has implications for speciation and species divergence. On the other hand de novo methods are efficient in identifying any repetitive sequence that is either specific to a given species or has mutated beyond recognition

that could render it unidentifiable using the mapping approach and therefore likely to be much more accurate for calculating copy numbers in species outside the reference species (see below).

2.3.3 Variation of TE copy number using *de novo* approaches

A major potential problem with studying interspecific variation using a mapping approach is that the reads from *Herrania* and *T. grandiflorum* are heterologously mapped to the *T. cacao* genome. Sequence variation between species affecting mapping quality could potentially have a considerable effect on the apparent frequency of TEs. We therefore also employed a *de novo* approach, using graph-based clustering. The graph-based clustering analysis (Figure 2.3) of the short reads shows considerable differences in the representation of the major families of repetitive elements in the genomes of the three species studied here. Figure 2.3 shows the four largest clusters generated by RepeatExplorer, their identity and genome percentage. The most striking difference is between the two genera, where *H. balaensis* has two extremely large low complexity clusters with combined genome percentage of 19.4%. The two *Theobroma* species are more similar, with their largest clusters containing about 1-2% of the reads used in the graph based clustering. However, cluster 2 in *T. cacao* is largely composed of LTR/Gypsy elements, whereas the top four clusters in *T. grandiflorum* are all identified as low complexity elements. Using graph based clustering it can be calculated that LTR/Gypsy and LTR/Copia elements make up 10% and 5% of the genome respectively in *T. cacao* (comparing well with figures derived from a mapping approach, Table 2.2), whereas in *Herrania* they are 4% and 7% and in *T. grandiflorum* 10% and 9%. This contrasts with the different figures arrived at for the other species using the mapping approach (see above). We conclude from comparing the results of

mapping with the *de novo* approach that mapping quality in interspecific comparison has an important effect on the results.

2.3.4 Intraspecific variation of TE abundance in *T. cacao* using short read mapping and PCA

The data on intraspecific relative abundance of TEs from short read mapping of the eight *T. cacao* accessions was analyzed using principal coordinates analysis (PCA) (Figure 2.4). The first two axes include 81% of the variance. Axis 1 (46%) separates most strongly the Stahel and Amelonado varieties from the Criollo22 variety. TEs with the highest loadings on this axis are DNA transposons such as DNA/hAT and LTR retrotransposon such as Copia. Axis 2 (35%) separates most strongly Scavina-6 from Criollo22. TEs that load most strongly onto this axis are DNA transposons such as Mutator and Harbinger and Gypsy retrotransposons. This analysis clearly separates most of the *T. cacao* samples and shows several clusters of cacao accessions with the first two axes, some of which accord well with phylogenetic relatedness. Due to concerns that these results might be due to a sampling artifact, particularly given that Scavina-6 has the lowest coverage, the PCA analysis was repeated individually on 49 sub-sampled datasets that were equal in size. However, the same general pattern described above was always observed (see A.1 in appendix A).

2.3.5 Sequence conservation of transposable elements in *T. cacao*

Mapping of the short reads of the nine *T. cacao* libraries to the TE reference contigs revealed considerable levels of within species nucleotide variability, as calculated by number of nucleotide variants detected, divided by the length of the reference contigs. The nucleotide

variability in the class II DNA transposons was around 10 times less than in the class I retroelements (Figure 2.5). These results illustrate different modes of sequence evolution in class I retroelements and class II DNA elements in the *T. cacao* studied here.

Comparing the nucleotide variability in two classes of TEs is informative with regard to how these elements evolve on the whole but it sheds no light on what parts of individual elements are causing these differences. LTRDigest was able to identify characteristic features of LTR elements in 355 of the 650 class I families identified in (Argout et al., 2010). In those 355 families, 90% of the nucleotide variability lay outside of protein coding genes and the long terminal repeats (LTRs), while 5% was situated within the LTRs and 5% in genes (Figure 2.6). Reverse transcriptase (RT) was the largest contributor, containing about 3% of total nucleotide variability followed by integrase with about 1.5% and the three remaining genes all contributed less than 1% (see Figure 2.6). However these values are only informative of total variation not rates of variation, because they differ both in length and representation. LTRDigest does not identify all features in all the elements it interrogated. A better representation of the variability of the LTR genes and the long terminal repeat is to divide the number of nucleotide variants by the length of each feature, which yields a comparable estimate to the previously calculated nucleotide diversity. Those calculations show that the genes and the long terminal repeat all share a similar value, ranging from about 0.002 to 0.09, which are similar values to the average nucleotide diversity of the class I LTR retroelements shown in Figure 2.5.

2.4 Discussion

The study of transposable elements (TEs) has been revolutionized by the increased availability and lowered costs of next generation sequencing (NGS) technologies (Chaparro and Sabot, 2012). NGS methods have not only been applied in TE studies of plants with high quality whole genomic sequences available such as *Zea luxurians* (Tenaillon et al., 2011) and rice (*Oryza sativa)* (Sabot et al., 2011) but also in organisms with limited genomic resources available such as barley (*Hordum vulgare*), pea (*Pisum sativum*) and banana (*Musa acuminata*) (Macas et al., 2007; Wicker et al., 2008; Hribova et al., 2010). These studies demonstrate a strong correlation between copy-number estimation of TEs by traditional molecular methods and methods that count short reads from NGS experiments (Macas et al., 2007; Tenaillon et al., 2011). It was therefore not surprising that the copy-number estimation of TEs in this study fitted very well with previously published estimates in *T. cacao*, both in regard to the overall TE abundance in the genome, around 23%, and in the copy-number of the most abundant class I retroelement (Argout et al., 2010). Our study therefore confirms the utility and reliability of studying genomic repeats using short reads directly.

2.4.1 Different levels of nucleotide conservation in class I and class II TEs

The two major classes of TE, class I retroelements and class II DNA transposons, have been recognized for a long time as two fundamentally different groups of mobile elements probably present in all eukaryotic genomes (Wessler, 2006). The results presented in this study illustrate a considerable difference in the apparent conservation of the TEs in the genome of *T. cacao*, where the class I retrotransposons show significantly higher levels of heterogeneity, represented by an order of magnitude higher level of nucleotide diversity (Figure 2.5). This may be simply because

DNA transposons are more narrowly defined at the superfamily level. However, one possible biological explanation of the high levels of heterogeneity in class I retroelements results from their transposition mechanism, as described in detail in (Craig et al., 2002). Class I retrotransposons move about as a RNA intermediate, which is encoded into DNA before re-entry into the host genome by their endogenous reverse transcriptase enzyme, which is known to be low-fidelity, causing a high mutation rate (Gabriel et al., 1996; Craig et al., 2002).

2.4.2 Inter- and intraspecific differences in TE abundance in *H. balaensis*, *T. grandiflorum* and *T. cacao*

Transposable elements are known to cause large inter- and intraspecific differences in the size and composition of plant genomes, demonstrated in barley (*Hordeum vulgare*) (Vicient et al., 1999; Kalendar et al., 2000) and rice (*Oryza sativa*) (Huang et al., 2008). However, our study only found relatively subtle intraspecific differences of the overall TE abundance in T. cacao. Nevertheless this slight intraspecific variation in TE copy number does potentially contribute to the variable genome sizes of different Cacao accessions reported in the supplementary material in the *T. cacao* genome paper and other sources (Figueira et al., 1992; Marie and Brown, 1993; Argout et al., 2010). Furthermore using a PCA approach to differentiate accessions based on TE abundance, wide separations do occur (Figure 2.4). The ability to separate cacao accessions according to TE composition despite the fact that they are all closely related, some being of recent hybrid origin (Argout et al., 2010). As massively parallel sequencing (MPS) costs fall, there is interest in using MPS to identify accessions, and such use has been called “ultra-barcoding” (Kane et al., 2012). This paper shows that data generated for ultra-barcoding could

also be used for “transposon composition fingerprinting” of cacao accessions (i.e. identification based on a unique spectrum of transposon composition).

2.4.3 Mapping vs. *de novo* approaches to studying TEs from short reads

Our results (Table 2.2) suggest that the mapping approach, while reliable within the reference species (*T. cacao*), is unreliable in interspecific comparisons, at least for some TE families. The mapping approach reports considerable differences in the composition of repetitive elements in the three species studied (Figure 2.2). Apparently the genomes of *T. grandiflorum* and *H. balaensis* are significantly deficient in many LTR retrotransposon families that are very abundant in T. cacao (Figure 2.3). However this difference may be at least partly caused by low interspecific mapping quality of the short reads, since our reference contigs originate from the genome of *T. cacao*. The LTR retrotransposon families in particular have high nucleotide diversity (Figure 2.5), which is likely to cause problems in the mapping of the short reads.

The evidence for the failure of the mapping approach in interspecific comparisons comes from the *de novo* approach of graph based clustering using RepeatExplorer. This demonstrates that in both *T. grandiflorum* and *H. balaensis* the LTR TE families are more abundant than the mapping approach suggested (Table 2.2 and Figure 2.3). More importantly the graph based clustering showed that the composition of *H. balaensis* and *T. grandiflorum* is quite different from *T. cacao*. Therefore we conclude that mapping based approaches are well suited to look at TE evolution in an intraspecific manner whereas de novo methods, such as graph based clustering, are much more useful in the exploration of differences in repetitive elements across species boundaries.

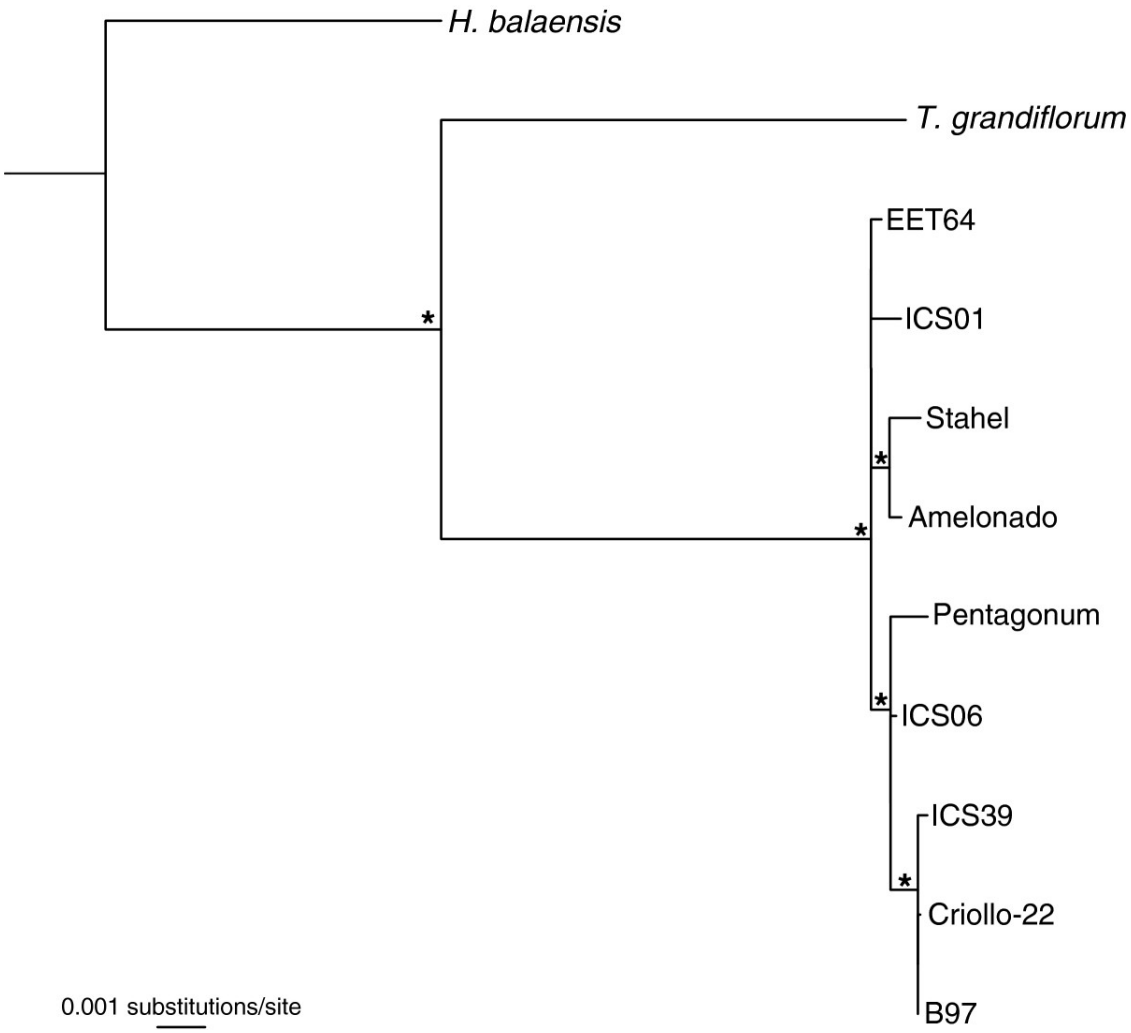


Figure 2.1 Phylogeny of *Herrania balaensis*, *Theobroma grandiflorum* and nine of the *T. cacao* varieties. The phylogenetic tree was constructed using partial sequence data of 97 ultra conserved orthologus sequences (UCOS). *Theobroma cacao* cv. Scavina-6 was excluded from the phylogenetic analysis due to low sequencing coverage. Nodes marked with asterisk have high bootstrap support (>90%).

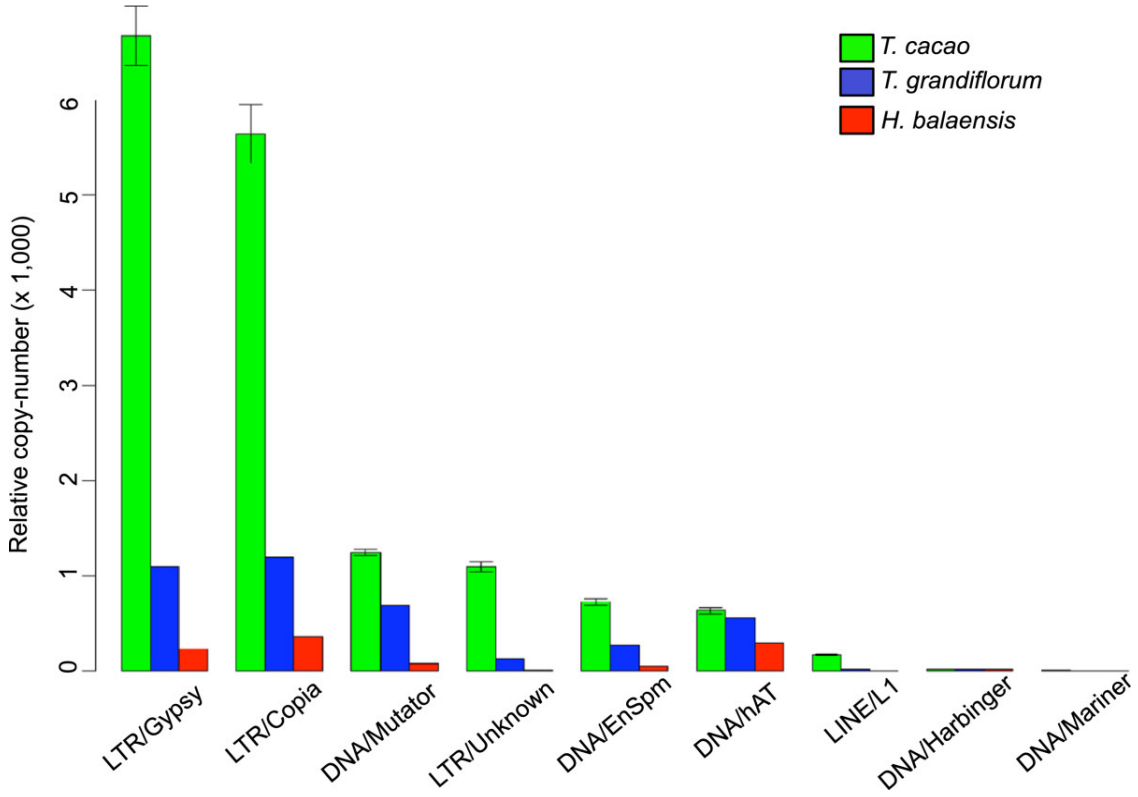


Figure 2.2 Relative copy-number of transposable elements using reference based mapping. Relative copy-numbers of the TE super-families in the three species represented with bar plots. Relative copy-number was calculated by dividing the total coverage of each super-family, within a sample, by the sample's mean UCOS coverage. The much lower recovery of transposable elements in the other species is apparently due to mapping failure as the graph based clustering indicates that TE copy numbers are comparable in all species. Error bars represent standard deviation and correspond to intraspecific variation.

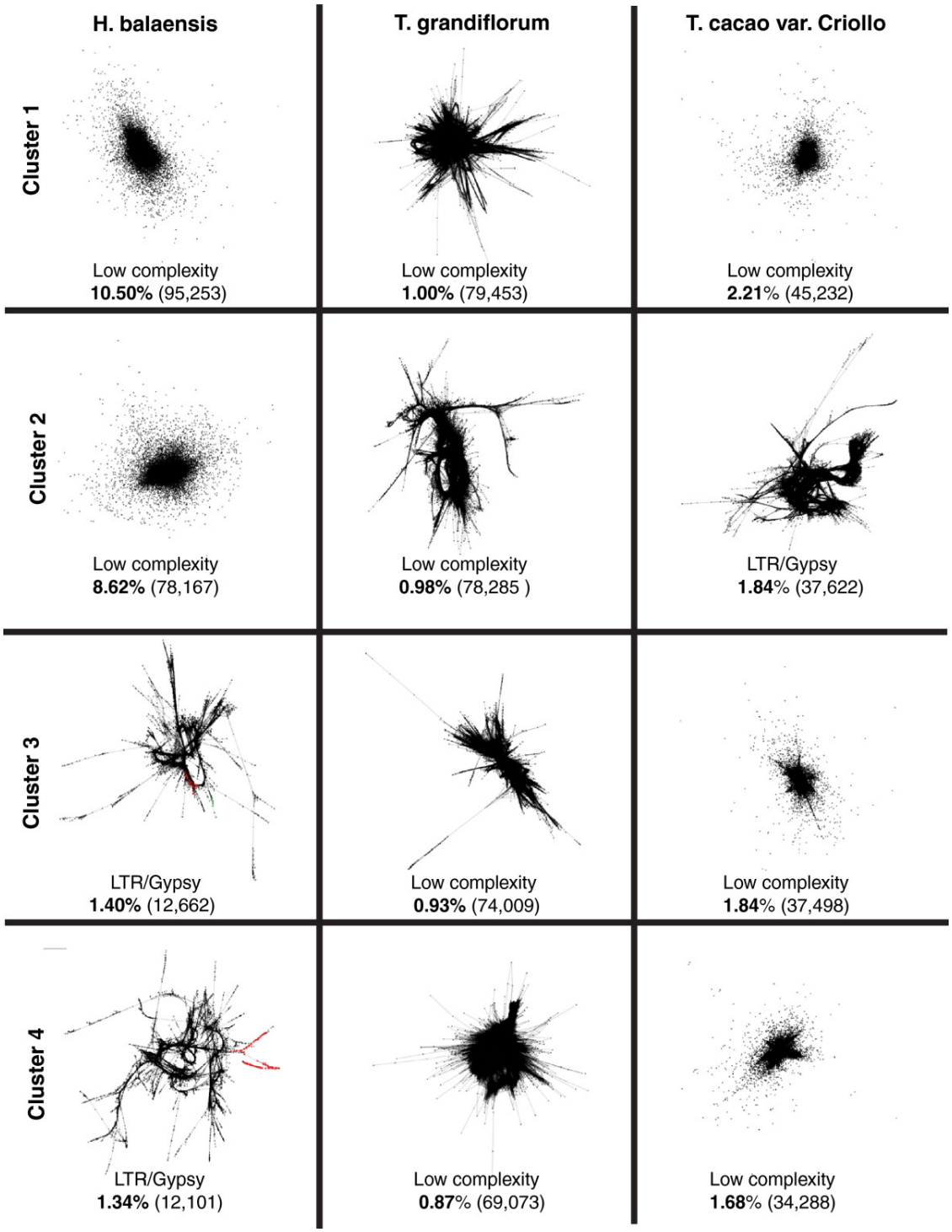

H. balaensis
T. grandiflorum
T. cacao var. Criollo
Cluster 1
Low complexity
10.50% (95,253)
Low complexity
1.00% (79,453)
Low complexity
2.21% (45,232)
Cluster 2
Low complexity
8.62% (78,167)
Low complexity
0.98% (78,285)
LTR/Gypsy
1.84% (37,622)
Cluster 3
LTR/Gypsy
1.40% (12,662)
Low complexity
0.93% (74,009)
Low complexity
1.84% (37,498)
Cluster 4
LTR/Gypsy
1.34% (12,101)
Low complexity
0.87% (69,073)
Low complexity
1.68% (34,288)

Figure 2.3 Graph based clustering analysis of repetitive elements in the three species. Graph layouts of the four largest clusters of repetitive elements detected in the graph based clustering analysis. *Herrania balaensis* is shown on the left, *T. grandiflorum* in the middle and *T. cacao* cv. Criollo on the right. Clusters are ordered by size, with largest at the top and fourth largest at the bottom. Below each graph layout is the class of the repetitive element, the genome percentage of each cluster and number of paired reads belonging to it in parentheses. Coloured regions in the some graphs represent conserved domains identified by RepeatExplorer. A total of 11,243,224 reads were used in the graph based clustering.

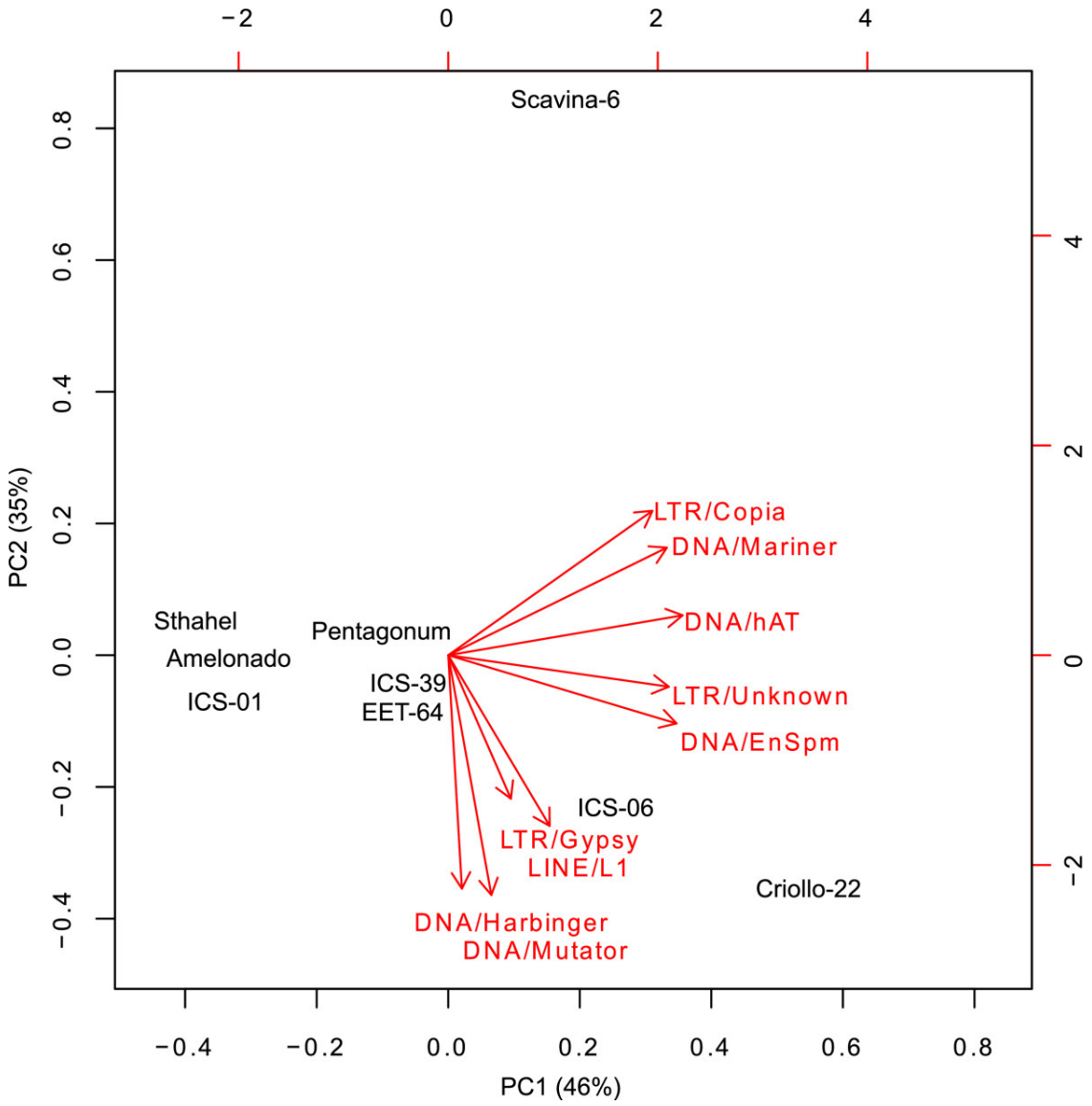


Figure 2.4 PCA of the transposable element composition in the *Theobroma cacao* genotypes. A biplot from a principal component analysis (PCA) using the standardized abundance of each TE super-family as explanatory variables. Percentage of the explained variance is shown in parentheses in the legend of the x- and y-axis.

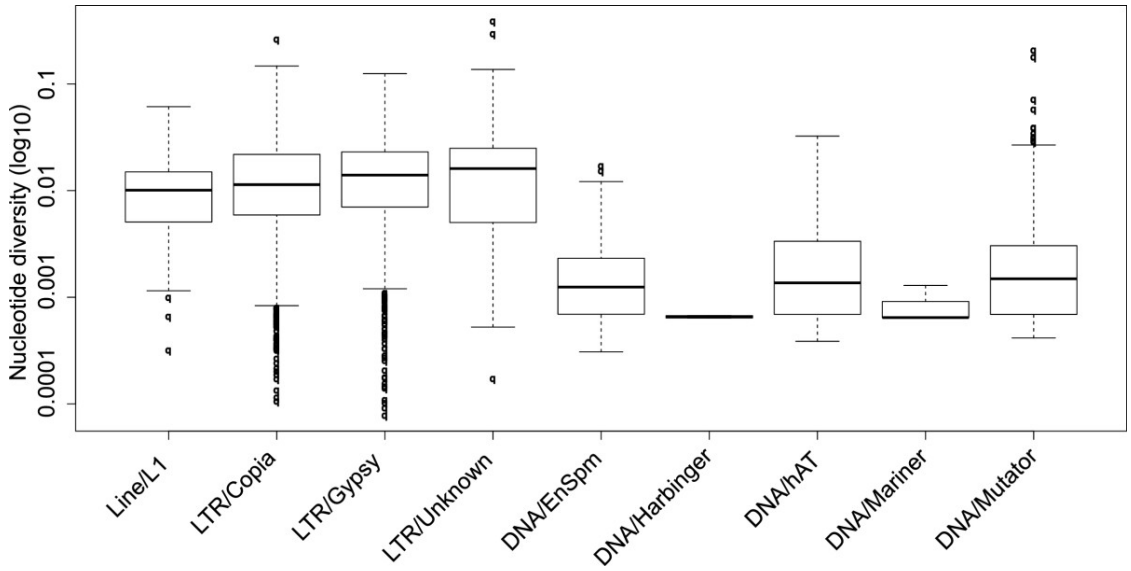


Figure 2.5 Nucleotide variability of transposable elements in *Theobroma cacao*. Box plot showing the nucleotide diversity across the super-families in *T. cacao*. This shows that DNA transposons have less variation at the superfamily level (see Discussion). Analyses were performed on standardized data sets (Methods) and values are presented transformed to a log10 scale.

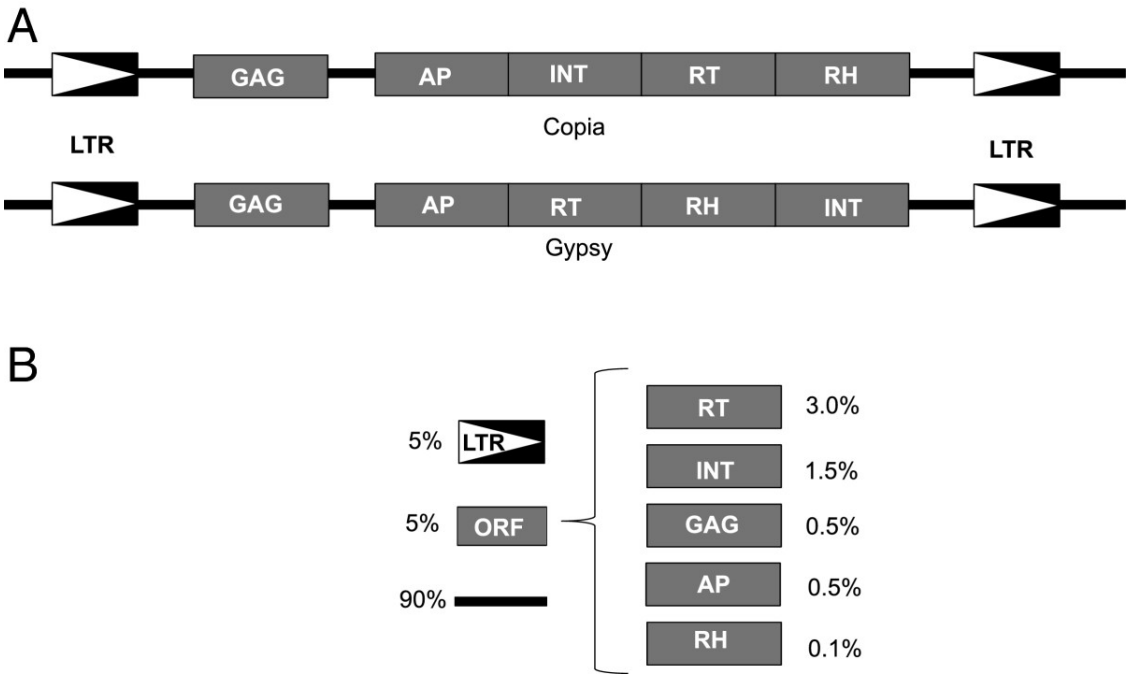


Figure 2.6 Nucleotide diversity of LTR/Copia and LTR/Gypsy elements in *Theobroma cacao*. (A) Schematic diagram of the structure of the two most common LTR retrotransposons super-families in the *T. cacao* genome. (B) Partitioning of nucleotide variation is shown as percentage values next to each of the retrotransposon components. The white arrows with black background represents the long terminal repeat (LTR), black line regions in between open reading frames (ORFs) and LTRs and grey boxes represent the following open reading frames: Reverse transcriptase (RT), integrase (IT), capsid protein (GAG), aspartic proteinase (AP) and Rnase H (RH).

Name	Chloroplast haplotype[1]	Read length (bp)	No. reads after trimming	UCOS coverage (SE*)
EET-64 (*T. cacao*)	Criollo	60	6.5E+07	5.1 (0.1)
Criollo-22 (*T. cacao*)	Criollo	60	4.2E+07	4.2 (0.1)
Stahel (*T. cacao*)	Criollo	60	5.7e+07	4.8 (0.1)
Pentagonum (*T. cacao*)	Criollo	80	5.5E+07	4.7 (0.1)
ICS39 (*T. cacao*)	Criollo	80	6.0E+07	5.4 (0.1)
Amelonado (*T. cacao*)	Forastero	60	6.8E+07	5.6 (0.1)
ICS06 (*T. cacao*)	Forastero	80	6.1E+07	5.7 (0.1)
ICS01 (*T. cacao*)	Forastero	60	4.9E+07	4.2 (0.1)
Scavina-6 (*T. cacao*)	Forastero	60	2.9E+07	1.7 (0.03)
T. grandiflorum (Cupuaçu)	na	60	6.8E+07	5.3 (0.1)
H. balaensis	na	80	7.4E+07	8.3 (0.2)

1: See Kane et al. (2012)
*SE: Standard error

Table 2.1 Sequence summary statistics. Illumina sequence summary statistics and observed average coverage of the UCOS contigs for *Theobroma cacao*, *T. grandiflorum* and *Herrania balaensis* based on Burrows-Wheeler Aligner (BWA) alignments.

Reference based mapping			
	T. cacao	*T. grandiflorum*	*H. balanensis*
LTR-Gypsy	9%	2%	0.6%
LTR-Copia	7%	2%	0.5%
Graph based clustering			
	T. cacao	*T. grandiflorum*	*H. balanensis*
LTR-Gypsy	10%	10%	4%
LTR-Copia	5%	9%	7%

Table 2.2 LTR retrotransposon frequencies in the three species estimated with two different methods. Comparison of estimated LTR retrotransposon frequencies as percentages of the genome, calculated with reference based mapping (upper half) and graph based clustering (lower half). Within *T. cacao* there is little discrepancy between the methods. Heterologous mapping between species produces different results suggesting that graph-based clustering may be more appropriate for inter-species comparisons (see Discussion).

Chapter 3: Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics

3.1 Introduction

Polyploidy, the duplication of whole genomes, is an important evolutionary event that is especially prevalent in plants (Otto and Whitton, 2000). A recent study has revealed that all angiosperms have undergone at least two rounds of ancient whole-genome duplication (Jiao et al., 2011) in addition to several younger, lineage-specific events (Jiao et al., 2012). These events are thought to have been very important in the evolutionary diversification of flowering plants (Adams and Wendel, 2005; Soltis et al., 2009; Jiao et al., 2012). In addition to these ancient polyploidy events, recent whole-genome duplications are very common in most extant plant lineages (Otto and Whitton, 2000; Adams and Wendel 2005; Wood et al., 2009). This is especially the case for the majority of the world's most important crop species, in which polyploidy seems to be particularly prevalent (Adams and Wendel, 2005). However, the genome complexity caused by genome duplications can be troublesome for crop genomics, for instance in genome-wide association studies of polyploid crops (Harper et al., 2012). It is therefore of considerable importance to characterize fully the genome history of crop species in order to take this into account in crop research.

The flax genus (*Linum*) contains about 180 species that are spread across six continents. It is thought to have originated about 46 million years ago (MYA), making it a relatively old genus (McDill et al., 2009; McDill and Simpson, 2011). The genus is divided into numerous sections. A large, predominately blue-flowered clade contains the sections Dasylinum and Linum, while the group of yellow-flowered flaxes contains the sections Cathartolinum, Linopsis

and Syllinum (McDill et al., 2009; McDill and Simpson, 2011). Cultivated flax (*Linum usitatissimum*) is an important source of high-quality fibers (Mohanty et al., 2000) and seed oil (Green, 1986). The oil has industrial uses as well as considerable perceived health benefits (Singh et al., 2011). Its genome was recently sequenced (Wang et al., 2012), resulting in the discovery that it had undergone a whole-genome duplication (WGD) around 5–9 MYA. As an extension to this finding we wished to examine the possibility that another, older, *Linum* specific polyploidy event might have occurred some time earlier in the evolutionary history of cultivated flax. Until now, this hypothesis could not be tested due to insufficient genomic data from related species. In this study we use transcriptome sequences of 11 *Linum* species to identify and characterize whole-genome duplication events in the evolutionary history of cultivated flax.

3.2 Material and methods

3.2.1 Illumina sequencing and *de novo* assembly

Total RNA from several tissue types was extracted and used to make Illumina sequencing libraries using methods described by Johnson et al. (2012). The libraries were multiplexed and sequenced on an Illumina HiSeq platform using paired-end chemistry. Quality trimming of the Illumina reads is described in the Methods section of Johnson et al. (2012). All species used in this study had a single library constructed from a pooled RNA sample from various tissue types, except for *Linum usitatissimum* and *L. perenne*, which had certain tissue types separated into individual libraries. However, they were later combined into single-species libraries, since analyses of individual tissue libraries did not change any findings of the paper. A total of 12 species were used in this study: 11 Linum species and *Bischofia javanica* (Phyllanthaceae),

which was used as an outgroup in the phylogenetic analyses due to its relatively close relationship with the Linaceae (Xi et al., 2012).

The short Illumina reads were assembled in a de novo fashion using Trinity v.r2013-02-25 (Grabherr et al., 2011) with the program's default settings. In order to reduce the number of almost identical sequences in the assembly, contigs with a sequence similarity higher than 98 % were clustered together using the CD-HIT-EST program in the CD-HIT package v.4.6 (cd-hit-est flags: -c 0·99 -l 299 -d 0) (Li et al., 2001; Li and Godzik, 2006). This step also removed all contigs shorter than 300 bp from the assembly.

3.2.2 Identification of orthologues and phylogenetic analyses of *Linum*

Transcripts were then translated into their corresponding amino acid sequences using prot4EST (Wasmuth and Blaxter, 2004). Prot4EST uses BLASTX (Altschul et al., 1997) for certain parts of its amino acid translation pipeline and therefore requires a BLAST database in amino acid format. For this purpose we used the annotated protein sequences from the published flax (*L. usitatissimum*) genome (Wang et al., 2012), which can be downloaded from Phytozome (Goodstein et al., 2012). The best amino acid transcript from each cluster was determined by its length and number of internal stop codons, using a Python script, and was chosen as the cluster's representative in the assembly. OrthoMCL v.2.0.3 (Li et al., 2003) was used to identify orthologous sequences in the assemblies. An E-value of 1E-5 was used in the all versus all BLASTP step of the OrthoMCL pipeline, and percentMatchCutoff was set to 50 and evalueExponentCutoff to −5 in the orthomclPairs step. Other parts of OrthoMCL were executed with the pipeline's default settings.

A phylogeny of the *Linum* species used in this study was constructed using a selected subset of the orthologue groups generated by OrthoMCL. A Python script was used to extract orthologue groups that matched the following criteria: (1) the orthologue group had to contain contigs from all species but (2) could not contain more than one contig from each species (i.e. putative single copy genes). These singleton orthologue groups were used in the subsequent phylogenetic analyses. The amino acid and nucleotide sequences of transcripts in these orthologue groups were extracted using a Python script. MAFFT v.7.0.5.b (Katoh and Standley, 2013) was used to align the amino acid sequences using the -auto flag. The alignments were then converted to their corresponding nucleotide sequence using RevTrans v.1·4 (Wernersson and Pedersen, 2003). Leading and trailing gaps were removed from the alignments using a Python script and then trimmed further with trimAl v.1.2 (Capella-Gutiérrez et al., 2009) with the -automated1 flag. Alignments smaller than 300 bp were discarded from further analyses. The appropriate model of nucleotide substitution for the trimmed nucleotide alignments was determined with jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) using the Akaike information criterion (AIC). These alignments were used to generate species phylogenies using two methods.

All nucleotide alignments were concatenated into a single matrix that was analysed with MrBayes v.3.2.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Each alignment was defined as a separate partition in the matrix and a Python script was used to incorporate the jModelTest outputs into the MrBayes block of the matrix, ensuring that the appropriate base substitution model was used for every alignment. Model parameters were unlinked across all partitions. *Bischofia javanica* was used as the outgroup to confirm the rooting of the analyses. Two runs were started with 2,000,000 generations and burn-in set at 25 %.

Convergence of the analyses was checked by running two independent chains and manually inspecting traces using Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

A second species phylogeny was constructed using the STAR method (species tree estimation using average ranks of coalescences) (Liu et al., 2009b), which is based on the multispecies coalescent model (Rannala and Yang, 2003). The STAR method uses the average ranks of coalescent events in a collection of gene trees to construct a single species topology using a distance method. First we generated individual gene trees for each of the previously mentioned trimmed nucleotide alignment, using RAxML v.7.2.6 (Stamatakis, 2006) with 10 search replicates. *Bischofia javanica* was set as the outgroup to confirm the rooting of the phylogenetic analysis. The appropriate model of nucleotide substitution for each alignment was parsed from jModelTest outputs using a Python script. The star.sptree function in the phyBASE v.1.3 package (Liu and Yu, 2009), under R v.2.15.3 (R Core Team, 2014), was used to generate a single species topology from all gene trees. Branch lengths were estimated and added to the STAR tree using GARLI v.2.0 (Zwickl, 2006), by optimizing the model parameters of the concatenated matrix, which were initially estimated by jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012). Node support of the STAR tree was acquired by generating 1000 multilocus bootstrap replicates (Seo, 2008) with the bootstrap.mulgene function in phyBASE, analysing each bootstrap replicate with PhyML v.3.0 (Guindon et al., 2010) and constructing 1,000 STAR trees in phyBASE. A consensus of the 1,000 STAR trees was generated using the consense program in the PHYLIP package v.3·69 (Felsenstein, 2005). The transcriptome to phylogeny methods described above are implemented in a pipeline (T2Phy) under development by one of us (S.S.).

3.2.3 Whole-genome duplication inference from age distributions of paralogues

The DupPipe pipeline (Barker et al., 2008) was used to look for evidence of whole-genome duplication events in the 11 *Linum* species. The pipeline inferred pairs of paralogues within the transcriptome assemblies and estimated their divergence based on synonymous substitution rates (Ks), which are used as a proxy for the age of the duplicated gene pair. DupPipe uses a discontinuous MegaBlast (Zhang et al., 2000; Ma et al., 2002) within each assembly to cluster contigs into gene families based on 40 % sequence similarity over at least 300 bp. The appropriate reading frame was estimated by comparing each pair of sequences with a large set of protein sequences (Wheeler et al., 2007) using BLASTX (Altschul et al., 1997). Each nucleotide sequence was then paired with its best protein hit and only genes with a minimum of 30 % similarity over at least 150 sites were retained for further analyses. Genwise 2.2.2 (Birney et al., 2004) was used to align each gene to its best protein hit in order to determine the correct reading frame. The nucleotide contigs were then converted to their corresponding amino acid sequence using the highest-scoring DNA–protein alignment from GeneWise. Alignments of duplicated gene pairs were generated using MUSCLE 3.6 (Edgar, 2004) and the aligned amino acids were converted back to DNA sequences using RevTrans 1.4 (Wernersson and Pedersen, 2003). Ks values (synonymous substitution rates) for each duplicate gene pair were estimated using the codeml program in the PAML package (Yang, 1997) under the F3 × 4 model (Goldman and Yang, 1994). Duplicated gene pairs that could be identified as transposable elements were removed from the dataset. Due to concerns that extremely low Ks values were noise caused by alternative splicing or assembly artefacts, all Ks values lower than 0.001 were removed from the dataset. Furthermore, we excluded all Ks values larger than 2 in order to minimize saturation effects (Vanneste et al., 2013).

The number of significant features, i.e. peaks, in the age distributions of gene duplicates was estimated using SiZer (Chaudhuri and Marron, 1999). SiZer looks for significant changes in kernel density by performing a Gaussian smoothing with a wide range of bandwidths and has been extensively used in the investigation of paleopolyploidy (Barker et al., 2008, 2009; Shi et al., 2010). SiZer identified two significant peaks in some of the transcriptome assemblies, which were then analysed further using mixture models. EMMIX (McLachlan et al., 1999) was used to fit a model with two normal distributions using maximum likelihood in order to estimate the position of the two peaks in the dataset. Mixture models are useful in characterizing peaks generated by paleopolyploidy events, since their distribution is expected to be roughly Gaussian (Blanc and Wolfe, 2004; Schlueter et al., 2004). We used the EMMIX function in the EMMIX R package v.1·0·18 [downloaded from http://www.maths.uq.edu.au.ezproxy.library.ubc.ca/~gjm/mix_soft/EMMIX_R/index.html (19 December 2013)] on log-transformed Ks values, with two normally distributed components.

3.2.4 Inference of a whole-genome duplication event from phylogenetic analysis of paralogues

To establish phylogenetic evidence for the putative paleopolyploidy event inferred by DupPipe, phylogenies of all orthologue groups containing a paralogue pair with a Ks value between 0.4 and 1.0 were constructed. This range was established based on two lines of evidence. First, results from the mixture models produced by EMMIX (McLachlan et al., 1999) demonstrated that the median of the paleopolyploidy event was around 0.68. We incorporated this estimate in the modelling of the effects a polyploidy event at that time would have on the Ks distribution, using the R scripts provided in Cui et al. (2006). The modelling showed that the effects of a

polyploidy event around Ks 0.68 would be most significantly noticed around Ks 0.4–0.8. Second, we examined the results from Gaussian smoothing with SiZer, which were largely in agreement with the modeling results. However, in order to be inclusive, we increased the upper Ks limit by 0.2 in order to investigate more paralogues phylogenetically. Amino acid and nucleotide sequences of orthologue groups containing the paralogues were extracted from all species using a Python script and amino acid alignments were generated using MAFFT v.7.0.5.b (Katoh and Standley, 2013) with the -auto flag. The alignments were converted to their corresponding DNA sequences using RevTrans v1.4 (Wernersson and Pedersen, 2003) and trimmed using trimAl v.1.2 (Capella-Gutiérrez et al., 2009) with the -automated1 flag. Phylogenies from the trimmed nucleotide alignments were inferred using RAxML v.7.2.6 (Stamatakis, 2006), with ten search replicates and the GTR + gamma model of nucleotide substitution. Node support of each tree was estimated using the 50 % majority rule consensus of 100 bootstrap replicates from RAxML. Phylogenies were converted to NEXUS format and combined into a singe file using a Python script, which could be analysed using Dendroscope v.3.2.8 (Huson and Scornavacca, 2012).

The paralogue gene trees were inspected manually and split into two major groups based on the phylogenetic pattern observed. We first separated phylogenies that were uninformative regarding a polyploidy event in the *Linum* genus. Those phylogenies included orthologue groups that contained multiple gene families and paralogues that were likely generated by multiple tandem duplications or over-assembly. These also included orthologous groups that had a large number of missing taxa and showed strong phylogenetic conflict within the yellow- or blue-flower clades. The remaining phylogenies showed strong phylogenetic patterning of paralogues and are therefore potentially informative regarding a polyploidy event in the evolutionary history

of the *Linum* species. The only consistent pattern observed in these kinds of trees was the occurrence of a single clade of yellow-flowered species and two blue-flowered clades. We excluded any of these phylogenies that had excessive numbers of missing taxa, i.e. fewer than two yellow-flowered species, or if either of the blue-flowered clades contained fewer than three species.

3.3 Results

3.3.1 Illumina sequencing and *de novo* assembly

The Illumina sequencing yielded between 19 and 83 million high-quality paired-end reads per species (Table 3.1). The *de novo* assembly of these reads resulted in 22,416–48,269 contigs larger than 300 bp per library.

3.3.2 Phylogenetic analysis of *Linum*

A total of 34,894 orthologous groups were identified by OrthoMCL, of which 413 were used to generate a phylogeny for the 11 *Linum* species. In these 413, all species had a single contig representing each group. The average alignment contained 546 characters (s.d. 212.97) and the combined length of all alignments was 225,891 characters. The two methods used (Bayesian analysis of a concatenated matrix and the coalescence-based STAR method) retrieved identical topologies, very similar branch lengths and a 100 % support value for every node in the phylogeny. As the two trees were very similar, only the STAR tree is presented here (Figure 3.1). It is of interest to note that this phylogeny is almost identical to the most recently published *Linum* phylogeny (McDill et al., 2009). The 11 *Linum* species were split by a long branch into two major clades: (1) a mostly yellow-flowered clade containing *L. flavum, L. macraei, L.*

strictum and *L. tenuifolium*; and (2) a predominantly blue-flowered clade that comprises all other species (Figure 3.1). There were two exceptions to this flower colour rule, one in each clade, where *L. tenuifolium* flowers are white or pale pink and *L. grandiflorum* is red-flowered. However for simplicity's sake, all subsequent reference to these clades will be based on their dominant flower colour: yellow (y) and blue (b).

3.3.3 Paralogue age distributions

The possible presence of paleopolyploidy events in the evolutionary history of the 11 *Linum* species was first investigated by identifying paralogues within each transcriptome assembly and analysing their age distribution. An average of 3,017 paralogue pairs (s.d. 1,809) were identified in each of the transcriptome assemblies. Visualization of the duplicate age distribution revealed two types of patterns in genome evolution. Seven out of 11 *Linum* species showed a noticeable increase in Ks value frequency (i.e. a shallow peak) around 0.6 (Figure 3.2E–K), compared with the L-shaped curve expected in species that have not undergone a whole-genome duplication recently in their evolutionary history (Blanc and Wolfe, 2004). However, four of our species showed a contrasting pattern, with no visible change of slope around Ks 0.6 (Figure 3.2A–D). When this pattern is compared with the *Linum* phylogeny, it becomes clear that the peak at Ks 0.6 is restricted to species belonging to the blue clade (Figure 3.2), while being uniformly absent in the yellow clade. The presence of this peak in all blue clade species around the same Ks value strongly suggests a unique polyploidy event in an ancestor of this clade some time after the split from the yellow-flowered *Linum* species.

The statistical significance of the changes in slopes in the duplicate age distributions was tested with SiZer (Chaudhuri and Marron, 1999). The results of these analyses are shown in the

form of SiZer plots, positioned underneath their corresponding duplicate distribution in Figure 3.2A pattern consistent with a paleopolyploidy event was observed in all of the blue-flowered Linum species but in none of the species belonging to the yellow-flowered clade (Figure 3.2) (see the legend of Figure 3.2 for a detailed description of the SiZer plots). The evidence for a paleopolyploidy event is represented by blue areas in the SiZer plots in the bottom half of Figure 3.2E–K. These blue areas correspond to a significant peak ($P < 0.05$) in the frequency of duplicate pairs in all species that is centered around Ks 0.6.

In order to determine the age of this putative paleopolyploidy event more precisely, mixture models with two normally distributed components were fitted to each of the paralogue age distributions of the blue-flowered *Linum* species. The mixture models fitted one component to the first peak in the age distribution, around Ks 0.01, which reflected the continuous birth and death model of gene evolution generated by frequently occurring single-gene duplications (Blanc and Wolfe, 2004). The second component was fitted to the older peak with an average median of Ks 0.68. Using this Ks value and two commonly used measures of the synonymous mutation rate (Koch et al., 2000; Lynch and Conery, 2000), the polyploidy event may be estimated to have occurred 23–42 MYA.

3.3.4 Phylogenetic analysis of paralogues

A total of 767 orthologous groups containing paralogues (henceforth referred to simply as paralogue groups) were extracted and their phylogenies manually inspected for evidence of whole-genome duplication. Most of the paralogue groups (587) were uninformative with regard to the presence or absence of a polyploidy event as the pattern of duplication showed no obvious phylogenetic pattern. About 260 paralogue groups clearly consisted of multiple gene families

clustered together into a single group. The source of paralogy in the remaining 327 uninformative groups was likely to be from (1) multiple tandem duplications or assembly artefacts (239 groups), (2) incorrect orthology inference (62) or (3) numerous missing taxa (26) (see Table 3.2).

The remaining 180 (23.5 %) trees showed clear phylogenetic patterning of gene duplication and were therefore potentially informative of whole-genome duplication events. Remarkably, all 180 trees were consistent with the whole-genome duplication event proposed here in that they divided into three clades, the yellow-flowered species (clade y) and two duplicate clades of the blue-flowered species (clades b-I and b-II), each reflecting the organismal phylogeny of the component species and implying a gene duplication in the blue species but not in the yellow. Figure 3.3 shows an example of such a phylogeny (from a chaperone-like gene) and it shows three well-supported clades: one is composed of the yellow-flowered *Linum* species and two blue-flowered *Linum* clades are observed. Furthermore, the blue-flowered clades are sisters to each other, and when combined are sister to the yellow-flowered *Linum* species. Finally, individual species within each of the three clades followed the species phylogeny in Figure 3.1. This phylogeny did not, however, show evidence for the later polyploidy event that is specific to *L. bienne*/*L. usitatissimum* (for expected reasons, see Discussion). It is also important to note that very few of the 180 paralogue phylogenies were as complete as shown in Figure 3.3. Most of them had some missing genes or taxa, which is not surprising due to the nature of transcriptomic data. Nevertheless, all were consistent with a shared whole-genome duplication event in the evolutionary history of the blue-flowered *Linum* species, and this is the most likely cause for the large number of phylogenies supporting the pattern in Figure 3.3.

3.4 Discussion

3.4.1 Consistency of date estimation

Our results demonstrate that blue-flowered *Linum* species (sections Linum and Dasylinum) (McDill et al., 2009) underwent a whole-genome duplication after they split from the rest of *Linum*. This split occurred near the base of the genus and has previously been estimated to have occurred 41–46 MYA (McDill et al., 2009; McDill and Simpson, 2011). The former study also estimated the age of the most recent common ancestor of the blue-flowered clade to be 29–32 million years. We estimate the duplication event to have occurred 23–42 MYA, which is consistent with the independently derived phylogenetic estimates of divergence within the genus *Linum*. The difficulty of accurately inferring the age of ancient duplication events is well established (Doyle and Egan, 2010; Vanneste et al., 2013), so the close correspondence between our observations and the dates given by McDill et al. (2009) provides some confidence in the reliability of the dating. It also raises the possibility that species diversification in the blue-flowered clade may have been driven, at least in part, by the whole-genome duplication event, as has been suggested for other groups (Vamosi and Dickinson, 2006; Soltis et al., 2009).

3.4.2 Relationship between the two polyploidy events in the evolution of cultivated flax (*Linum usitatissimum*)

It is important to note that the paleopolyploidy event described here was not obvious in the single-species analysis of duplicated genes using the fully sequenced genome of cultivated flax (Wang et al., 2012). This highlights the importance of using a phylogenomic approach, as has been shown elsewhere (Jiao et al., 2011; Van de Peer, 2011). Cultivated flax belongs to the blue-flowered clade and therefore shares the whole-genome duplication event. The Ks and SiZer plots

of *L. usitatissimum* and *L. bienne*, its sister species (Figure 3.2J, K) do indeed show clear peaks around Ks 0.68, but they are not as large as in the other blue-flowered species (Figure 3.2E–I). We do not know the reason for this, but it may be related to the fact that *L. usitatissimum* and *L. bienne* underwent an independent polyploidy event 5–9 MYA (Wang et al., 2012). It would be interesting to investigate whether this later 'mesopolyploidy event' (sensu Guerra, 2008; Schubert and Lysak, 2011) caused accelerated duplicate gene loss, thereby attenuating the signal from the palaeopolyploidy event. If this is true, it follows that it might be more difficult to pinpoint individual polyploidy events in lineages that have undergone multiple whole-genome duplication events.

There is little evidence in the present study for the later polyploidy event specific to *Linum bienne*/*Linum usitatissimum*. This is to be expected as the present study focused on the detection of ancient events in multiple species, and used entirely Illumina short-read data in contradistinction to the analyses included in Wang et al. (2012), which used the completely assembled flax genome and full-length cDNA. Focusing on ancient events allowed us to be conservative in excluding very closely related duplicates in case of assembly error or other artefacts (important because of our use of relatively low-depth short-read data), and this is likely to have had the effect of diminishing or eliminating the signal from the more recent event. This illustrates the importance of specifically targeted studies to discover genomic events at different periods of evolutionary history.

3.4.3 Polyploidy and chromosome number

This ancient polyploidy event described here is not evident from an examination of the published chromosome numbers of the species (see Table 3.1). However, as this is an ancient event (~30 MYA) the absence of a chromosomal signature is not surprising. Whole-genome duplication events have been divided into three classes based on chromosomal repatterning (Guerra, 2008; Schubert and Lysak, 2011). As chromosome repatterning requires time, these classes are usually correlated with age. The classes are: (1) neopolyploidy, in which chromosomes are still in multiples of related diploids, with little if any, chromosome repatterning (usually very recent, Holocene, <11.000 years ago, to Pleistocene, <2.5 MYA); (2) mesopolyploidy, in which there may be some chromosome number reduction and chromosome repatterning, but polyploidy is still suggested by higher chromosome number (usually early Pleistocene to late Tertiary, <10 MYA); and (3) palaeopolyploidy, in which there is complete diploidization and considerable chromosome number reduction (usually Tertiary or older, >10 MYA). The whole-genome duplication event described here obviously falls into the last category and chromosome number reduction is to be expected as a normal pattern of palaeopolyploids.

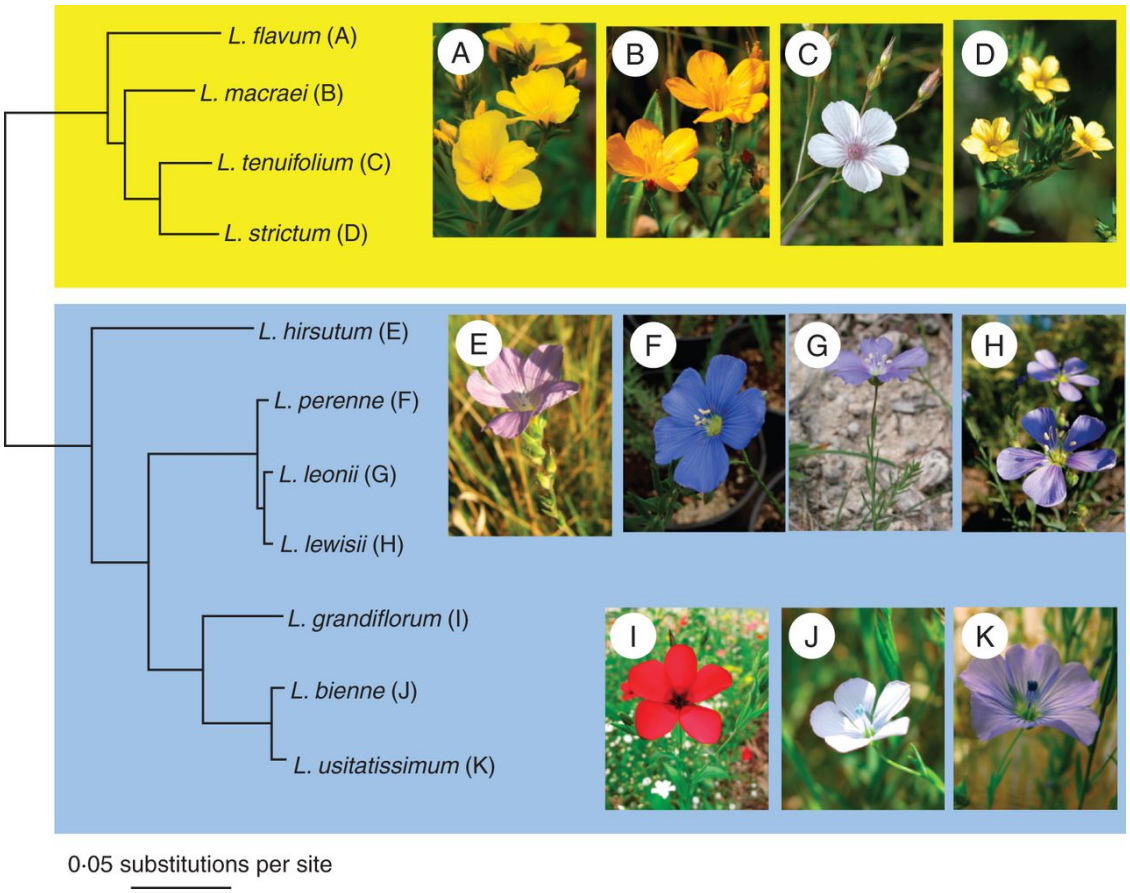


Figure 3.1 STAR phylogeny of the 11 *Linum* species constructed from 413 gene trees. Branch lengths were estimated with GARLI and all nodes on the tree have 100 % bootstrap support. The tree shows the two major clades of Linum species studied here and their dominant flower colour. The tree was rooted with *Bischofia javanica* (Phyllanthaceae; taxon not shown here).

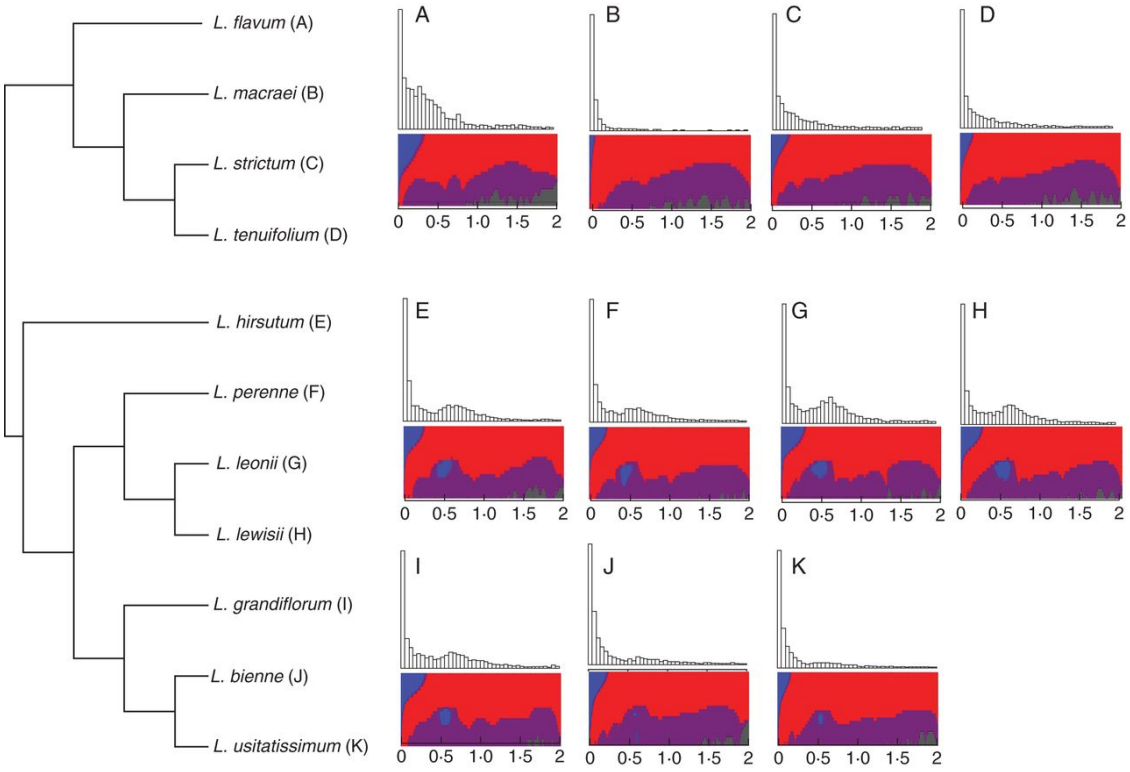


Figure 3.2 Cladogram, estimated with the STAR method, of the 11 *Linum* species (left), with their corresponding duplicate age distributions and SiZer plots (right). The SiZer plots are placed underneath each paralogue age distribution (the x-axis being Ks values). Different bandwidths used in the Gaussian smoothing of the Ks values are plotted on the y-axis of the SiZer plots and an optimal binning of the Ks values is plotted on the x-axes. These plots are composed of four colours: blue represents a significant increase in Ks value density, red represents a significant decrease in density, purple represents regions where there is no significant increase or decease in density, and grey areas represent insufficient data. Peaks in duplicate age distributions generated by paleopolyploidy events are characterized by the SiZer plots as blue areas flanked by red and purple areas, generally located in the middle of the y-axis. The position on the x-axis depends on the age of the duplication event. The blue areas around Ks 0.68 in all the SiZer plots of the blue-flowered Linum species represent statistical evidence supporting the occurrence of a polyploidy event.

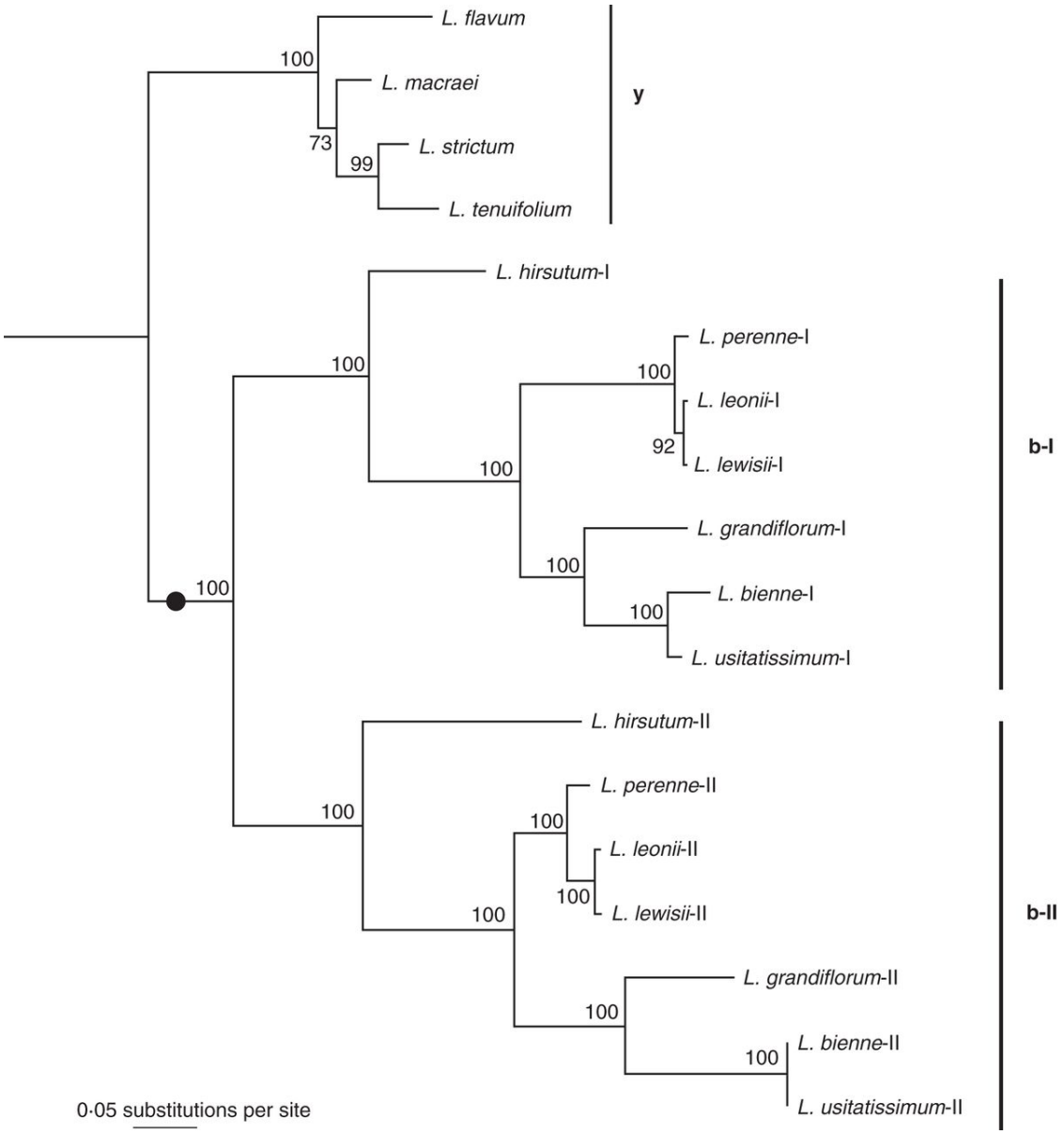


Figure 3.3 A phylogeny of orthologous groups that is consistent with a polyploidy event occurring on the branch leading to the blue-flowered *Linum* (black dot). The species relationship within each of the clades is consistent with the species phylogeny (Figure 3.1). The tree was rooted on its midpoint for visualization purposes and bootstrap support is shown near the nodes of the phylogeny. Y indicates the yellow-flowered clade, b indicates the blue-flowered 1 clades. Based on a BLASTX search on Phytozome (Goodstein et al.,

2012) using the L. usitatissimum contigs, this gene appears to be a Co-chaperone-like protein in the GrpE family. L. usitatissimum-I corresponds to the gene Lus10029654 and L. usitatissimum-II matches Lus1002803 in the genome assembly of cultivated flax (*L. usitatissimum*). The best hit of both paralogues in the *Arabidopsis thaliana* genome assembly is AT5G17710.

Species **Chromosome number**[**]	**Clade (Family)**	**Tissue sequenced**	**Number of reads (read length)**
Bischofia javanica NA	NA (Phyllanthaceae)	Young leaves	2.33E+07 (90 bp)
Linum flavum 2n = 28	Yellow (Linaceae)	Stem apex, shoot, leaves	2.42E+07 (90 bp)
Linum macraei NA	Yellow (Linaceae)	Stem apex, shoot, leaves	2.65E+07 (90 bp)
Linum strictum 2n = 18	Yellow (Linaceae)	Stem apex, shoot, leaves, flowers	2.76E+07 (90 bp)
Linum tenuifolium 2n = 18	Yellow (Linaceae)	Stem apex, shoot, leaves	2.79E+07 (90 bp)
Linum hirsutum 2n = 16	Blue (Linaceae)	Stem apex, shoot, leaves	2.97E+07 (90 bp)
Linum perenne[*] 2n = 18/36	Blue (Linaceae)	Stem apex, shoot, leaves, flower	5.62E+07 (90 bp)
Linum lewisii 2n = 18	Blue (Linaceae)	Stem apex, shoot, leaves	2.92E+07 (90 bp)
Linum leoni NA	Blue (Linaceae)	Stem apex, shoot, leaves	2.78E+07 (90 bp)
Linum grandiflorum 2n = 16	Blue (Linaceae)	Stem, flower buds, leaves, flowers	1.89E+07 (90 bp)
Linum bienne 2n = 30	Blue (Linaceae)	Stem apex, shoot, leaves, flowers,	2.28E+07 (90 bp)
Linum usitatissimum[*] 2n = 30	Blue (Linaceae)	Stem apex, stem	8.31E+07 (90 bp)

*Species with more than one library sequenced (two *L. perenne* and three *L. usitatissimum*)
**Chromosome numbers retrived from the Index to Plant Chromosome Numbers (IPCN)

Table 3.1 Number of reads acquired per species, in addition to tissue type info.

Group	Putative source of paralogy (comments)	N (frequency)
WGD* uninformative	More than one gene family	260 (33.9%)
-	Small scale duplication	239 (31.1%)
-	Unknown (strong phylogenetic conflict, likely due to multiple tandem duplications or assembly artefact)	62 (8.0%)
-	Unknown (too many missing taxa to determine pattern of paralogy)	26 (3.4%)
WGD* informative	WGD (strong phylogenetic pattern of paralogy consistent with a WGD event)	180 (23.6%)
Total		**767**

*Whole genome duplication (WGD)

Table 3.2 Summary of the patterns observed in paralogue phylogenies.

Chapter 4: Evidence for the origin of veiny pea (*Lathyrus venosus* Muhl. ex Willd., Fabaceae) by autopolyploidy

4.1 Introduction

The genus *Lathyrus* comprises around 160 herbaceous legume species (Kenicer, 2008). It is widely distributed in the temperate regions of the world, with its origin and main centre of diversity in the Mediterranean region (Schaefer et al., 2012). In North America the genus is also quite species rich, particularly in the west (Table 4.1). There are about 26 species endemic to the continent (Broich, 1989) and two more widespread native species also found in Eurasia (*L. palustris*, *L. japonicus*). Rather few species occur in eastern North America, but these include *L. venosus* and *L. ochroleucus*, which are widespread with partially overlapping distributions (Figure 4.1). They are somewhat similar vegetatively and can be confused when not in flower, although florally they are highly distinctive (Figure 4.2).

The species delimitation and phylogenetic relationships among some of the North-American taxa is still problematic, despite the application of both early experimental taxonomic methods (Senn, 1938) and later molecular techniques (Asmussen and Liston, 1998; Kenicer et al., 2005; Schaefer et al., 2012). These molecular studies have been based on extensive taxon sampling but a relatively small set of genomic regions, typically one or more plastid spacer regions and the internal transcribed spacer (ITS) region of ribosomal DNA.

Increased availability and lowered cost of massively parallel sequencing (MPS) have given new opportunities to the field of molecular systematics (Harrison and Kidner, 2011). It has enabled researchers to sequence entire plastid genomes, hereafter referred to as plastomes, and

ribosomal subunits of multiple species with relatively low cost (Straub et al., 2012). These regions are in high copy numbers in plant cells and can easily be assembled from low coverage, whole genome shotgun sequencing. This method has been called ultrabarcoding (Kane et al., 2012) or genome-skimming (Straub et al., 2012) and it has proved useful to identify, or to resolve phylogenetic relationships among closely related species (Eserman et al., 2013). Furthermore, this approach has been used to uncover the parents of an allopolyploid sunflower species (Bock et al., 2014). However, massively parallel sequencing is potentially challenging in *Lathyrus* because of the gigantic genome sizes common in the genus, which range up to 14 Gb in *Lathyrus vestitus* (Narayan, 1982; Bennett and Smith, 1991). *Lathyrus* therefore provides an interesting test of the ultrabarcoding approach in what have been called giga-genomes (Mackay et al., 2012).

Whole plastid genomes are particularly tractable for analysis, since their major genome organization and gene order is generally conserved, even among distantly related taxa (Palmer, 1991). However, exceptions to conserved plastome structures exist and one of the best known example is pea (*Pisum sativum*). Pea is a close relative of *Lathyrus* and its plastome has undergone numerous rearrangements (Palmer and Thompson, 1982). The grass pea plastome (*Lathyrus sativus*) is also known to have rearrangements (Magee et al., 2010) but these rearrangements have not been studied in any other species of *Lathyrus*. The sequencing of more whole plastomes in *Lathyrus* is therefore of some interest to assess the extent of structural variation in the plastid.

Polyploidy is remarkably rare in *Lathyrus* and is only known to occur in seven of the 160 described species (Broich, 1989; Chalup et al., 2012). Three of these species have both diploid and autopolyploid races (Broich, 1989; Khawaja et al., 1995; Khawaja et al., 1997). *Lathyrus*

venosus is a tetraploid (2n=4x=28), North-American species widely distributed in the northern part of the continent (Figure 4.1). It is the only *Lathyrus* species that has been suspected to be of hybrid origin, a hypothesis attributed to G.L. Stebbins by Gutiérrez et al. (1994). The Stebbins hypothesis suggests that *L. venosus* is an allopolyploid species, derived from a hybridization event between the marsh pea (*L. palustris*) and the creamy pea (*L. ochroleucus)*. Stebbins' hypothesis was based on both morphology and the current geographical distribution of the three species (J. F. Gutiérrez, personal communication). Gutiérrez et al. (1994) marshalled support for an allopolyploid origin of *L. venosus*, and specifically the Stebbins hypothesis, based on chromosomal C-banding and isozymes. However a more recent study suggests that *L. venosus* is more likely an autopolyploid species, due to the high frequency of multivalent chromosomes during meiosis (Khawaja et al., 1997). It is important to note that hybridization is generally thought to be extremely rare in *Lathyrus,* as natural hybrids have very rarely been reported and never been unequivocally substantiated. Furthermore, the production of hybrids by artificial interspecific crosses is extremely difficult (Senn, 1938; Hammett, 1989; Yunus and Jackson, 1991). Given these apparent barriers to hybridization in the genus it would be of especial interest if the hybrid origin of *L. venosus* could be substantiated. The aim of this study is therefore to clarify the origin of *L. venosus* using an ultrabarcoding approach, where low pass whole genome sequencing is used to compare *L. venosus* with its proposed parental species in addition to several other North-American *Lathyrus* species and several outgroup species.

4.2 Material and methods

4.2.1 Source of plant material and mapping data

The plant material for this study came from three sources. First we collected live plants in the field, transplanted into UBC's greenhouse facilities. Secondly we obtained seeds from a commercial provider, Roger Parsons Sweet Peas (Chichester, UK). Thirdly we received seeds from the USDA germplasm collection at Pullman, Washington (W6). All plants were grown in greenhouse facilities at UBC. In all cases where plants required critical determination they were grown until flowering, and herbarium voucher specimens were then collected (UBC). Further details are given in appendix B.1. Confirmation of the polyploidy of *L. venosus* versus the diploid status of other accessions was made using flow cytometry following the methods detailed in Dolezel et al. (2007), where *Vicia faba* (2C = 26.90) was used as the internal standard. As multiple cytotypes have been recorded in *L. palustris*, we additionally checked the diploid status of our material using chromosome counts. Root tips were stored for 24-26 hours at 4°C, in order to increase the metaphase index of dividing cells and fixed in 1:3 v/v glacial acetic acid and absolute ethanol for 2 h at room temperature. Squashes were preformed in 45% acetic acid and stained using a Feulgen solution (Feulgen and Rossenbeck, 1924). Metaphase cells were examined at 400X magnification using optical microscope. The maps (Figure 4.1) are based on records in The Global Biodiversity Information Facility (GBIF): http://data.gbif.org (accessed: October, 2013), search: *Lathyrus venosus* and North America/ *Lathyrus ochroleucus* and North America, multiple datasets. The records were manually inspected and dubious records omitted from the maps.

4.2.2 Construction of Illumina sequencing libraries

Total DNA was extracted from fresh leaf material using a modified version of the CTAB protocol (Doyle and Doyle 1987). We preformed RNase treatments following the suppliers protocol (cat. 19101, QIAGEN, Germantown, MD). DNA quality was assessed based on 260/280 nm ratios from NanoDrop measurements (ND-2000, Nano-Drop Technologies, DE) in combination with visual inspections on 1% agarose gels. Only high quality DNA was used in the Illumina library preparations, which were performed using the NEXTflex^{TM} DNA sequencing kit (100 bp Paired-End reads) (cat: 5140-02, Bioo Scientific Corp, TX). We followed the manufacturer's protocol and c. 400 bp DNA fragments were size selected using Agencourt AMPure Xp^{TM} magnetic beads (cat. A63880, Beckman Coulter Genomics, MA). Finished libraries were pooled and sequenced on a partial lane of the Illumina HiSeq-2000 platform.

4.2.3 Plastome assembly

Raw Illumina reads were quality trimmed using Trimmomatic v.0.3 (Lohse et al., 2012) using the following flags: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36. The plastome of each species was assembled using the following iterative approach. *De novo* assemblies of were generated from the quality trimmed reads, using the CLC Genomic Workbench v.6.5.1. Contigs containing plastid sequence were identified by BLAST search using the available grass pea plastome (*Lathyrus sativus)*, published in Magee et al. (2010) (GenBank: NC_014063.1). Nucleotides represented by Ns in the plastome assemblies were manually corrected by retrieving sequence information directly from the quality trimmed reads. When needed, multiple contigs containing plastid sequence were joined by eye using information from the quality trimmed reads. The quality of each plastome assembly was verified visually by

inspecting a BWA mem pileup, v. 0.7.5a (Li and Durbin, 2009), of the trimmed reads using Tablet v.1.13.12.17 (Milne et al., 2013). Finally all plastome assemblies were annotated using DOGMA (Wyman et al., 2004).

4.2.4 Assembly of ribosomal subunits

Two ribosomal tandem repeat units were assembled using an iterative approach: (i) the 45S rDNA repeat unit, containing the 18S, 5.8S and 26S rDNA subunits and (ii) the 5S rDNA repeat unit, containing a single rDNA unit in addition to a small spacer region. With a few exceptions (Galián et al., 2012) these two tandem repeat units are found as separate loci within the genome, as in legumes. For the assembly of these units, we used the recently developed pipeline, aTRAM (Automated Target Restricted Assembly Method, available at https://github.com/juliema/aTRAM), which takes advantage of the paired end feature of Illumina reads. In short, aTRAM preforms a BLAST search using a given reference sequence as its query in a custom blast database that contains paired end Illumina reads (Allen and Huang, pers.com.). aTRAM then extracts all the short reads that blast to the reference sequence and then reunites then with their mate pairs in order to perform *de novo* assembly. By default aTRAM uses Velvet (Zerbino and Birney, 2008) for *de novo* assembly and uses the longest Velvet contig as the reference in a new iteration. However, we used Trinity v. r2013-08-14 (Grabherr et al., 2011) as the *de novo* assembler, since we found it to work better for the 45S rDNA. As an initial reference sequence for the aTRAM assembly, we used the previously published (Kenicer et al., 2005) and publically available ITS sequence of *Lathyrus venosus* (GenBank: JX506154). It took about 10 aTRAM iterations to assemble the full 45S rDNA sequence, as well as some manual adjustments at both ends of the assembled contig. For the assembly of the small ribosomal subunit, we used

the published (Ellis et al., 1988) and publicly available 5S rDNA sequence (GenBank: AY499178) of pea (*Pisum sativum*). We checked the quality of our assembly by inspecting the BWA pileup of the trimmed reads (Li and Durbin, 2009) using Tablet (Milne et al., 2013). Finally we identified and coded ambiguous positions in our assembly using the UnifiedGenotyper in the GATK package v. 2.4.9 (McKenna et al., 2010), where the ploidy flag was set either set to 2 or 4 depending on the ploidy of the species.

An initial annotation of the 45S rDNA was performed using RNAmmmer v.1.2 Web Server (http://www.cbs.dtu.dk/services/RNAmmer/), described in Lagesen et al. (2007). Since RNAammmer did not identify the 5.8S rDNA gene, the external transcribed spacer (ETS) or internal transcribed spacer (ITS), we aligned some publically available sequences to annotate these regions (ETS, GenBank: Z92949, Bena et al., 1998; ITS and 5.8S, GenBank: JX506154, Schaefer et al., 2012). We adjusted the annotation of the 26S gene using previously published rDNA sequence from *Eucryphia lucida* (Kuzoff et al., 1998). Regions that fell outside the annotated genes and transcribed spacers were designated as IGS (Intergenic Spacer). For the small ribosomal subunit we based our annotation on the *P. sativum* 5S rDNA reference sequence (Ellis et al., 1988).

4.2.5 Interspecific plastome and 45S rDNA sequence variation

Interspecific variation in plastomes and the ribosomal subunits were investigated by mapping the quality trimmed reads of all *Lathyrus* species (i.e. excluding *Pisum sativum*) to reference sequences of *Lathyrus venosus* using BWA (Li and Durbin, 2009). Single nucleotide polymorphisms (SNPs) were characterized using the UnifiedGenotyper in the GATK package v. 2.4.9 (McKenna et al., 2010). We used the default settings with the exception of the ploidy flag,

which was set to 1 in the plastome SNP calling and either 2 or 4 in the 45S rDNA mapping, depending of the ploidy of the species. The SNP calls were then filtered with the VariantFiltration tool in the GATK, using the following filter expression: "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0". Variable positions were summed in 1000 bp and 100 bp bins for the plastome and 45S rDNA respectively, using a custom Python script which utilized the digitize and bincount functions in numpy v. 1.6.1 (http://www.numpy.org/). Interspecific variation in the plastome and 45S rDNA was visualized using Circos (Krzywinski et al., 2009).

4.2.6 Analysis of plastome rearrangements

Comparison of the *de novo* plastome assemblies from *Lathyrus* revealed extensive rearrangements, where large plastome segments in closely related species had been translocated and/or inverted. In order to describe these rearrangements and to compare the plastome structure among species, we split the coding region into 22 gene blocks. Each block was defined as a gene or a set of genes that were consistently in the same within-block order among all species. We frequently observed inversions of whole gene blocks with the component gene set presented in the reverse order. The gene blocks were manually defined by inspecting GenomeVx (Conant and Wolfe, 2008) visualization of annotated plastomes by eye and comparing them among species. A more complete gene block analysis, using an expanded dataset, in given in Chapter 6.

4.2.7 Phylogenetic analysis

Multiple sequence alignments of the 5S – and 45S rDNA sequences were generated with MAFFT v.705b (Katoh and Standley, 2013). The MAFFT alignments were manually adjusted

using BioEdit (Hall, 1999). Due to the extensive rearrangements observed in the plastomes (see previous paragraph), we restrict our plastome phylogenetic analysis to protein coding genes. We developed a simple phylogenetic pipeline that extracts gene coding regions from DOGMA annotated plastomes, aligns individual gene with MAFFT, trims gaps using trimAl v.1.2 (-auto flag) (Capella-Gutiérrez et al., 2009) and generates a concatenated alignment of all genes. The pipeline, Plast2phy, written in Python, is available at https://github.com/saemi/plast2phy. Models of base substitution were tested using jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012). Using the Akaike information criterion (AIC), we determined the GTR+G+I model optimal for the concatenated plastome alignment and the 45S rDNA dataset while the HKY + I model was optimal for the 5S rDNA matrix. We analyzed all datasets with maximum likelihood (ML; Felsenstein, 1973) using GARLI (Zwickl, 2006) and under a Bayesian framework using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). We ran GARLI v. 2.0 with default settings, using ten independent searches and 100 bootstrap replicates. Bootstrap consensus was calculated using SumTrees v. 3.3.1 in the DendroPy package (Sukumaran and Holder, 2010). The Bayesian analyses were performed on a multithreaded version of MrBayes v.3.2.2, using 4 separate runs and 10 million generations. A default burnin of 25% was used and the log files inspected using Tracer (http://tree.bio.ed.ac.uk/software/tracer/) to verify that runs had reached equilibrium. Trees from phylogenetic analysis were drawn using FigTree v.1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/), rooted with *Pisum sativum* and provided with bootstrap values were using Inkscape (http://inkscape.org/).

Phylogenetic hypothesis testing was performed using the Shimodaira-Hasegawa (SH) (Shimodaira and Hasegawa, 1999; Goldman et al., 2000) test implemented in the program package CONSEL (Shimodaira, 2001). GARLI was used to generate the sitewise log-likelihood

values for each individual site, which is the input for CONSEL. The sister species relationship of *Lathyrus venosus, L. ochroleucus* and *L. graminifolius* were tested, due to observed phylogenetic incongruences and our interest in the polyploid origin of *L. venosus*. We tested the monophyly of *L. ochroleucus* and *L. venosus* as well as the monophyly of *L. graminifolius* and *L.venosus* in both datasets.

4.2.8 Examination of nucleotide additivity in the 45S rDNA sequence of *L. venosus*

Infra-individual site polymorphisms (2ISPs; Potts et al. 2014) can result from the persistence of polymorphisms resulting from lineage-specific mutation, or from nucleotide additivity resulting from hybridisation. If polymorphisms are a combination of fixed differences between putative parental species (nucleotide additivity), hybridization is a likely explanation, and this has been used in several studies to detect hybidization and suggest parentage (e.g. Sang et al., 1995; Whittall et al., 2000; Bock et al., 2014). Traditionally nucleotide additivity was determined using Sanger sequencing (Sang et al., 1995; Whittall et al., 2000), but more recently is has also been inferred using genome skimming methods (Bock et al., 2014). We examined the 45S rDNA sequences for ambiguous nucleotide calls, where the two bases are characteristic of separate species and hence evidence of nucleotide additivity and hybridization. Ambiguous nucleotide calls were obtained from the GATK UnifiedGenotyper, previously described in the 'Interspecific plastome and 45S rDNA sequence variation' section in the methods above. Each ambiguous call was verified by manually inspecting the alignment of the short reads to the 45S rDNA reference sequence using Tablet (Milne et al., 2013). Ambiguous positions covered by reads with two alleles present in a roughly 50/50 ratio, were coded using the IUPAC ambiguity codes.

4.2.9 Extraction of exomic nuclear data

Despite the overall low coverage of the genome, we determined that some genic regions had high enough coverage to be mappable to a reference transcriptome of *L. odoratus*. We used a coverage threshold of x8 for the mapping to ensure reliable base calling. The islands of higher coverage may result partly from chance as these regions differed in different libraries. However, systematic bias in Illumina sequencing methodology may also paly a role. Whatever the reason, we could reliably call SNPs in these ≥ x8 coverage fragments and consequently devised a strategy for interrogating the nuclear genome in order to provide a contrasting exomic dataset for comparison to the plastome and rDNA. The reference exomic contigs for *L. odoratus* were generated in the following fashion. Total RNA was extracted from immature floral tissue of *L. odoratus* cv. Cupani ("wild type"). Flower buds were collected in the greenhouse and snap frozen in liquid nitrogen. Standard petals were removed from the bud and total RNA extracted using Plant RNA Reagent (Invitrogen). The RNA extraction was washed twice with Turbo DNase (Ambion). Library construction and Illumina sequencing was performed by Cofactor Genomics (St. Louis, MO, http://www.cofactorgenomics.com/). Sequencing libraries were constructed using cDNA which had been enriched for nuclear genes using PolyA selection. The library was then normalized and finally sequenced on an Illumina Genome Analyzer II. The sequencing yielded about 45 million 60-bp paired end reads. The reads were trimmed for quality prior to assembly, where bases below quality of 20 were removed from the ends of each read. Trimmed reads were assembled in a *de novo* fashion using Trinity v. r2013-08-14 (Grabherr et al., 2011). Contigs smaller than 300 bp were removed from the assembly and contigs with an overall similarity of 99% and greater were clustered together using CD-HIT-EST program in the CD-HIT package v.4-6 (cd-hit-est flags: -c 0.99 -l 299 -d 0) (Li et al., 2001; Li and Godzik,

2006). After clustering, there were 47,988 contigs with a combined length of 34,030,315 bp. These contigs were used as reference sequences for short read mapping (see next section).

4.2.10 Read depth analysis of the exomic region in whole genome shotgun sequencing data

We investigated the patterns of read depth in the exomic region of *L. odoratus* by mapping of the reads from the low coverage whole genome shotgun sequencing. We sequenced the same cultivar of *L. odoratus* as was used for the generation of the transcriptome with a raw read depth of 0.144X. Raw read depth is calculated by dividing the number of sequenced bases by haploid genome size (Bennett and Leitch, 2012). We also attempted to assess the genome size of *L. odoratus,* using the observed k-mer frequency estimated using Jellyfish v. 2.1.3 (Marçais and Kingsford, 2011). Those results were consistent with a large genome size of *L. odoratus*. The reads were then mapped to exomic contigs using the bwa mem algorithm v. 0.7.5a (Li and Durbin, 2009) with the programs default settings. The output alignments were sorted, duplicate reads marked using Picard (http://picard.sourceforge.net) and coverage per reference nucleotide extracted using the genomecov program in the bedtools package (Quinlan and Hall, 2010). We compared the observed frequency distribution of coverages using various functions in the numpy Python package (http://www.numpy.org/) and R (R Core Team, 2014). The null-expectation is that the observed coverage from mapped reads should be randomly distributed following a Poisson distribution with a lambda (i.e. mean and variance) equal to the raw read depth (http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf). However, we determined that the observed frequency distribution of read depths in the exomic region of *L. odoratus* clearly did not conform to a random Poisson distribution, as had a greater proportion of highly covered sites than expected (see Table 4.4).

4.2.11 Mining low coverage whole genome shotgun data of exomic SNPs

Due to the unexpectedly high frequency of relatively highly covered positions in the exomic region of *L. odoratus* (see previous section), we decided to extract single nucleotide polymorphisms from our *Lathyrus* species of interest. Quality trimmed reads from each sequencing library were mapped to the *L. odoratus* exomic contigs using the bwa mem algorithm with the default settings v. 0.7.5a (Li and Durbin, 2009). The output alignments were sorted and duplicate reads marked using Picard (http://picard.sourceforge.net). Variants were called using the UnifiedGenotyper, which is a part of the GATK package v. 2.4.9 (McKenna et al., 2010). We used the default settings, except the ploidy flag which was either set to 2 or 4, depending of the ploidy of the species. We coded heterozygous sites using the appropriate IUPAC amibiguity codes. Low quality variant calls were removed using the VariantFiltration tool in GATK, with the following filter expression: "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 173.0 || MappingQualityRankSum < -12.5 || ReadPosRankSum < -8.0". The genomecov program in the bedtools package (Quinlan and Hall, 2010) was used to retrieve coverage information for site in the exomic contigs. We removed sites which were covered less than eight times and where the UnifiedGenotyper called three or more alleles in a single position. The reference base from the *L. odoratus* reference position was used for sites without a variant call and coverage of at least eight. Sites passing these filters were used in the subsequent analysis.

4.2.12 Estimation of genetic distances from exomic SNPs

Pairwise genetic distances were estimated using PLINK v1.07 (Purcell et al., 2007). We used custom Python scripts to generate the PLINK input file, which excluded all non-biallelic markers. We used a minimum allele frequency of 4% (--maf 0.04 flag) and the maximum

amount of per-SNP missing data was set to 10% (--geno 0.10). A matrix of pairwise genetic distances was generated using the following PLINK commands: --cluster --distance-matrix. The distance calculated by PLINK is standard p-distance, where the number of alleles that differ among a pair of individuals are summed and divided by two times (for diploids) the number of loci. This scales the distance to a number from 0 to 1, where genetically similar individuals have a number close to 0. For the calculation of the standard p-distance, sites from the tetraploid *L. venosus* were coded, where possible, to a diploid format, (i.e. AATT to AT and AAAA to AA). Sites that could not be coded in that fashion were excluded (i.e. AAAT or ATTT). In order to amplify relatedness between hybrids and their parents we also calculated a modified p-distance in a pairwise manner. Here the total number of loci that shared an allele was summed up for each pair of individuals and divided by the number of loci. Both distance measures were used in Principal Coordinates Analysis (PCoA) as a convenient way of displaying the patterns of genetic relationships between species. The multivariate analyses were performed using software Ginkgo in the VegAna package (Bouxin, 2005).

4.2.13 Phylogenetic inference based on exomic SNPs

We used SNAPP (Bryant et al., 2012) to infer phylogenetic relationships based on the SNPs that were extracted from exomic regions (see previous section). SNAPP implements a full coalescent model in a Bayesian framework to estimate species trees from biallelic SNPs. We used the program's default parameters, which included a chain length of 10 million generations and tree sampling every 1,000 generation resulting in 10,000 trees. The burnin was set to 15% based on an inspection of SNAPP's log file using Tracer v.1.6 (Rambaut and Drummond, 2009). Frequencies of observed tree topologies were summarized using treesetanalyser and species

relationships visualized using DensiTree v. 2.0.1 (Bouckaert, 2010). Both treestatanalyzer and DensiTree are part of the SNAPP package. We also preformed a network analysis of the exomic SNPs using SplitsTree v.4.11.3 (Huson and Bryant, 2006).

4.2.14 Dating the divergence of *L. ochroleucus* and *L. venosus*

In order to estimate the divergence time of *Lathyrus venosus* and *L. ochroleucus* we calculated the synonymous substitution ratio (Ks), i.e. the numbers of silent mutations per silent sites, among the two plastomes. Using custom Python scripts, we extracted protein-coding regions from the DOGMA annotated plastomes, aligned the amino acid sequences using MAFFT, converted the amino acid alignment to its corresponding coding sequences using RevTrans (Wernersson and Pedersen, 2003), generated a concatenated alignment of all plastid genes and calculated the Ks ratio using MEGA v. 6.0 (Tamura et al., 2013). Divergence times were then estimated using previously estimated plastome mutation rates for herbaceous species, which range from 1.1 - 2.9 silent substitutions per billion (10^9) years (Wolfe et al., 1987).

4.3 Results

4.3.1 Illumina sequencing, de novo assembly and depth of coverage

We retrieved about 11 – 20 million high quality reads per library (Table 4.2). That translates to around 1-2 Gbp of sequence and about 0.09 – 0.2x average nuclear genome coverage, based on our own estimates of genome sizes. However the plastome and rDNA regions we used for phylogenetic reconstruction are in high copy number and had about three orders of magnitude higher read depth. The assembled plastomes ranged from 120-126 kb in size and their read depth varied from 110x to 380x (GenBank: KJ806192-KJ806203). The large ribosomal subunit (45S),

about 8 kb long, had a read depth of 145 – 723x and the small ribosomal subunit (5S), about 280bp in length, was covered about 162 – 3,072 times (Table 4.2). We were unable to assemble the 5S rDNA sequence from *Lathyrus sativus*. Even though the overall coverage of the nuclear genome in our sequencing was low (ca. 0.1x) we found that coverage of the nuclear genome was uneven and consequently a surprising amount of putatively single copy nuclear sequence could be reliably retrieved. These results are given in more detail below.

4.3.2 Interspecific variation of plastomes and rDNA in *Lathyrus*

Sequence variation in plastomes and ribosomal subunits among *Lathyrus* species was investigated by short read mapping and SNP calling, where *L. venosus* was used as a reference (Figure 4.3) The larger ribosomal repeat unit (45S) was quite variable, in particular the transcribed and intergenic spacers (ETS, ITS and IGS) (Figure 4.3A.). The *Lathyrus* 5S rDNA is composed of a 91 bp ribosomal gene and a 188 bp spacer region species. No variable sites were identified within the 5S ribosomal gene but a total of 10 SNPs were characterized in the spacer region. Furthermore we discovered a 50 bp deletion in the *Lathyrus* 5S rDNA spacer region, compared to the pea (*Pisum sativum*) reference sequence (AY499178). The 45S rDNA contains three ribosomal genes, which are very conserved among these closely related species (Figure 4.3A.). The 26S rDNA was the most variable ribosomal gene, 0.010 SNPs/site followed by the 5.8S and 18S rDNA, both with 0.006 SNPs/site. This is not the case for the 45S rDNA spacer regions, which contain a substantial amount of sequence variation. The intergenic spacer (IGS) and external transcribed spacers (ETS) were the most variable regions within the large ribosomal subunit, with a variability of 0.10 SNPs/site and 0.16 SNPs/site respectively. The two internal

transcribed spacers (ITS1 and ITS2) were moderately variable, with an average of 0.086 SNPs/site.

We identified a total of 5,845 variable positions in the plastomes of analyzed *Lathyrus* species. Roughly half of the sequence variability of the *Lathyrus* plastomes lies within intergenic regions (Figure 4.3B). However the intergenic region is about 20 kb smaller than the coding region, which makes the spacer region as a whole more variable on average. Most plastid genes are either highly conserved or moderately variable, with variability ranging from 0 to 0.174 variants/site. The most variable protein coding gene was *clpP*, with 0.174 variants/site. Other variable genes were *ycf1* and *accD* with 0.133 and 0.123 variants/site respectively. We also examined the intraspecific plastome variation within *L. japonicus* and *L. ochroleucus* and found very little variation. A total of 89 variable positions were identified in *L. japonicus* and 7 in *L. ochroleucus.*

4.3.3 Plastome rearrangements within *Lathyrus*

When we compared the *de novo* assembled plastomes from *Pisum sativum* and the *Lathyrus* species, major differences in gene order were observed. To understand and visualize these differences, we split the plastome into 22 gene blocks, where each gene block consists of genes that are in the same order in all species (Figure 4.4 and Table 4.3). Figure 4.4 shows that the order of these 22 gene blocks are highly variable within *Lathyrus*. Despite this structural variability, some closely related species have the same gene order. We therefore defined major plastotypes (MPt), where plastomes with identical gene order were designated the same MPt (Figure 4.4). An example is the MPt Lathyrus-01, which is shared by four related species: *Lathyrus venosus, L. ochroleucus, L. littoralis* and *L. japonicus.* A total of six MPt were

identified, five in *Lathyrus* (Figure 4.4A-D) and one for *P. sativum* (Figure 4.4A). Each of the MPt observed in *Lathyrus* was compared to the Lathyrus-01 (Figure 4.4). That comparison reveals that the differences among MPts within *Lathyrus* can be explained by one or more translocation and/or inversion events, represented by coloured rectangles and arrows. Furthermore, the numbers of these events reflect the phylogenetic distances among species (see Figure 4.5). A practical consequence of these major plastome structural differences within *Lathyrus*, is that identifying homologous intragenic sequences among species becomes challenging. We therefore decided to restrict plastid phylogenetic inference to protein coding genes, where an inference of primary homology is straightforward and can be made from gene annotation (see next section).

4.3.4 Phylogenetic relationships among North-American *Lathyrus*

We performed phylogenetic analysis separately on the plastid, 45S - and 5S rDNA datasets using maximum likelihood (ML) and Bayesian methods. The Bayesian analysis gave identical trees as the ML and similar support values. The phylogenetic tree based on 5S rDNA phylogeny is shown in appendix B.2 and will not be discussed further due to missing data and insufficient phylogenetic resolution. When we analyzed the full 45S rDNA alignment, we retrieved trees with overall low bootstrap support and numerous polytomies. We decided to exclude the IGS (intergenic spacer) from our alignment, due to its extremely high variability (Figure 4.3A) and probable homoplasy. This drastically improved the resolution and support values of our tree. For the plastid data, we decided to exclude all non-coding regions from the phylogenetic inference due to the extensive rearrangements in the *Lathyrus* plastomes (see previous section).

Both 45S rDNA and plastid data place *L. sativus* and *L. pubescens* outside the East-Asian (*L. davidii*) and the North-American *Lathyrus* clade, which was expected. However there is an incongruence in the placement of these taxa between the two phylogenies (Figure 4.5). The 45S rDNA data is furthermore unable to unequivocally resolve the relationship of *L. davidii* and the N-American *Lathyrus*, resulting in a polytomy (Figure 4.5A). The plastid data however places *L. davidii* outside the North-American taxa with 100% bootstrap support (Figure 4.2B). Within North-American *Lathyrus*, both plastid and 45S rDNA datasets, split the remaining species into two clades: (1) *Lathyrus japonicus* and *L. littoralis* and (2) *L. palustris, L. graminifolius, L. ochroleucus* and *L. venosus.* Within clade 2 there is another incongruence, where *Lathyrus ochroleucus* is sister to *L. venosus* in the plastid data but *L. graminifolius* groups with *L. venosus* in the 45S rDNA with low bootstrap support (<50%). The true phylogenetic relationship of *L. venosus* was of particular interest to the authors of this paper and therefore we decided to investigate the significance of this incongruence further.

The statistical significance of the *L. ochroleucus* and *L. venosus* incongruence among the plastid and 45S rDNA tree topologies was tested using a Shimodaira-Hasegawa (SH) test. We ran maximum likelihood analyses, where the topology of North-American *Lathyrus* from the best 45S rDNA tree was enforced on the plastid dataset, and vice versa. Log-likelihood ratio tests showed that when the plastid phylogenetic inference is constrained with the 45S rDNA topology, it results in a tree with a significantly worse log-likelihood ($p < 0.001$). This is not the case for the 45S rDNA phylogeny. There we find no statistical difference among phylogenies inferred with the plastid topology of North-American *Lathyrus* enforced compared to an unconstrained run ($p > 0.05$). These results demonstrate that the incongruence among the plastid and 45S rDNA topologies is insignificant and probably only due to poor coalescence among the rDNA

sequences. The phylogenetic evidence is consistent with the suggestion that *L. ochroleucus* and *L. venosus* are true sister species, as seen in the plastid phylogeny (Figure 4.5B). The sister species relationship among these species is furthermore supported by the Bayesian multispecies coalescent tree, generated by *BEAST from the plastid and 45S rDNA datasets (appendix B.3).

4.3.5 Lack of nucleotide additivity in *L. venosus*

We found very little evidence of any nucleotide additivity in the 45S rDNA sequence of *L. venosus*. Only a single position in the entire alignment was determined to be of that nature. That SNP position was coded as R in *L. venosus*, representing A/G, where the G allele was present in *L. graminifolius, L. palustris* and *L. ochroleucus.* The A allele was present in *L. japonicus* and *L. littoralis*. This SNP position is therefore likely to be homoplastic in nature. Eight additional heterozygous positions were determined in the *L. venosus* 45S rDNA alignment, but none of these was additive in nature.

4.3.6 Frequency distribution of read depths in the exomic region of *L. odoratus*

We discovered an unexpected pattern of read depths when *Lathyrus odoratus* whole genome reads were mapped to a transcriptome of *L. odoratus* (Table 4.4). The read depth per base in a whole genome shotgun sequencing experiment is expected to follow a Poisson distribution if the reads are randomly distributed over the genome. However we observed far more highly-covered positions than would be expected from random sequencing. For example there were about 1.9 million positions covered 8 or more times, where the null-expectation is 0 (Table 4.4), indicating that there is a very strong selective bias in Illumina sequencing. This biased coverage in the exomic region was also detected in other sequenced libraries. The non-random distribution of

sequence coverage proved extremely useful for nuclear SNP discovery in the studied *Lathyrus* species (see next section).

4.3.7 Exomic divergence between *L. venosus* and other species

A total of 53,869 exomic positions were extracted from the mapping of reads from WGS sequencing to the exomic region of *L. odoratus* and used for estimation. These sites had coverage of at least 8x. After removing invariable sites (minor allele frequency < 0.04) and sites with more than 10% missing taxa, we were left with 23,159 positions. We used these loci to estimate the similarity of the nuclear genome among our species of interest. In particular, we wanted to see whether we would find any evidence supporting the allopolyploid origin of *L. venosus* through the hybridization of *L. ochroleucus* and *L. palustris*, previously referred to as the "Stebbins hypothesis". If the Stebbins hypothesis is correct, *L. venosus* should have a smaller genetic distance to its proposed parental species compared to other North-American species. Both distance methods that were employed, standard p-distance and modified p-distance, show a strong similarity between *L. venosus* and *L. ochroleucus* (Figure 4.6) but *L. palustris* is no closer to *L. venosus* than *L. venosus* is to other-North American species.

4.3.8 Phylogenetic inference using exomic SNPs

The phylogeny of the North-American *Lathyrus* inferred from the exomic SNPs using SNAPP (Figure 4.7) was largely congruent with the plastid tree (Figure 4.5B), with one exception. In the SNAPP phylogeny, *L. davidii* and *L. pubescens* group together as sister lineages (appendix B.4), whereas *L. davidii* is sister to the North-American *Lathyrus* in the plastid tree. Figure 4.7A shows a densitree representation of the three most frequently observed tree topologies, each

shown in a different color (Figure 4.7A-D). The differences among the top three tree topologies are mostly caused by the unstable positions of *L. palustris* and *L. graminifolius*. The most likely cause for observed incongruences is differences in the coalescence among the nuclear markers. In all post-burnin trees produced by SNAPP, *L. venosus* and *L. ochroleucus* come out as sister species. There is no signal suggesting an alternative relationship, between *L. venosus* and another species (e.g. *L. palustris*) coming from part of the genome. The results from the network analysis are in agreement with the SNAPP analysis (Figure 4.8).

4.3.9 The divergence time of *Lathyrus ochroleucus* and *L. venosus*

The divergence time of *Lathyrus ochroleucus* and *L. venosus* was estimated from the synonymous substitution ratio (Ks) of plastid genes. We extracted the coding sequence from the two sister species, aligned them and estimated the number of silent substitutions per silent sites (Ks) and used published mutation rates for plastid genes (1.1 - 2.9 /10^{-9} year, Wolfe et al., 1987) to estimate the divergence time. We extracted a total of 75 protein coding genes from the plastome assemblies of the two species, which had a total of 21,060 codon positions (63,180 bp). The Ks value was estimated to be 0.004 indicating that *L. ochroleucus* and *L. venosus* diverged around 0.7 – 1.8 million years ago.

4.4 Discussion

4.4.1 Polyploid origin of *Lathyrus venosus*

Two contrasting hypotheses have been put forward for the polyploid origin of *L. venosus*: (1) allopolyploid speciation, from a hybridization event between *L. palustris* and *L. ochroleucus* (Gutiérrez et al., 1994), and (2) simple autopolyploidy (Khawaja et al., 1997). Our data supports

the autopolyploidy of *L. venosus,* since no marked incongruence was found between the exomic, nuclear ribosomal and plastid datasets, and there is no evidence of a dual signal of *L. venosus* relationship coming from the ribosomal DNA or the exomic snps. These results fit well with what is known about the biology of *Lathyrus,* notably the tendency of *Lathyrus* species to produce infraspecific autopolyploid races (Broich, 1989; Khawaja et al., 1995) and the rarity of hybridization in the genus as a whole (Hammett, 1989; Yunus and Jackson, 1991). We therefore conclude that *L. venosus* is an autopolyploid species, closely related to *L. ochroleucus,* and that these species diverged some 0.7 – 1.8 million years ago. The close relationship between *L. venosus* and *L. ochroleucus* fits well with the morphology of two species, previous phylogenetic work (Schaefer et al., 2012) and current geographical distribution (Figure 4.1). Autopolyploidy is a "Cinderella" of plant evolution: often underestimated and downplayed in importance (Soltis et al, 2007). In *Lathyrus venosus* we have an example of an autopolyploid species that is as widespread and ecologically successful as its diploid sister taxon.

4.4.2 Plastome rearrangements within *Lathyrus*

The complete plastome sequence of the studied *Lathyrus* species proved very useful for the phylogenetic reconstruction of investigated taxa (see previous section). However the extent of the major plastome rearrangements noticed in *Lathyrus* (see Figure 4.4) were somewhat unexpected. Rearrrangments in the plastid genome of *L. nissolia* had been previously hypothesized (Asmussen and Liston, 1998). Land plant plastid genomes generally have a conserved genome organization (Palmer, 1991; Raubeson and Jansen, 2005), although exceptions are known (e.g. Weng et al., 2014). Two relatives of *Lathyrus, Pisum sativum* and *Trifolium subterraneum* (both in the same major group: the inverted repeat lacking clade, IRLC)

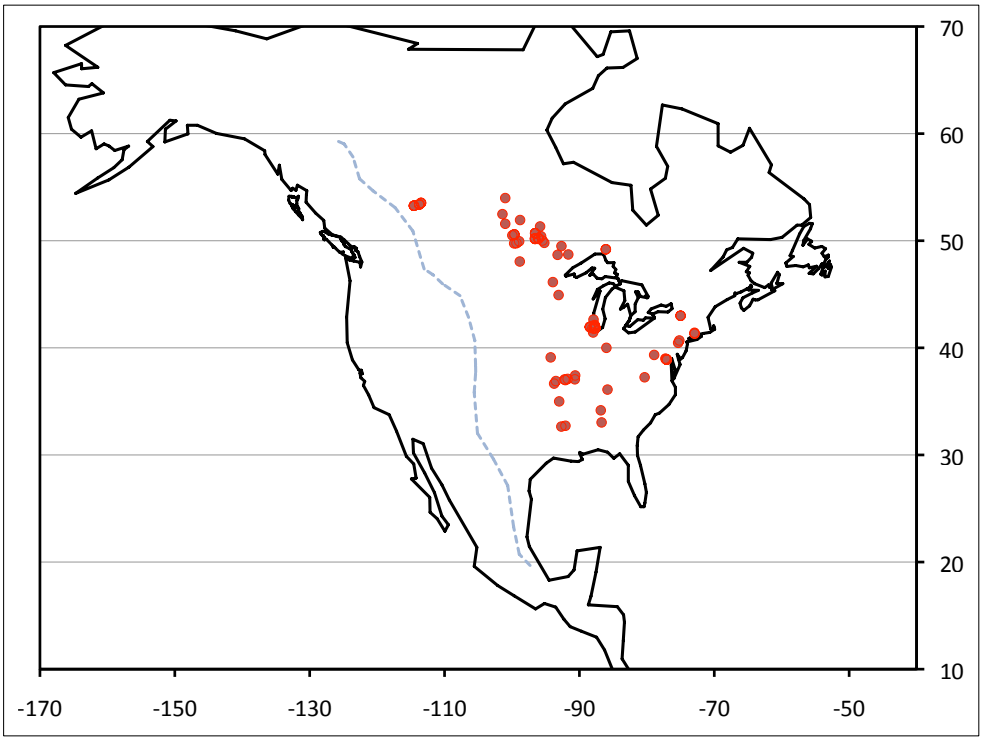
are also known to have highly rearranged plastomes (Palmer and Thompson, 1982; Cai et al., 2008). However the complexity of plastid genome rearrangements in a single genus, *Lathyrus*, is remarkable and indicates the likely magnitude of this phenomenon in the Fabeae (investigated further in Chapter 6). However, despite the rearrangements, we noted gene blocks that were consistently held together and never broken up by transposition or inversion. These blocks likely represent co-regulated units, i.e. the fundamental operon structure of the plastome (Sugita and Sugiura 1996; Ghulam et al., 2013; Stoppel and Meurer, 2013).

4.4.3 Phylogenetic potential of low depth whole genome sequencing for North-American *Lathyrus*

The North American members of section Orobus include numerous closely related taxa that result from a recent radiation. Some issues of species delimitation and species relationships are not yet fully resolved, although there have been a number of broad-scale phylogenetic surveys of the genus (Asmussen and Liston, 1998; Kenicer et al., 2005; Schaefer et al., 2012). Our study demonstrates that despite limited taxa sampling, it is possible to retrieve a well-supported phylogeny of these closely related species using whole plastome data (Figure 4.5B). With decreased costs of next generation sequencing and new methods that enable researchers to sequence tissue collected from herbarium specimens and assemble whole plastomes (Stull et al., 2013), it is clear that whole genome sequencing has great promise for providing a very detailed resolution of species relationships in this, or indeed any, genus. One problem is that some plastid intergenic spacer regions may be problematic for comparative sequencing due to the very extensive gene order rearrangements in this group (see previous section). Our solution is to use

only plastid coding regions for phylogenetics and the phylogenetic pipeline we developed for this purpose, plast2phy, may be of some general utility.

a)


70
60
50
40
30
20
10
-170
-150
-130
-110
-90
-70
-50

b)

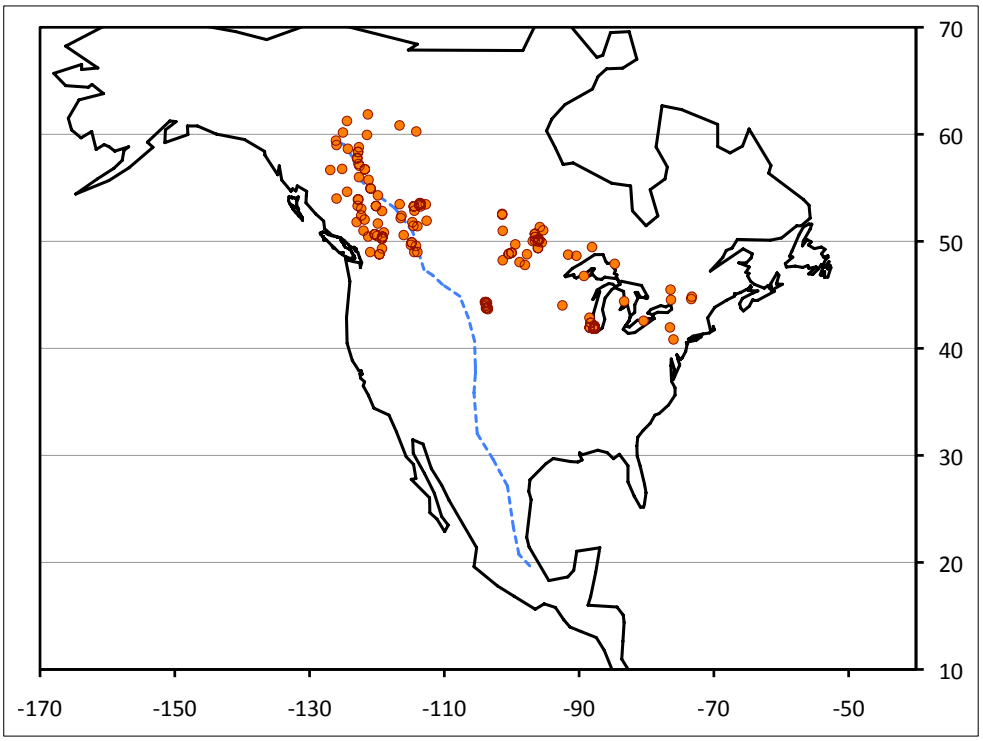

70
60
50
40
30
20
10
-170
-150
-130
-110
-90
-70
-50

Figure 4.1 Maps of the distribution of *Lathyrus ochroleucus* and *L. venosus*. (A) locations of representative herbarium specimens of *L. venosus*, representing the main distributional area of the species. *Lathyrus venosus* is a predominantly eastern species that does not occur west of the rocky mountains. The dashed line marks the boundary of the mountain west (the eastern edge of the western cordilleras). (B) Map showing selected herbarium records of *L. ochroleucus*, indicating the main distributional area of the species. *L. ochroleucus* has a transcontinental-northern distribution, and in western Canada it occurs on both sides of the Rocky Mountains.

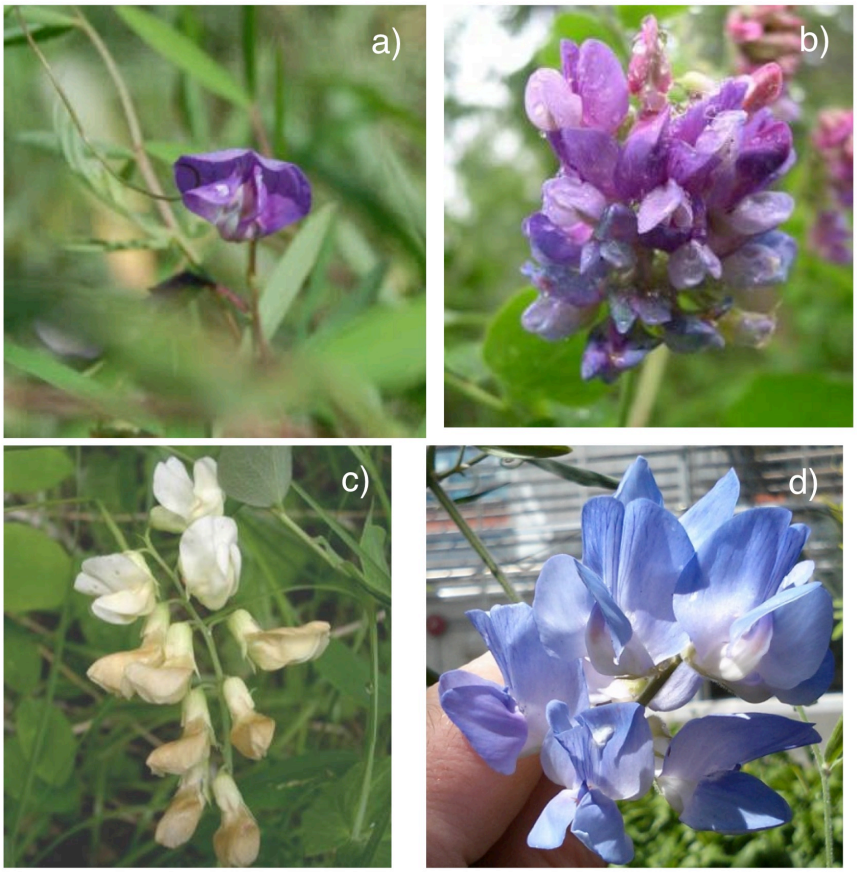


Figure 4.2 Photographs of some of the studied *Lathyrus* species: a) *Lathyrus palustris,* b) *L. venosus*, c) *L. ochroleucus* and d) *L. pubescens.*

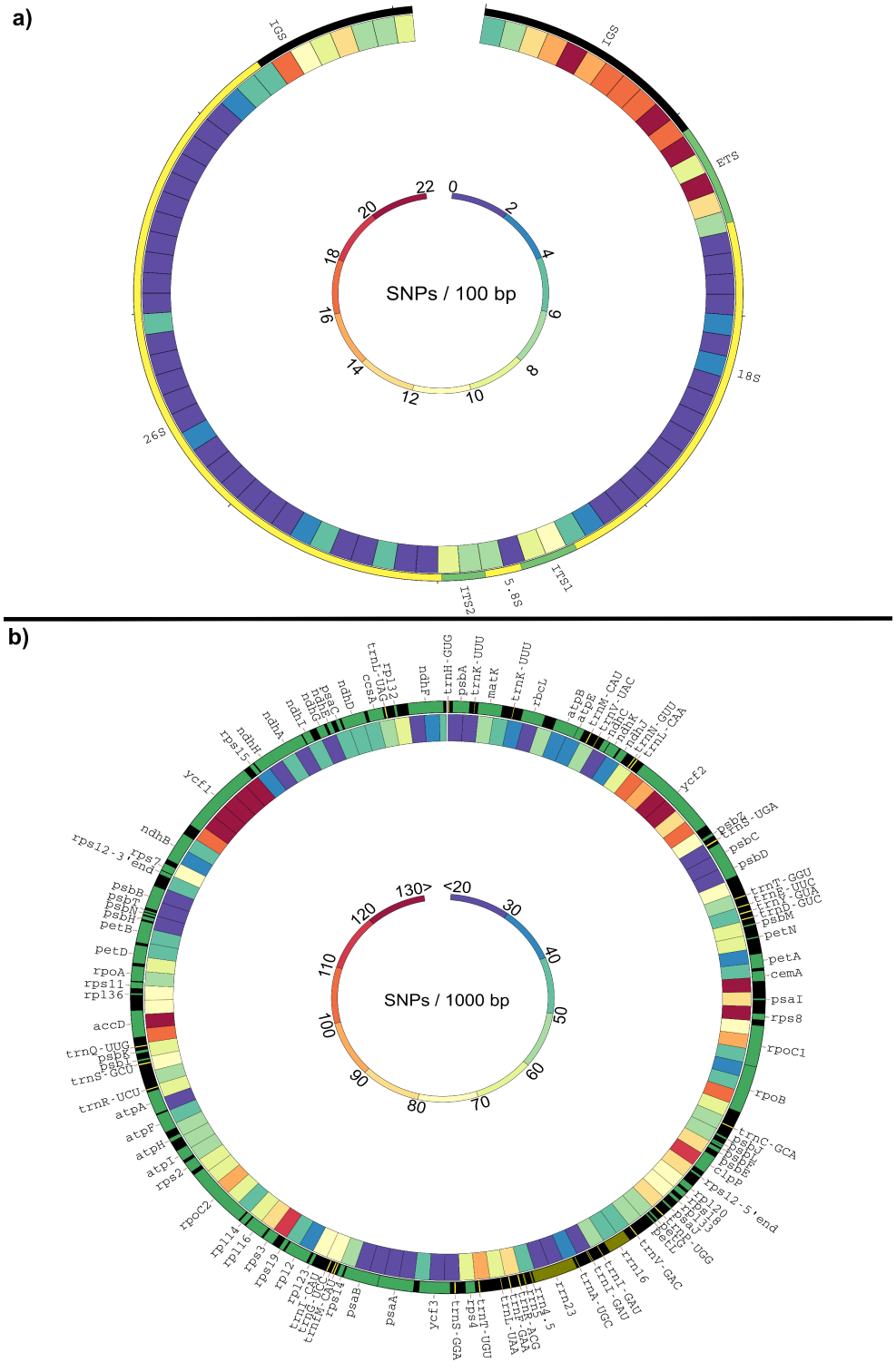

a)
IGS
ETS
18S
5.8S
ITS1
ITS2
26S
IGS
SNPs / 100 bp
0
2
4
6
8
10
12
14
16
18
20
22
b)
SNPs / 1000 bp
<20
30
40
50
60
70
80
90
100
110
120
130>

Figure 4.3 Circular representations of sequence variability among *Lathyrus* species within the large ribosomal subunit (A) and the plastome (B). Number of variable sites were binned into 100 bp (A) or 1,000 bp (B) blocks and variability visualized on a heatmap. The number of variable sites that each colour represents is shown in the small circular legend at the centre of each figure. (A) The ribosomal genes (18S, 5.8S and 26S) are in yellow, the intragenic – and external transcribed spacers (ITS and ETS) are in light green and the intergenic region is shown in black. (B) Protein coding genes are shown in dark green, tRNA are in yellow, rRNAs are in olive green and intergenic spacers are in black.

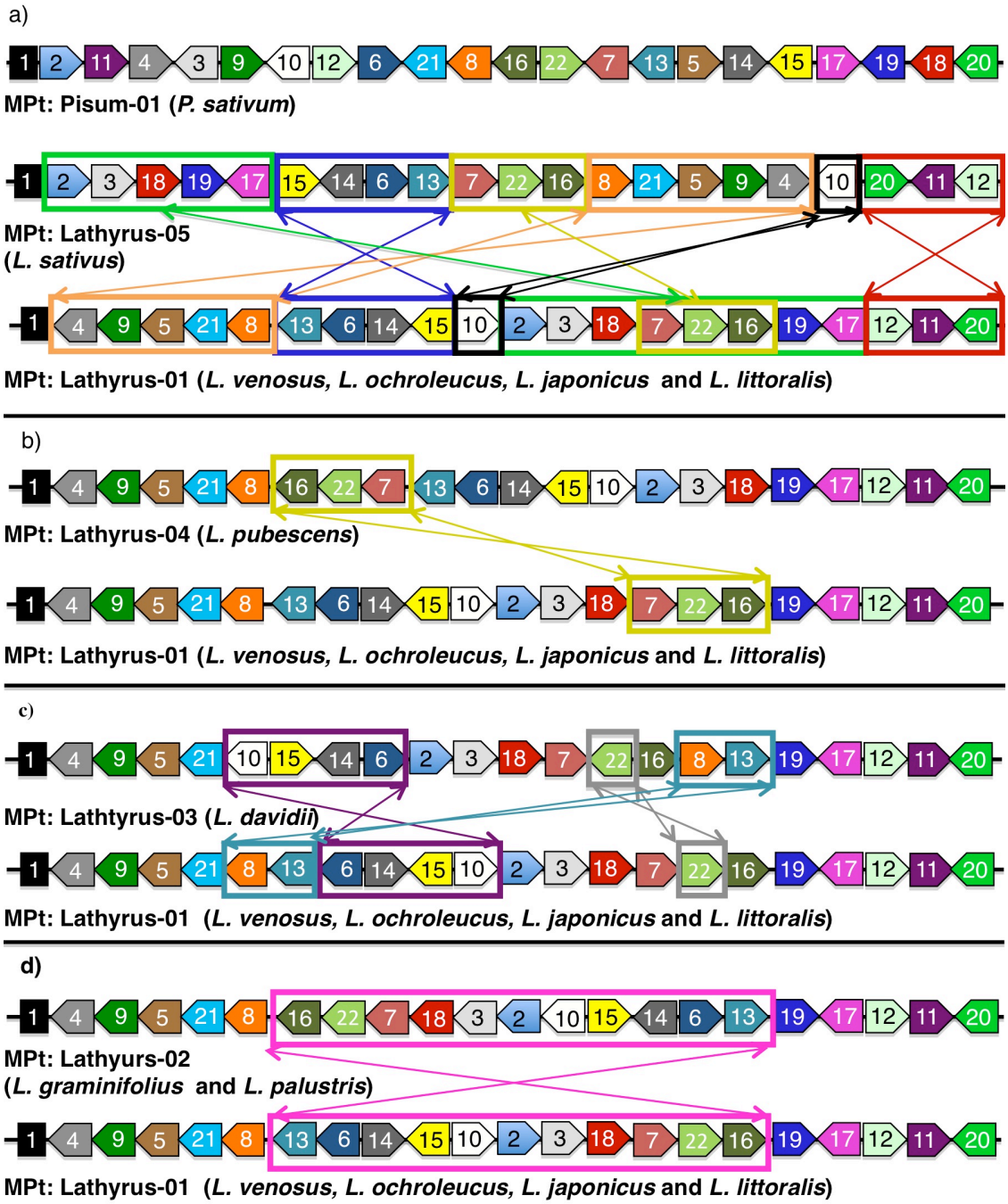

a)
MPt: Pisum-01 (P. sativum)
MPt: Lathyrus-05 (L. sativus)
MPt: Lathyrus-01 (L. venosus, L. ochroleucus, L. japonicus and L. littoralis)
b)
MPt: Lathyrus-04 (L. pubescens)
MPt: Lathyrus-01 (L. venosus, L. ochroleucus, L. japonicus and L. littoralis)
c)
MPt: Lathtyrus-03 (L. davidii)
MPt: Lathyrus-01 (L. venosus, L. ochroleucus, L. japonicus and L. littoralis)
d)
MPt: Lathyurs-02 (L. graminifolius and L. palustris)
MPt: Lathyrus-01 (L. venosus, L. ochroleucus, L. japonicus and L. littoralis)

Figure 4.4 Linear representation of gene order in the *Lathyrus* and *Pisum sativum* plastomes. Each numbered box corresponds to a set of plastid genes in a particular order (see Table 4.3). Major plastotypes (MPt), which were defined based on gene order, are shown underneath each plastome, with the species that sharing each MPt shown in parentheses. The orientation of individual arrowed boxes indicates in what direction (5'-> 3' or 3'<-5') majority of the genes within each block are transcribed in *L. venosus*. Inversions of entire gene blocks between major plastotypes are represented by a horizontal flip of the coloured boxes. Structural similarity among *Lathyrus* MPt was examined by comparing Lathyrus-02 –Lathyrus-05 to Lathyrus-01 (a-d). Possible inversions and/or translocations events of syntenic gene blocks are shown using coloured rectangles and arrows in the same color. A single arrow indicates a translocation event, where two crossing arrows indicate an inversion or a translocated inversion.

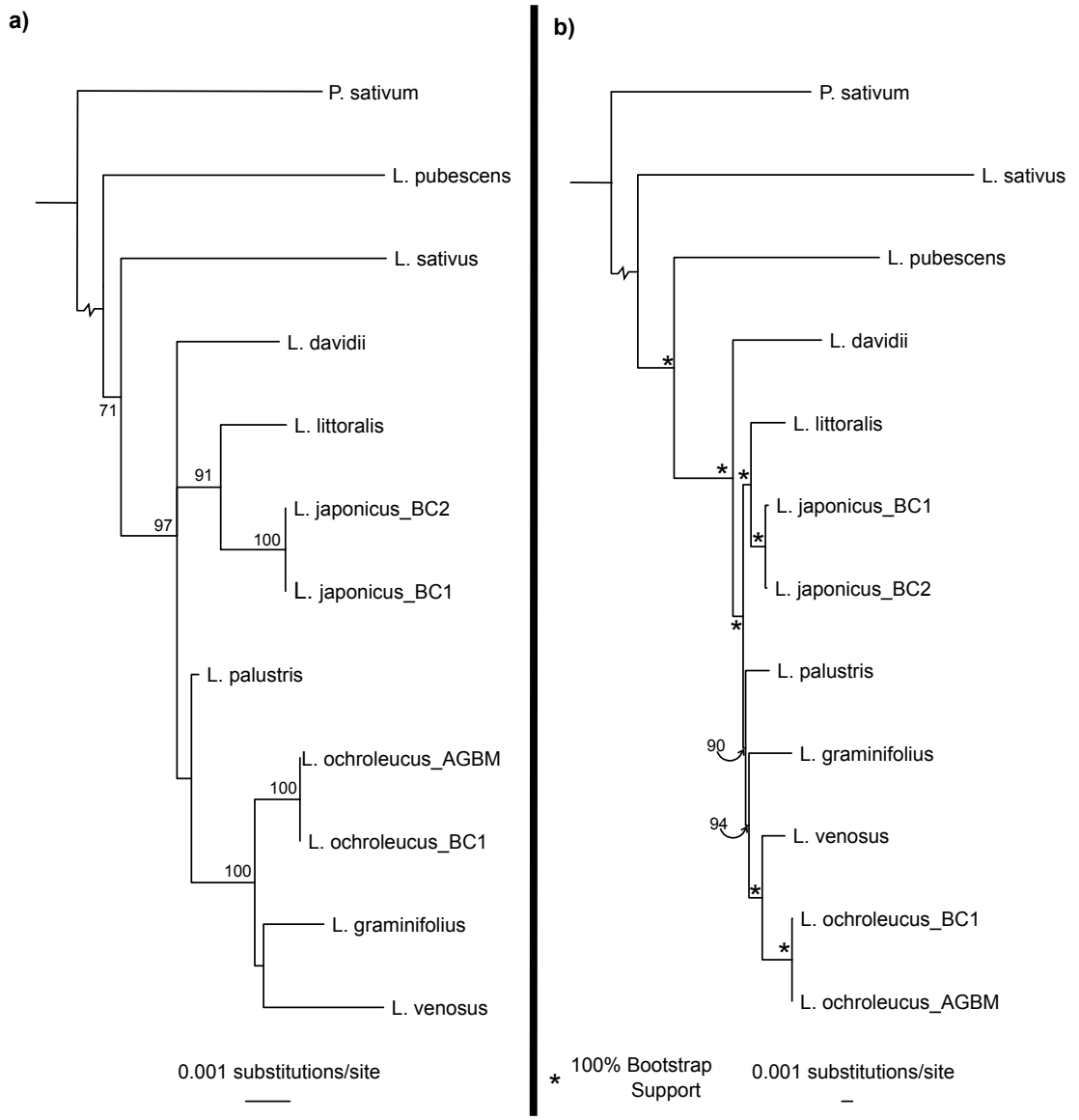
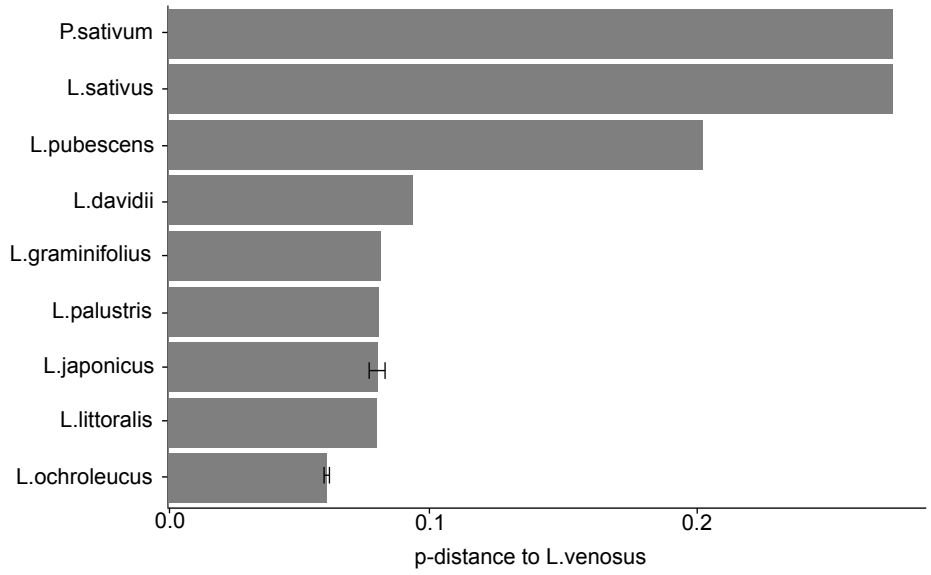


Figure 4.5 Phylogenetic trees representing the relationships among the sequenced *Lathyrus* species, based on the transcribed region of the large ribosomal subunit (A) and protein coding regions within the plastome (B). Trees were inferred using maximum likelihood and the support values, that are drawn on the nodes, are based on a 100 bootstrap searches. The long outgroup branch, splitting *P. sativum* and *Lathyrus,* was replaced with a wrinkle to save space. (A) Support values below 50 are omitted and (B) star indicates a 100% bootstrap support (*).

a)


P.sativum
L.sativus
L.pubescens
L.davidii
L.graminifolius
L.palustris
L.japonicus
L.littoralis
L.ochroleucus
0.0
0.1
0.2
p-distance to L.venosus

b)

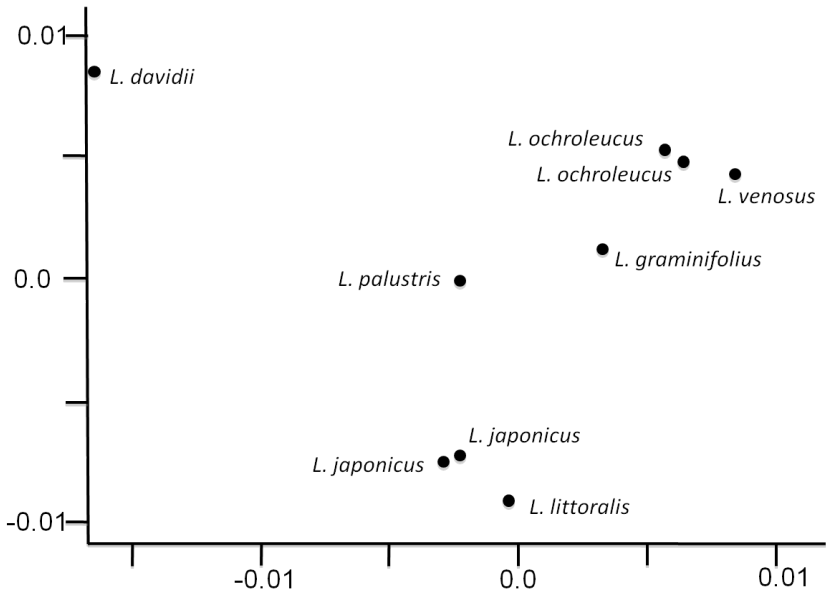

0.01
0.0
-0.01
L. davidii
L. ochroleucus
L. ochroleucus
L. venosus
L. graminifolius
L. palustris
L. japonicus
L. japonicus
L. littoralis
-0.01
0.0
0.01

Figure 4.6 Pairwise genetic distances calculated from exomic SNPs. (A) Standard p-distance between *Lathyrus venosus* and all other species, error bars are +/- 2*standard deviations. (B) A Principal Coordinate Analysis (PCoA) of the pairwise modified p-distance (see methods). *Pisum sativum* and *L. sativus* were omitted from the analysis, in order to enhance the resolution among genetically similar taxa.

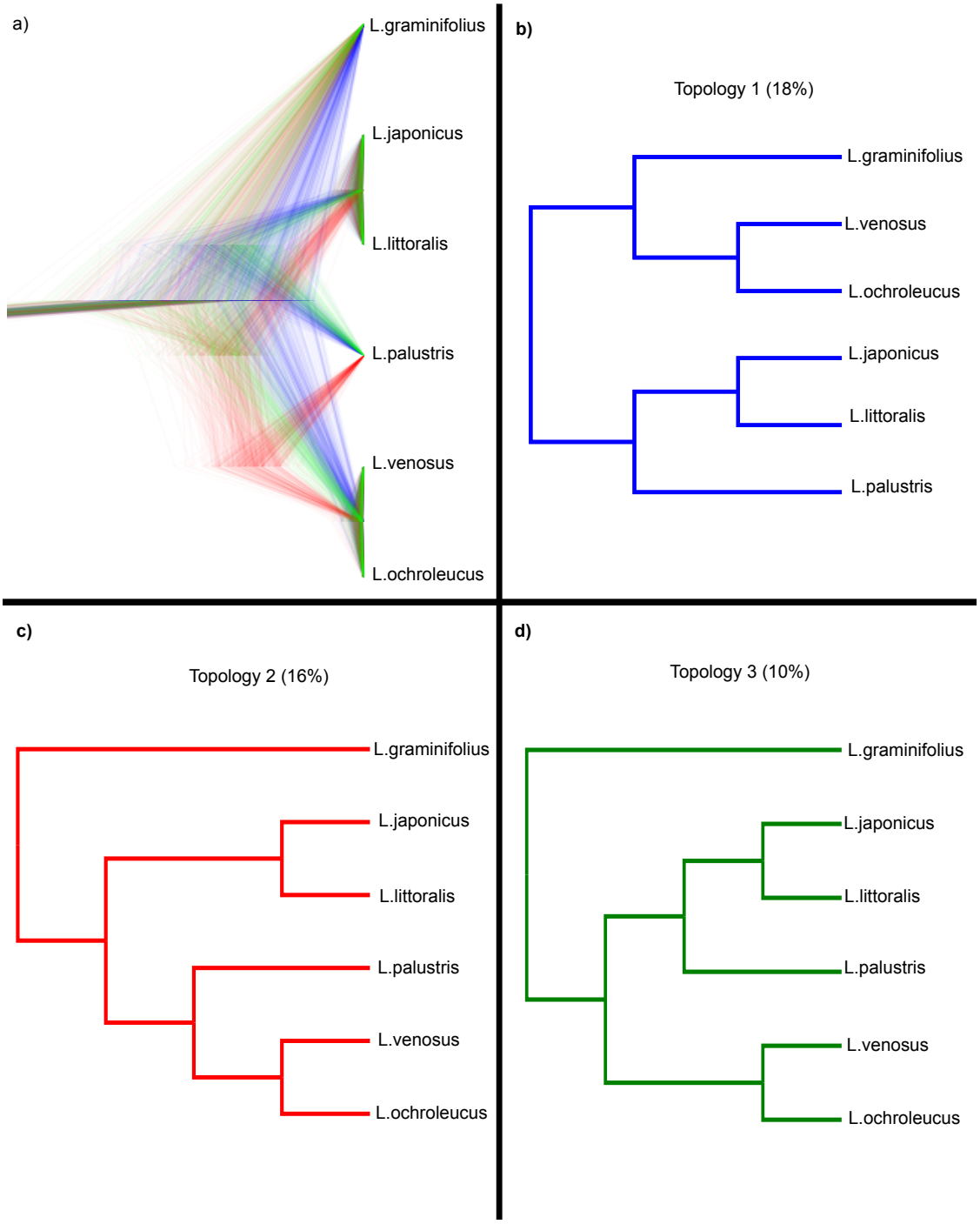

a)
L.graminifolius
L.japonicus
L.littoralis
L.palustris
L.venosus
L.ochroleucus
b)
Topology 1 (18%)
L.graminifolius
L.venosus
L.ochroleucus
L.japonicus
L.littoralis
L.palustris
c)
Topology 2 (16%)
L.graminifolius
L.japonicus
L.littoralis
L.palustris
L.venosus
L.ochroleucus
d)
Topology 3 (10%)
L.graminifolius
L.japonicus
L.littoralis
L.palustris
L.venosus
L.ochroleucus

Figure 4.7 A DensiTree visualization of the three most frequently observed topologies of the North-American *Lathyrus,* produced by SNAPP (A), and individual topogies with their corresponding frequency in parenthesis (B-D). The color of trees B-D matches each topology in the densi-tree (A).

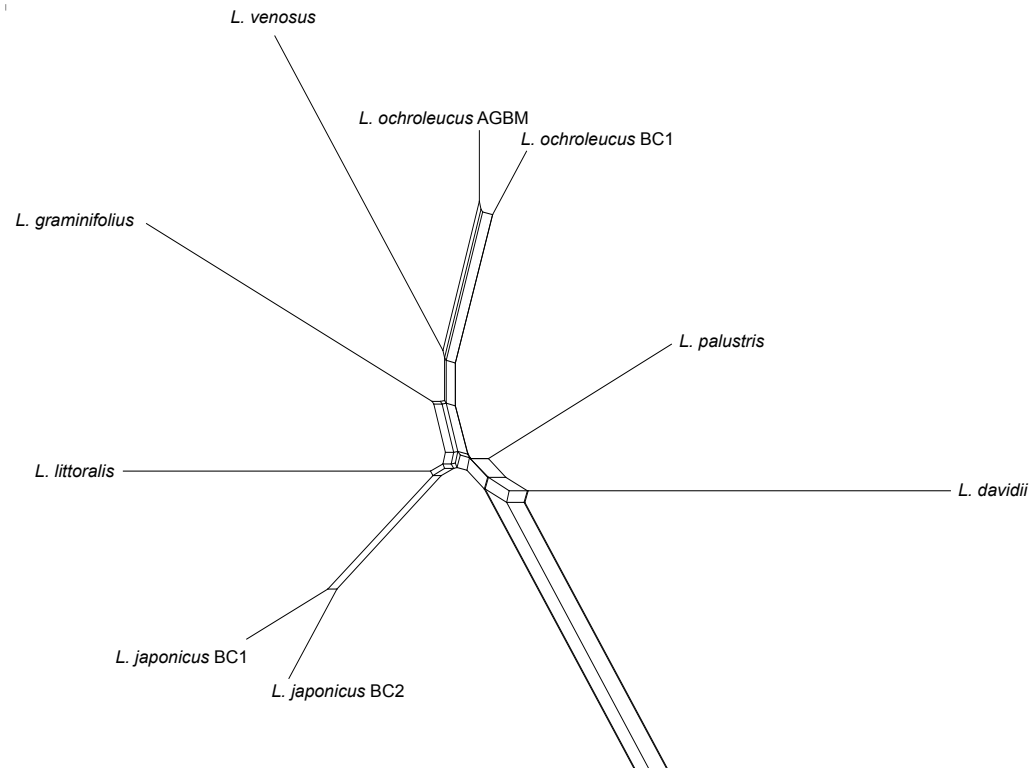


Figure 4.8. A network analysis of the North-American *Lathyrus* species and *L. davidii* using SplitsTree, based on the exomic SNPs. Two outgroup species, *L. pubescens* and *L. sativus*, were also included in the analysis but are not shown here.

SPECIES OF THE PACIFIC SLOPE AND COAST	*L. japonicus* Willd. *L. palustris* L. *L. jepsonii* Greene *L. delnorticus* C. L. Hitchcock *L. biflorus* T. W. Nelson & J. P. Nelson *L. vestitus* Nuttall in J. Torrey & A. Gray *L. splendens* Kellogg *L. nevadensis* S. Watson *L. holochlorus* (Piper) C. L. Hitchcock *L. tracyi* Bradshaw *L. sulphureus* W. H. Brewer ex A. Gray *L. polyphyllus* Nuttall in J. Torrey & A. Gray *L. glandulosus* Broich *L. torreyi* A. Gray *L. littoralis* (Nuttall) Endlicher ex Walp.
SPECIES OF THE MOUNTAIN WEST	*L. eucosmus* Butters & H. St. John (*L. decaphyllus* auct.) *L. brachycalyx* Rydberg *L. grimesii* Barneby *L. lanszwertii* Kellogg, *sens. lat*. (incl. *L. bijugatus*, *L. leucanthus*) *L. hitchcockianus* Barneby & Reveal *L. ochroleucus* Hooker *L. nevadensis* S. Watson

SPECIES OF THE MOUNTAIN WEST cont.	*L. laetivirens* Greene ex Rydberg *L. brownii* Eastwood *L. pauciflorus* Fernald *L. rigidus* T. G. White *L. graminifolius* (S. Watson) T. G. White
SPECIES OCCURRING EAST OF THE WESTERN CORDILLERAS	*L. japonicus* Willd. *L. decaphyllus* Pursh (*L. polymorphus* Nuttall) *L. palustris* L. *L. venosus* Muhlenberg ex Willd. *L. ochroleucus* Hooker *L. pusillus* Elliott

Table 4.1 Native *Lathyrus* species in North America. The genus is very diverse in the west. *Lathyrus venosus* is one of a small number of eastern species. The taxonomic treatment in this table follows that of S. Broich (pers. comm.).

Species_population (origin)	**NO reads**	**Plastome length (bp) (coverage) GenBank accession**	**45S rDNA length (bp) (coverage)**	**5S rDNA length (bp) (coverage)**
Lathyrus davidii (RP)	1.12E+07	123,896 (220) KJ806192	7,984 (146)	263 (635)
L. graminifolius (USDA)	1.57E+07	122,439 (180) KJ806193	7,463 (212)	279 (1,367)
*L. japonicus*_BC1 (BC, Can)	1.49E+07	124,243 (205) KJ806194	7,829 (299)	283 (2,775)
*L. japonicus*_BC2 (BC, Can)	1.89E+07	124,291 (280) KJ806195	7,819 (319)	283 (2,289)
L. littoralis (WA, USA)	1.45E+07	123,735 (140) KJ806196	7,940 (286)	286 (3,157)
L. ochroleucus_AGBM (BC, Can)	1.83E+07	123,912 (345) KJ806197	7,576 (330)	279 (1,213)
*L. ochroleucus*_BC1 (BC, Can)	1.98E+07	124,263 (380) KJ806198	7,590 (723)	279 (1,906)
L. palustris (BC, Can)	1.31E+07	124,288 (172) KJ806199	7,834 (264)	283 (513)
L. pubescens (RP)	1.41E+07	126,422 (210) KJ806200	8,106 (227)	275 (162)
L. venosus (SK, Can)	1.44E+07	125,460 (110) KJ806202	8,283 (149)	279(3,072)
L. sativus (RP)	1.39E+07	120,816 (213) KJ806201	8,037 (1,077)	na
P. sativum (cv Carrera)	1.41E+07	122,150 (343) KJ806203	8,058 (1,373)	329 (na)

Table 4.2 Illumina sequencing and *de novo* assembly summary.

Gene Block	Approximate size (kb)	Gene(s) that make up the block	note
1	15.3	rpl32, ndhF, psbA, matK, rbcL, atpB, atpE, ndhC, ndhK, ndhJ	Same gene order in all species
2	3.6	rps4, ycf3	
3	4.9	psaA, psaB, rps14	+psbM in Pisum
4	5.7	ycf2	
5	0.1	petN	
6	0.8	psbE, psbF, psbL, psbJ	
7	10.4	rpoC2, rps2, atpI, atpH, atpF, atpA	
8	2.0	cemA, psaI	+ycf4 in Pisum
9	3.3	psbZ, psbC, psbD	
10	7.3	rrn5, rrn4,5, rrn23, rrn16	
11	2.5	ndhE, psaC, ndhD	
12	14.8	rps12-3'exon, rps7, ndhB, ycf1, rps15, ndhH, ndhA, ndhI, ndhG	
13	6.0	rpoB, rpoC1, rps8	
14	2.8	clpP, rps12-5'exon, rpl20	
15	2.9	rps18, rpl33, psaJ, petG, petL	
16	1.8	accD	
17	5.4	petD, petB, psbH, psbN, psbB	
18	5.4	rpl23, rpl2, rps19, rps3, rpl16, rpl14	
19	2.0	rpl36, rps11, rpoA	
20	1.0	ccsA	
21	1.0	petA	
22	0.7	psbI, psbK	

Table 4.3 Composition and size of plastid gene blocks shown in Figure 4.3.

Coverage	Actual number of transcriptome nucleotide positions	Expected from Poisson distribution
0x	24,219,125	29,466,433
1x	5,695,230	4,243,166
2-3x	1,782,726	320,172
4-7x	455,530	543
8-15x	317,932	0
16-31x	357,562	0
32-63x	198,325	0
>63x	1,003,885	0
Total:	34,030,315	34,030,315

Table 4.4 Coverage of reference transcriptome from whole genome sequencing at raw read depth of 0.144X.

Chapter 5: Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*)

5.1 Introduction

The increased availability and lowered costs of various massively parallel sequencing technologies have resulted in a dramatic expansion of fully sequenced plastid genomes (Moore et al., 2006). There are currently 512 plastome sequences listed at NCBI's Organelle Genome Resource website (http://tinyurl.com/ncbi-plastid-genomes, 17 April 2014), and half of them have been made available since 2012. Their structure, gene order and gene content is generally highly conserved across most flowering plants, where most plastomes have two copies of a highly conserved inverted repeat (IR) and two conserved single copy regions (see Wicke et al., 2011). However, there are some well known and striking exceptions from these conserved plastome structures in several photosynthetic angiosperm lineages, such as the Geraniaceae (Guisinger et al., 2011) and Campanulaceae (Haberle et al., 2008).

Subterranean clover (*Trifolium subterraneum*) is also known to have an unusual plastid genome structure (Cai et al., 2008). First it lacks one copy of the inverted repeat, but that is a character shared with a large group of papilionoid legumes (Wojciechowski et al., 2000) designated the inverted repeat lacking clade (IRLC). Secondly its plastome has undergone more than a dozen rearrangements, i.e. translocations and/or inversions, compared to the plastid genome of *Medicago truncatula* (Cai et al., 2008). However, highly rearranged plastomes also occur in the related tribe Fabeae, represented by *Lathyrus sativus* and *Pisum sativum* (see Magee et al., 2010). Finally, and very unusually, its plastome contains about fourfold the amount of repeated DNA compared to related legume species and is consequently about 20 kb longer (Cai

et al., 2008). Up until now it has not been known whether this unusual repeat-rich plastome structure is unique to *T. subterraneum*, a general feature of *Trifolium* species, or characteristic of some part of the genus *Trifolium*. A recent study has shown that these repetat-rich plastomes seem to be a feature of the subgenus *Trifolium* (Sabir et al., 2014). However, plastome sequences from only two of the seven sections within *Trifolium* are available. To investigate this further, we performed a low coverage whole genome shotgun sequencing of nine strategically sampled *Trifolium* species and were able to assemble eight plastid genomes. These plastomes were then analyzed to elucidate the phylogenetic distribution of plastome variation in the genus.

5.2 Material and methods

5.2.1.1 Plant material and Illumina sequencing

Sampling was strategically placed across the genus using the sectional classification of Ellison et al. (2006). Total DNA was extracted from fresh leaf material of plants that had been grown from seeds in a greenhouse (at UBC), following a modified version of the CTAB protocol (Doyle and Doyle, 1987). The seeds were obtained from the United States Department of Agriculture (USDA) National Plant Germplasm System, more specifically the National Temperate Forage Legume Genetic Resources Unit in Prosser, WA. Plants were grown until they flowered, the material was critically determined and herbarium specimens collected (of all but one accession, see Table 5.1). RNase treatments were performed (cat. 19101, QIAGEN, Germantown, MD) and DNA quality was assessed by visual inspections on 1% agarose gels. Illumina sequencing libraries were constructed from high quality DNA, using the NEXTflexTM DNA sequencing kit (100 bp Paired-End reads) (cat: 5140-02, BiooScientific Corp, TX). We followed the manufacturer's protocol and c. 400 bp DNA fragments were size selected using Agencourt

AMPure XpTM magnetic beads (cat. A63880, BeckmanCoulter Genomics, MA). Completed libraries were pooled and sequenced on a lane of the Illumina HiSeq-2000 platform.

5.2.2 Plastid genome assemblies and annotation

Trimmomatic v.0.3 (Lohse et al., 2012) was used to trim and remove low quality Illumina reads, with the following flags: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15MINLEN:36. High quality reads were used in all subsequent analysis and singlet reads, i.e. reads without a paired end, were discarded. We used the *de novo* method implemented in CLC Genomic Workbench v.7.0.2 to generate assemblies for each species, using the default settings. Contigs of plastid origin were identified by a BLAST search (Altschul et al., 1997) to the available plastid genome of *Trifolium subterraneum,* published in Cai et al., (2008) [NCBI Reference Sequence: NC_011828]. These were generally the largest and most highly covered contigs in the *de novo* assembly and always had an E-value of 0 when blasted to the *T. subterraneum* plastome. Regions with nucleotides represented as Ns were manually resolved by retrieving sequence information directly from the quality trimmed reads. For three out of eight species, the *de novo* assembly returned a single large plastid contig. This was not the case for the plastid genome assembly of the remaining six *Trifolium* species, where the *de novo* assembly resulted in about five to 10 plastid contigs. In those cases we decided to extend both ends of the plastid contigs by manually appending sequence from the Illumina reads. The sequence of each contig was extended until an overlap of at least 50 bp could be identified at the end of another contig in the assembly, whereupon the contigs were joined. We identified a few short (1 - 2 kb) regions which were duplicated and hence could be joined to more than two other contigs in the *de novo* assembly. We resolved positions of these regions by adding them in the plastome assembly as

many times as was necessary for all the plastid contigs to be joined in a circle. This methodology worked well for joining all the plastid contigs into a well-supported hypothesis for a plastome assembly, with the exception of red clover (*Trifolium pratense*). The quality of each plastome assembly was verified visually by inspecting a BWA mem pileup, v. 0.7.5a (Li and Durbin, 2009), of paired end reads using Tablet v.1.13.12.17 (Milne et al., 2013). We ensured that the connections between manually joined contigs were supported by paired-end read mapping. Finally all plastome assemblies were annotated using DOGMA (Wyman et al., 2004). Maps of plastid genomes were generated using GenomeVX (Conant and Wolfe, 2008) and visually adjusted using inkscape (www.inkscape.org).

5.2.3 Identification and analysis of repeated DNA in the plastid genomes

Repeated segments were determined in each of the plastid genomes by a reciprocal BLASTN search (Altschul et al., 1997), using the NCBI's online BLAST service (http://blast.ncbi.nlm.nih.gov/), where each sequence was used as the query and subject. We used these BLAST outputs to generate dotplots, using R (R Core Team, 2014). We decided to analyze repetitive regions that the BLASTN search identified as 300 bp or longer and had an E-value of 0, by subjecting them to a BLASTX search using a custom database of plastid genes from closely related species. A BLASTX similarity cutoff was set by removing all hits with an E-value larger than 1E-06. We furthermore subjected all sequences that did not show significant BLASTX similarity to the *Trifolium* plastid genes, to a BLASTX and BLASTN search to NCBI's nucleotide collection (nt/nr) and non-redundant protein sequences (nr).

5.2.4 Phylogenetic analysis

Due to the extensive rearrangements observed in the plastomes (see chapter 4), we restricted our plastome phylogenetic analysis to protein coding genes. We used a custom phylogenetic pipeline, plast2phy, that extracted protein coding regions from DOGMA annotated plastomes, aligned individual gene with Mafft v. 7.0.5(-auto flag) (Katoh and Standley, 2013), trimmed alignment gaps using trimAl v.1.2 (-automated1 flag) (Capella-Gutiérrez et al., 2009) and finally generated a concatenated alignment of all genes. The pipeline, Plast2phy, written in Python, is available at https://github.com/saemi/plast2phy. We only included genes that were present in all species and showed no evidence of duplication. Models of base substitution were tested for the concatenated matrix using jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012). Using the Akaike information criterion (AIC), we determined the GTR+G+I model optimal for the concatenated plastome alignment. We analysed the dataset under maximum likelihood (ML; Felsenstein, 1973) using GARLI (Zwickl, 2006). We ran GARLI v. 2.0 with default settings, using ten independent searches and 100 bootstrap replicates. Bootstrap consensus was calculated using SumTrees v. 3.3.1 in the DendroPy package (Sukumaran and Holder, 2010). To obtain an approximate age for nodes on this tree by molecular dating we set the *Medicago-Trifolium* divergence at 24.1 MYA (Lavin et al., 2005) and estimated the other node ages under penalized likelihood (Sanderson, 2002) implemented in r8s v.1.8 (Sanderson, 2003). We chose the penalized likelihood algorithm since a likelihood ratio test, using constrained and unconstrained likelihood scores obtained from PAUP* (Swofford, 2003), rejected the molecular clock ($p < 0.001$). Confidence intervals were estimated by running dating analysis on 100 bootstrap resampled datasets generated by GARLI and are represented as two times the standard deviation. Trees from phylogenetic analysis were drawn using FigTree v.1.4.0

(http://tree.bio.ed.ac.uk/software/figtree/), rooted with *Medicago truncatula* [NCBI Reference Sequence NC_003119] and visually adjusted using Inkscape (http://www.inkscape.org/).

5.3 Results

5.3.1 Plastome assembly and structural variability

The plastomes of three sequenced *Trifolium* species, *Trifolium boissieri*, *T. strictum* and *T. glanduliferum*, were tractable, assembled easily and had no unusual structure, at least in the context of related species in the inverted repeat loss clade (IRLC), such as species of the tribe Fabeae (see Magee et al., 2010) (Figure 5.1A and 1C)[GenBank accessions: KJ788284, KJ788292 and KJ788285]. The remaining six *Trifolium* species had plastome structure similar to *T. subterraneum* (described in Cai et al., 2008), containing several short repetitive regions, which made them difficult to assembly. However, with one exception we were able to assemble all of them successfully, using careful analysis of the paired-end sequences (see below) (Figure 5.1B and 1D) [GenBank accessions: KJ788286 - KJ788289 and KJ788291]. The exception is *T. pratense*. The plastome structure of *T. pratense* appears to be highly complex and we were unable to complete a full assembly of it using the methods employed in this study. This is most likely due to the short read length of the Illumina sequences (100 bp) and their small insert sizes (ca. 250 bp). It is clear that this plastome contains several repeated regions, similar to *T. subterraneum*. We were, however, able to put together three large plastid contigs for *T. pratense*, with a combined length of 121 kb, which were annotated with DOGMA (Wyman et al., 2004) and used for phylogenetic analysis [GenBank accession KJ788290]. We generated draft plastome assemblies for the remaining five "difficult" species and verified them using mapping information from the paired end Illumina reads. We ensured that the entire assembly was highly

covered (at least 100x) and that the placement of adjacent plastid regions were supported by the mapping of paired end reads. However it is important to note, that despite these careful quality checks, assembly of some regions that have a particularly high frequency of repeats should be considered provisional.

5.3.2 Phylogenetic distribution of the refractory species

The six *Trifolium* species with plastomes that were rich in repeats differed markedly from the three species that were tractable for assembly. Furthermore, the six species and the previously sequenced *T. subterraneum* all belong to the same clade within *Trifolium*, which we call the "refractory clade" (i.e. resistant to assembly), comprising subg. *Trifolium* sections Lupinaster, Trifolium, Tricocephalum, Vesicastrum and Trifoliastrum (see Figure 5.2). The remaining three species, which have plastomes that assembled readily, are all outside this clade and represent subgenus *Chronosemium* and subgenus *Trifolium* section Paramesus (see Figure 5.2). Molecular dating analysis of the refractory clade revealed that it originated, i.e. had a most recent common ancestor (MRCA), about 12.4 – 13.8 MYA. The refractory clade has therefore had a considerable amount of time in which to accumulate a high diversity of repeat patterns. This compares to estimated MRCA ages of 16.1 Mya for *Trifolium* subgenus Trifolium and 18.4 Mya for the genus *Trifolium* as a whole. The phylogeny in Figure 5.2 was constructed inferred using maximum likelihood from a concatenated matrix of 68 protein coding plastid genes, with a combined aligned length of just over 48 kb. All nodes have 100% bootstrap support and the topology is consistent with the most recent phylogenetic treatment of *Trifolium* (see Ellison et al., 2006).

5.3.3 The nature of the duplicated regions

The amount of repetitive sequence in the *Trifolium* plastomes ranged from 1 % to 22 % (see Table 5.1). This was based on a reciprocal BLASTN search, where each plastome was used as the query and subject. Not surprisingly, plastomes belonging to the refractory clade contained about 5 - 20 times more repetitive DNA than *Trifolium* species in subgenus *Chronosemium* and subgenus *Trifolium* section Paramesus. Repetitive regions larger then 300 bp were subjected to a BLASTX search, in order to determine whether they contained any protein coding regions. A protein database was compiled of all annotated genes from the *Trifolium* plastomes. About 60% of the longer repetitive regions did not show significant (E-value ≤ 1E-06) similarity to any protein coding regions. The remaining 40% of showed similarity to one of the following plastid genes: *clpP*, *psaJ*, *psbK*, *psbN*, *rpl2*, *rpl23*, *rpoB*, *rps18*, *rps3*, *ycf1* and *ycf3*, indicating that at least partial duplication of genic regions has occurred. Patterns of genic duplication are highly variable, and no genes were found to be duplicated in more than two species. All repetitive gene regions consisted of only a partial reading frame and no duplicated genes were found with the entire reading frame intact in more than one copy. However a small number of gene duplicates appeared to have relatively large portions of open reading frames (ORF) that appeared to be intact (i.e. identical at the protein level to the full length functional copy). A BLASTX search of repetitive regions revealed that *psaJ, psbK*, *psbN, rpl2, ycf1* and *ycf3* had instances where they are 100% intact over 5 – 30% of the length of the corresponding complete ORF.

5.3.4 Instances of gene loss

The *Trifolium* plastomes have undergone some gene loss. *AccD* appears to missing from plastomes in the following sections of the subgenus Trifolium: Trifolium, Tricocephalum,

Vesicastrum and Trifoliastrum. This is in agreement with previous reports (Sabir et al., 2014), but our broad phylogenetic sampling shows that the plastid copy of *accD* is still functional in *T. lupinaster* (section Lupinaster of the subgenus *Trifolium*). We found *ndhB* to be missing in the plastome *T. meduseum.* That gene loss had previously been reported in the plastome of *T. subterraneum* (Cai et al., 2008). The loss of *ndhB* is likely a shared event in these two species, since they belong to the same section (Trichocephalum) (see Figure 5.2).

5.4 Discussion

5.4.1 The Phylogenetic distribution of plastome types

Two types of plastomes were observed among the *Trifolium* species in this study. Three species had no unusual structure in the context of IRLC plastomes, which are known to have numerous rearrangements (see Magee et al. 2010). The six remaining species had enlarged, repeat-containing plastome structure similar to *T. subterraneum* (described in Cai et al., 2008). These had several repeated regions and were 7 – 21 kb larger compared to *Medicago truncatula, T. boissieri*, *T. strictum* and *T. glanduliferum* (see Figure 5.2). Furthermore, we observed a strong phylogenetic clustering of these larger, more repeated plastomes. They comprise a single clade (Figure 5.2) and, strikingly, all the sampled species of this clade have the unusual repeat-rich plastomes. The North-American section Involucrarium also belongs to this clade (it is sister to section Trifoliastrum: Ellison et al., 2006), and is therefore likely to have similar "refractory" plastomes, although this has yet to be confirmed. This clade contains about 200 species or roughly 70% of the genus, which has a total of roughly 300 species. The evolutionary success of the clade with the repetitive plastomes leaves no doubt about the functional effectiveness of these plastomes despite their unusual structure.

5.4.2 Potential causes of genome instability and functional significance of the repeat regions

Some regions within the *T. subterraneum* plastome have been suspected to be of bacterial origin by lateral transfer (Cai et al., 2008), however those regions are not repetitive. The repetitive regions seem to be of plastid origin, as our BLASTN and BLASTX searches to NCBI's nucleotide collection (nt/nr) and non-redundant protein sequences (nr), did not indicate any obvious non-plastid sequence. Repeated sequences have been associated with the plastid rearrangments in other angiosperm lineages, such as the Campanulaceae (Cosner et al., 1997; Haberle et al., 2008) and Geraniaceae (Chumley et al., 2006; Guisinger et al., 2011). However, it is clear that plastomes can be highly rearranged without being highly repetitive, such as *Pisum sativum* and *Lathyrus sativus* (Fabaceae) (Magee et al., 2010).

It is not yet known what has caused repeated sequences to evolve within the plastomes of the *Trifolium* refractory clade although it is likely that this is under the control of nuclear genes. Plastid reorganization is a common process and there is evidence that reorganized molecules may comprise around 1% of the plastid complement of plant cells (Lilly et al., 2001). This variation may be caused by either homologous recombination (HR) acting on perfect repeats of >50bp, or microhomology-mediated break-induced replication (MMBIR) acting on microhomologous repeats of <30bp (Maréchal et al., 2009; Maréchal and Brisson, 2010). Various nuclear genes are known to be important in maintaining plastome stability through recombination control and surveillance, and these are therefore candidate genes for genome instability in certain lineages. The *RecA* gene has an important role in homologous recombination and there are plant copies that localise to the chloroplast (Cerutti et al., 1992; Cerruti and Jagendorf, 1993). Plastid-targeted *WHIRLY* genes are known to promote plastome

stability, apparently by preventing build up of abnormal molecules produced by MMBIR. Arabidopsis plants lacking functional copies of relevant *AtWHY* genes show increased accumulation of abnormal plastid DNA with irregular duplications, deletions and circularization events (Maréchal et al., 2009).

Most of these abnormal plastid forms are deleterious and transient, not contributing to plastid evolution in the majority of plant lineages. An important consideration therefore is not the genomic changes that cause an increased frequency of plastome reorganization, but what genome changes allow such major plastome changes to persist without being eliminated as deleterious. It may be that certain plastid genes are particularly sensitive to genome rearrangements and their removal to the nuclear genome makes the plastid genome more permissive to rearrangement. The plastid gene *accD* for instance, known to be essential for plant development (Kode et al., 2005) and with recombinationally active repeats (Gurdon and Maliga, 2014), has been moved to the nucleus independently in two lineages with highly rearranged plastomes: Campanulaceae (Rousseau-Gueutin et al., 2013) and *Trifolium* (Magee et al., 2010; Sabir et al., 2014).

One possible consequence of having repeated sequences in a circular molecule, such as the plastid genome, is that if recombinationally active they could potentially allow intramolecular recombination of the plastome into subgenomic molecules much in the manner of the plant mitochondrion, which is rich in recombinational repeats. This has been suggested as a reason for the repeat richness of certain algal plastomes (Maul et al., 2002). The *Trifolium* plastomes are much less rich in repeats than *Chlamydomonas*, or than the mitochondrial genome of land plants, but it is nevertheless interesting to consider the idea that the repeat accumulation in *Trifolium* plastomes is being driven by the advantage of maintaining substochiometric populations of specific ORFs in cellular subpopulations (Mackenzie and McIntosh, 1999).

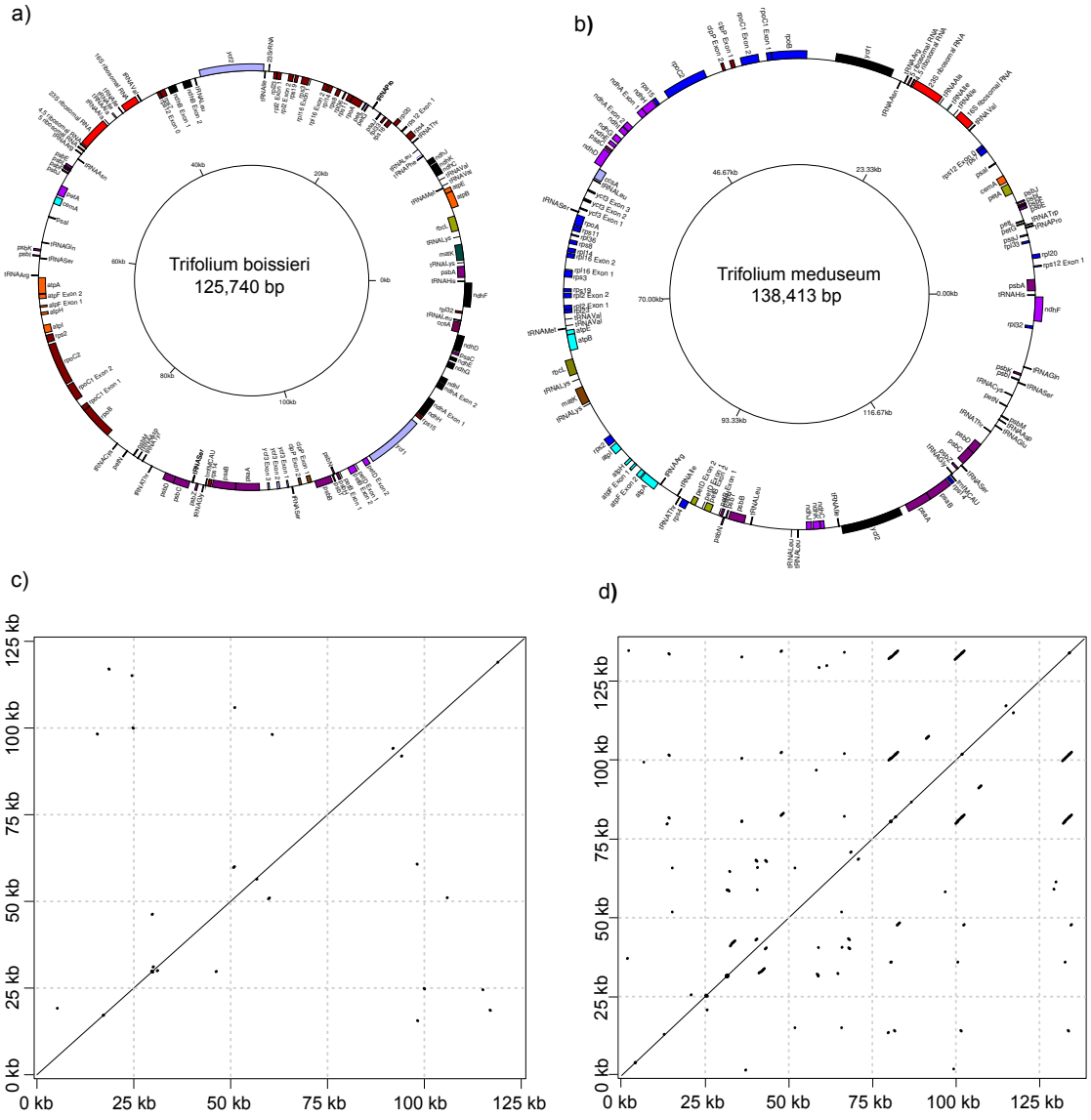


Figure 5.1 Gene map of two plastid genomes and dotplots illustrating repeated regions. Gene maps (a and b) and dotplots (c and d) of *Trifolium boisseri* (a and c) and *Trifolium meduseum* (b and d). The gene maps shows the postions and translation direction of protein coding genes, tRNAs and rRNAs in the plastid genome. The dotplots illustrate repeated sequences within the plastomes, identified by BLASTN, where aligned sequences

are shown as lines. Since the plastomes are aligned to themselves, the entire plastome is represented as a long and uninterupted diagonal line. Shorter lines and dots, which fall above and below the long line, are repeats.

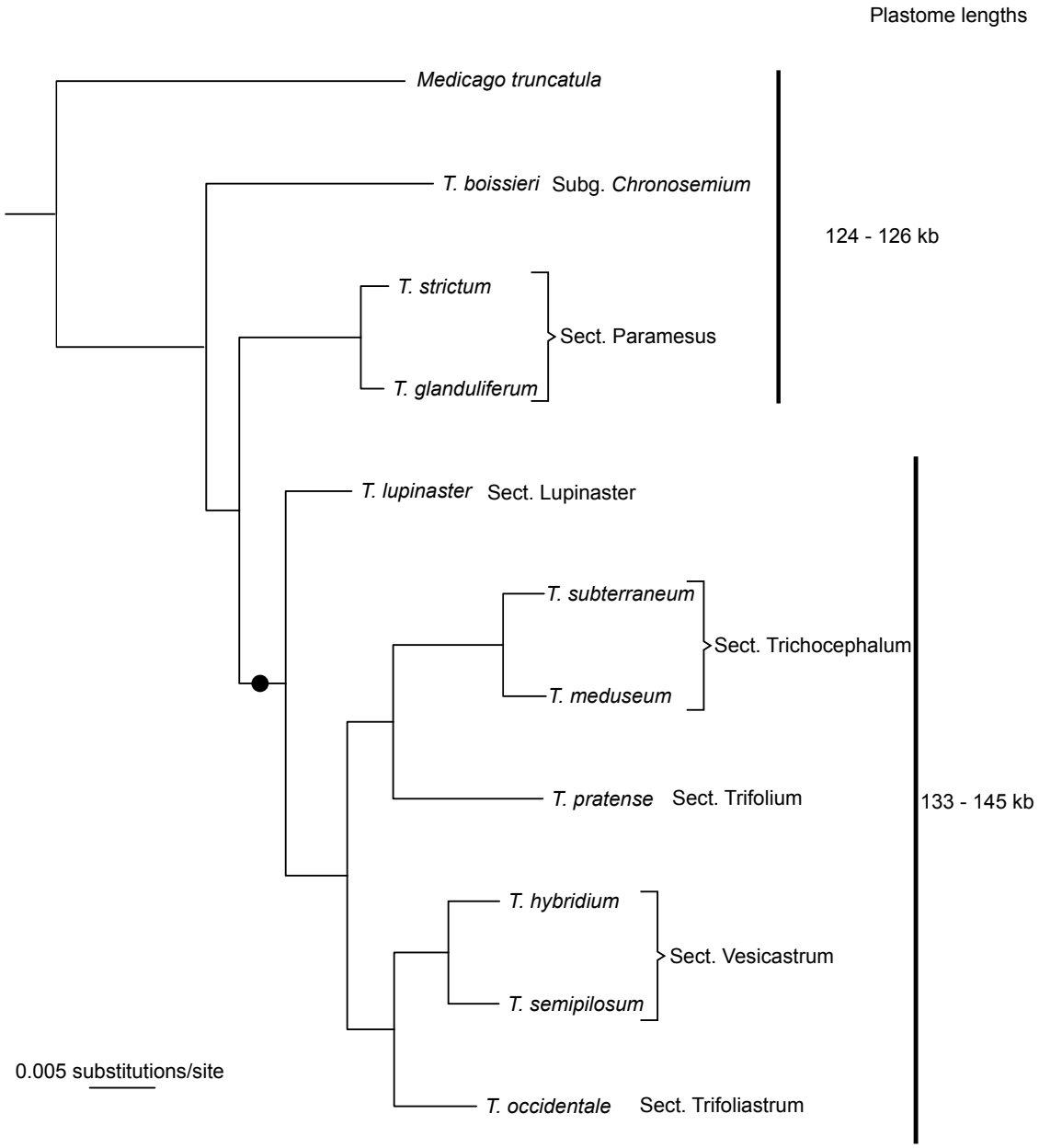


Figure 5.2 Phylogenetic relationships among the *Trifolium* species, generated using maximum likelihood from a concatenated matrix of 58 protein coding plastid genes with a combined aligned length of 48,058 characters. The black dot marks the clade with plastomes refractory to assembly and the black vertical lines indicate the

plastome size categories. The age of the most recent common ancestor (MRCA) of the refractory clade is estimated to be 12.4 – 13.8 million years old. All nodes have a 100% bootstrap support.

Species (USDA seed accession)	**Section**	**Plastome length (repetitive %)**	**NCBI accession**	**Herbarium voucher**
Medicago truncatula Gaertn. (NA)	NA	124,033 nt (2%)	NC_003119	NA
Trifolium boisseri Guss. (PI 369022)	Subg. Chronosemium	125,741 nt (1%)	KJ788284*	SS14-01 (UBC)
T. strictum L. (PI 369147)	Paramesus	125,835 nt (1%)	KJ788292*	SS14-02 (UBC)
T. glanduliferum Boiss. (PI 296666)	Paramesus	126,182 nt (1%)	KJ788285*	SSS14-03 (UBC)
T. lupinaster L. (PI 631632)	Lupinaster	135,077 nt (6%)	KJ788287*	SS14-04 (UBC)
T. meduseum Blanche (PI 369049)	Trichocephalum	138,441 nt (13%)	KJ788288*	SS14-05 (UBC)
T. subterraneum L. (NA)	Trichocephalum	144,763 nt (22%)	NC_011828	NA
T. pratense L. cv. Arlington (G 27569)	Trifolium	NA (NA)	KJ788290*	SS14-07 (UBC)
T. hybridum L. (PI 634109, as *T. pallescens*)	Vesicastrum	134,881 nt (8%)	KJ788286*	SS14-08 (UBC)
T. semipilosum Fresen. (PI 262238)	Vesicastrum	138,242 nt (11%)	KJ788291*	SS14-09 (UBC)
T. occidentale Coombe (PI 641363)	Trifoliastrum	133,806 nt (5%)	KJ788289*	NA

Table 5.1 Summary Illumina sequencing accessions and plastome assembly information. An asterisk marks plastome sequences newly reported in this thesis. Only a partial assembly of *T. pratense* was possible with our data (see text for explanation).

Chapter 6: Delimitation of conserved gene clusters in the scrambled plastomes of the IRLC legumes (Fabaceae: Trifolieae and Fabeae)

6.1 Introduction

The plastid genome, also known as the plastome, refers to the total genetic information of a single plant organelle, the plastid, which takes many developmental forms, the most notable being the chloroplast (Bock, 2007). Plastid genomes are circular structures of double stranded DNA, usually consisting of about 100-120 genes and are around 120-160 kb long in photosynthesizing plants (Bock, 2007). Their size, structure and gene content are highly conserved across land plants (Wicke et al., 2011). However there are exceptions, such as the Geraniaceae and Campanulaceae, which are two angiosperm families known to contain species with highly rearranged plastomes (Haberle et al., 2008; Guisinger et al., 2011). A dominating feature of plastid genomes is the presence of two copies of an inverted repeat (Wicke et al., 2011), however some plant groups have lost one copy of the repeat, one being a clade within papilionoid legumes, known as the inverted repeat lost clade (IRLC) (Wojciechowski et al., 2000).

Plants obtained their plastid organelles through an endosymbiosis event with a cyanobacteria-like organism, about 1.5 – 1.6 billion years ago (Margulis, 1970; Hedges et al., 2004). Its bacterial origin gives the plastid genome many prokaryotic features, such as small (70S) ribosomes and the absence of a mRNA 3' polyA tail (see Stern et al., 2010 for a review). An additional ancestral feature of the plastid genome is the organization of its coding region into multiple gene clusters, or operons (Sugita and Sugiura, 1996; Sugiura et al., 1998). These gene

clusters are stretches of the plastome consisting of several genes that are transcribed into di- or polycistronic units, which are then processed before translation (Stern et al., 2010). Several such clusters have already been identified in the plastid (Adachi et al., 2012; Ghulam et al., 2013; Stoppel and Meurer, 2013).

Several legume genera within the IRLC are known to harbor highly rearranged plastomes, as a result from multiple translocations and/or inversions: *Trifolium* (Cai et al., 2008; Sabir et al., 2014; chapter 5), *Pisum* (Palmer and Thompson, 1982; chapter 4), *Lathyrus* (Magee et al., 2010; chapter 4), *Lens* and *Vicia* (Sabir et al., 2014). The aim of this study is to analyze these rearrangements in these genera within IRLC, in order to investigate whether they can be used to study the organization of plastid genomes into operons.

6.2 Material and methods

6.2.1 Source of plant material

The plant material for this study came from three sources. First, live plants were collected in the field, transplanted into UBC's greenhouse facilities. Secondly, seeds were obtained from a commercial provider, Roger Parsons Sweet Peas (Chichester, UK). Thirdly, seeds were received from the USDA germplasm collection at Pullman, Washington (W6). A full list of germplasm used is given in Table 6.1. All plants were grown in greenhouse facilities at UBC. In all cases where plants required critical determination they were grown until flowering, and herbarium voucher specimens were then collected (UBC).

6.2.2 Illumina sequencing

Total DNA was extracted from fresh leaf material of plants that had been grown from seeds in a greenhouse (at UBC), following a modified version of the CTAB protocol (Doyle and Doyle, 1987). RNase treatments were performed (cat. 19101, QIAGEN, Germantown, MD) and DNA quality was assessed by visual inspections on 1% agarose gels. Illumina sequencing libraries were constructed from high quality DNA, using the NEXTflexTM DNA sequencing kit (100 bp Paired-End reads) (cat: 5140-02, BiooScientific Corp, TX). We followed the manufacturer's protocol and c. 400 bp DNA fragments were size selected using Agencourt AMPure XpTM magnetic beads (cat. A63880, BeckmanCoulter Genomics, MA). Completed libraries were pooled and sequenced on a lane of the Illumina HiSeq-2000 platform.

6.2.3 Plastid genome assemblies and annotation

Trimmomatic v.0.3 (Lohse et al., 2012) was used to trim and remove low quality Illumina reads, with the following flags: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15MINLEN:36. High quality reads were used in all subsequent analysis and singlet reads, i.e. reads without a paired end, were discarded. We used the *de novo* method implemented in CLC Genomic Workbench v.7.0.2 to generate assemblies for each species, using the default settings. Contigs of plastid origin were identified by a BLASTN search (Altschul et al., 1997) to a plastid genome of a closely related species. These were generally the largest and most highly covered contigs in the *de novo* assembly and always had an E-value of 0 when blasted to the reference plastome. Regions with nucleotides scored as Ns were manually resolved by retrieving sequence information directly from the quality-trimmed reads. For most species, the *de novo* assembly returned a single large plastid contig. When needed, multiple contigs containing plastid sequence

were joined by hand, using information from the quality-trimmed reads. The quality of each plastome assembly was verified by visually by inspecting a BWA mem pileup, v. 0.7.5a (Li and Durbin, 2009), of paired end reads using Tablet v.1.13.12.17 (Milne et al., 2013). We made sure that the connections between manually joined contigs were supported by paired-end read mapping. Finally all plastome assemblies were annotated using DOGMA (Wyman et al., 2004).

6.2.4 Phylogenetic analysis

Due to the extensive rearrangements observed in the plastomes (see chapter 4), we restricted our plastome phylogenetic analysis to protein coding genes. We used a custom phylogenetic pipeline, plast2phy, that extracted protein coding regions from DOGMA annotated plastomes, aligned individual gene with Mafft v. 7.0.5(-auto flag) (Katoh and Standley, 2013), trimmed alignment gaps using trimAl v.1.2 (-automated1 flag) (Capella-Gutiérrez et al., 2009) and finally generated a concatenated alignment of all genes. The pipeline, Plast2phy, written in Python, is available at https://github.com/saemi/plast2phy. Model of base substitution were tested for the concatenated matrix using jModelTest v.2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012). Using the Akaike information criterion (AIC), we determined the GTR+G+I model optimal for the concatenated plastome alignment. We analysed the dataset under maximum likelihood (ML; Felsenstein, 1973) using GARLI (Zwickl, 2006). We ran GARLI v. 2.0 with default settings, using ten independent searches and 100 bootstrap replicates. Bootstrap consensus was calculated using SumTrees v. 3.3.1 in the DendroPy package (Sukumaran and Holder, 2010). Trees from phylogenetic analysis were drawn using FigTree v.1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/), rooted with *Lotus japonicus* [NCBI Reference Sequence NC_002694], with the inverted repeat manually removed.

6.2.5 Identification of locally collinear blocks (LCBs) in plastid genomes and determination of gene clusters (GCs)

The progressiveAlignment method, implemented in the MAUVE v.2.3.1 package (Darling et al., 2010), was used with the default parameters to identify locally collinear blocks (LCBs) among the plastid genomes listed in Table 6.1. I excluded *Trifolium* species belonging to the 'refractory clade' (see chapter 5), due to their repetitive nature. In this study, a LCB represents a region within a plastid genome, that almost always occurs in different orientation and/or order among the plastid genomes, but is free from any internal rearrangements (see Darling et al., 2010). These regions are therefore putatively orthologous in nature. The LCBs are represented as colored boxes in Figure 6.2. I was particularly interested in finding LCBs that contained protein coding – and rRNA genes. I used two programs, projectAndStrip and makeBadgerMatrix (downloaded from http://gel.ahabs.wisc.edu/mauve/snapshots/), to generate a LCB boundary file, following this tutorial (http://gel.ahabs.wisc.edu/barphlye/#example). The LCB boundary file contains information on where each LCB starts and ends in each of the plastome analyzed. I used the plastome (*Cicer arietinum)* [NCBI Reference Sequence NC_011163] (Jansen et al., 2008), as the reference plastome. I compared the LCB boundary information with the positions of protein coding – and rRNA genes, using a custom Python script that I wrote, to determine which LCBs contained genes. These LCBs will be referred to from now on, on the basis of their gene content, as gene clusters (GCs). LCBs without any protein coding or rRNA genes were excluded from any further analyses.

6.3 Results

6.3.1 Phylogenetic distribution of scrambled plastomes within the IRLC

A total of 23 plastid genomes were used in the phylogenetic reconstruction of the IRLC legumes studied here (Table 6.1 and Figure 6.1). In order to assess the quality of our assembly methods, we compared the plastome sequences of three species that have an available plastid genome sequence on NCBI's genbank to our assemblies: *Lathyrus sativus, Lens culinaris* and *Pisum sativum.* We observed no major intraspecific plastid rearrangements and the differences observed within species were due to small indels and single nucleotide polymorphisms (data not shown). Figure 6.1 shows a phylogram, which was constructed using maximum likelihood of a concatenated matrix of 70 plastid protein-coding genes. The tree topology is mostly consistent with previous studies (see Figure 4.5, and Figure 5 in Wojciechowski et al., 2004) and the majority of nodes are well supported, with bootstrap support around 85% or higher. There are however exceptions, especially regarding the topology of the North-American *Lathyrus* and the grouping of *L. palustris* (compare with Figure 4.5). Despite this relatively minor phylogenetic incongruence, I feel that the analysis is sufficient to provide a phylogenetic context for the analysis of conserved gene clusters in the plastomes of the IRLC legumes, which is the main focus of this chapter.

I detected major rearrangements among some of the analyzed plastid genomes (see Figure 6.2 and 6.3). Furthermore, these rearrangements were restricted to five genera: *Trifolium, Lens, Vicia, Pisum* and *Lathyrus.* These genera form a monophyletic clade (see Figure 6.1 and Figure 5 in Wojciechowski et al., 2004). Outside this clade, the plastome structure is very stable. No rearrangements were detected among the *Medicago* spp., *Cicer arietinum* (Figure 6.2) and *Glycyrrhiza glabra* [NCBI Refseq NC_024038] (data not shown). The plastomes of *Glycyrrhiza*,

Cicer and *Medicago* were completely collinear to *Lotus japonicus* [NCBI Refseq NC_002694], which is an outgroup species (Figure 6.1), except that they lack the inverted repeat (IR).

6.3.2 Conserved gene clusters among the IRLC legume plastomes

Eighteen out the 21 plastomes analyzed here (see Figure 6.1) are highly rearranged. These structural rearrangements have involved multiple steps of translocations and/or inversions. However, gene content across the plastomes is very similar (data not shown) and their size is relatively constant (Table 6.1), ranging from about 120 to 126 kb. Closely related species tend to have more similar plastome structure than more distantly related species (Figures 6.1, 6.2 and 6.3). For example, *Lathyrus ochroleucus* and *L. venosus*, are sister species and have completely collinear plastomes (Figure 6.3). This is also the case for *L. japonicus* and *L. littoralis* (Figure 6.3) and *Trifolium glanduliferum* and *T. strictum* (Figure 6.2).

MAUVE identified a total of 34 localized collinear blocks (LCB) (Figure 6.2 and 6.3) in the 23 analyzed plastomes. Out of these 34 LCBs, 26 contained protein-coding genes and one LCB was made up of the plastid rRNA genes (see Table 6.2). These 27 gene clusters varied in size and number of genes that they encompass (see Table 6.2). Nine of the clusters contained only a single gene, 5 clusters were composed of 2 genes and the remaining 12 clusters consisted of more than 2 genes. The largest gene cluster (GC) is GC-1, 13.8 kb in length, containing the following genes: *rpl32, ndhF, psbA, matK, rbcL, atpE* and *atpB* (Table 6.2). The smallest gene cluster detected was GC-7, about 1.2 kb in length, containing only a single gene, *petN.* Many of the gene clusters have previously been recognized as plastid operons (i.e. transcriptional units), such as GC-2, 5, 14, 19, 20, 27, 28 and 34 (see Sugita and Sugiura, 1996), which suggests that

the delimitation of these clusters is not random and is under functional constraint (see discussion).

6.4 Discussion

6.4.1 The rearrangements of plastomes in IRLC legumes

Plastid genomes of analyzed *Trifolium* and the Fabeae (*Lens, Vicia, Pisum* and *Lathyrus*) species are highly rearranged, as a result from multiple rounds of translocations and/or inversions (Figure 6.2 and 6.3). These rearrangements have previously been reported (Palmer and Thompson, 1982; Cai et al., 2008; Magee et al., 2010; Sabir et al., 2004). Plastid genomes tend to be quite conserved structurally across land plants (see Wicke et al., 2011). However, besides the legumes there are other well-known exceptions, such as Geraniaceae (Guisinger et al., 2011) and Campanulaceae (Haberle et al., 2008). The plastome rearrangements described here for the Fabeae appear to be most similar to those reported in *Trachelium caeruleum* (Campanulaceae), since they do not involve proliferation of repeated elements, such as in certain *Trifolium* species (see chapter 5) or in the Geraniaceae (Guisinger et al., 2011; Weng et al., 2013). The functional cause of these rearrangements is not known. The stability of plastid genomes is maintained through recombination mechanisms, which are controlled by a large number of nuclear genes (see Maréchal and Brisson 2010). Whatever the functional changes that result from this unprecedented genome instability, it is clear that it offers a unique opportunity for the study the organization of transcriptional units within the plastid genomes of flowering plants.

6.4.2 Do conserved blocks in otherwise rearranged plastomes represent operons?

The gene space of plastid genomes is organized into transcriptional units, similar to operons in the genome of their cyanobacterial ancestors (Stern et al., 2010). However it is important to note that despite being of bacterial origin, most aspects of the regulation of gene expression are radically different in the plastid, mainly due to its interactions with the nuclear genome (Stern et al., 2010). Nevertheless, it is well established from functional studies that many plastid genes are organized into dicistronic or polycistronic operon-like units, i.e. co-regulated gene blocks, also known as transcriptional units (Sugita and Sugiura, 1996). It is therefore reasonable to assume that any structural rearrangements that would break up these transcriptional units would be very detrimental to the plastid and be selected against.

My results are in agreement with that assumption, as many of the gene clusters that I observe are known plastid polycistronic operons (Table 6.2 and Table 2 in Sugita and Sugiura, 1996). Examples of this are: (i) Gene Cluster 17 (GC-17, Table 6.2) that seems to correspond to the *psbB* operon, which has been extensively studied (Stoppel and Meurer, 2013); (ii) GC-6 which contains the same genes as the *psbD/C/Z* operon, which has been characterized in tobacco (Adachi et al., 2012); (iii) GC-24 which contains all the rRNA genes, which are necessary to construct the plastid 70S ribosome. The numerous genes that are not associated with any other and freely translocate independently, are likely to represent single gene transcriuptional units, i.e. monocistronic operons. Gene Cluster 1 was very interesting, since it contains seven cistrons that previously were thought to be transcribed independently (i.e. as monocistronic units) or belong to different operons (see Table 2 in Sugita and Sugiura, 1996). My results are very suggestive that that GC-1 is a conserved plastid operon, at least in the legume species analyzed here. Six LCBs without any annotated protein-coding – or rRNA genes were also identified. They varied

in size, smallest being around 150 nt and the largest was 1.3 kb. It is possible that some of these apparently empty LCBs contain unrecognized functional elements or even unannotated protein sequences.

These results demonstrate that identification of conserved gene clusters in this clade of rapid structural evolution is a powerful way of provide evidence for previously described plastid operons and potentially to find new ones. Such is the extent of the genic reorganization in the sampled species that it may be argued that the persistence of multiple intact gene blocks is implausible unless these units (Table 6.2) are inviolable as they represent the fundamental regulatory architecture of the legume plastid.

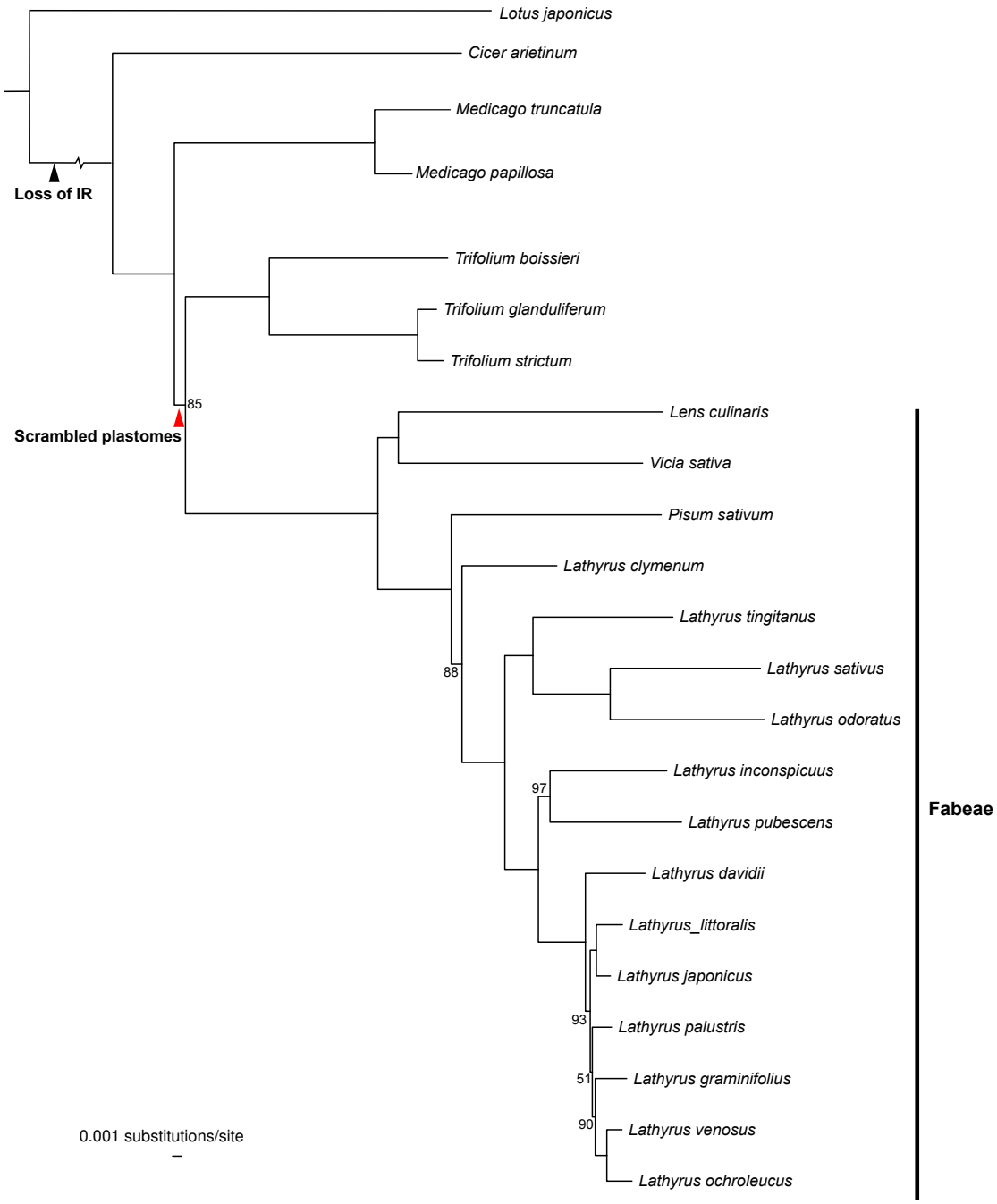


Figure 6.1 A phylogram showing relationships among the analyzed IRLC legume species. The phylogeny was generated using maximum likelihood from a concatenated matrix of 70 protein coding plastid genes with a

combined aligned length of 62,525 characters. The black triangle marks the loss of the inverted repeat and the red triangle marks the evolutionary origin of highly rearranged plastomes (i.e. scrambled, see results). When bootstrap support was lower then 100%, it written next to the corresponding node.

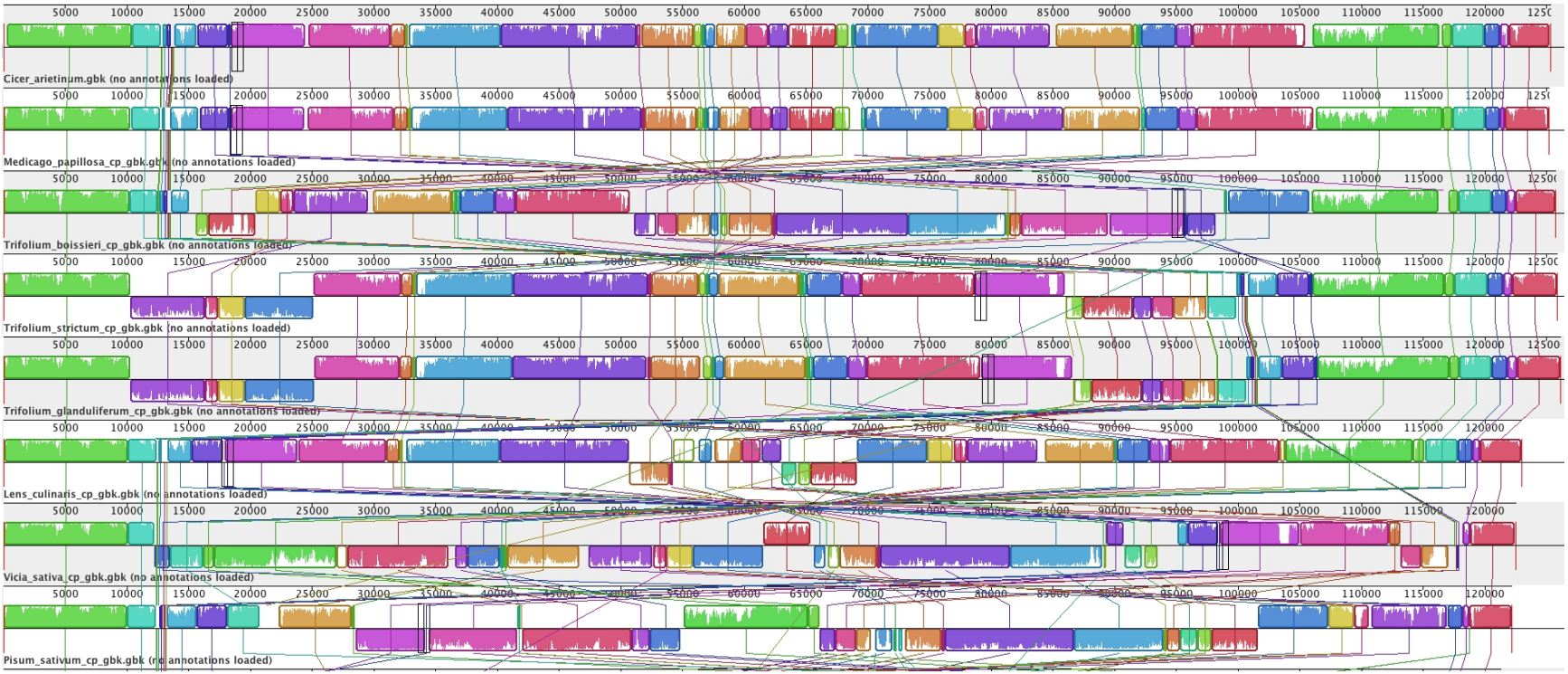


Figure 6.2 A MAUVE alignment showing the boundaries of plastid identified locally collinear blocks (LCBs) and their rearrangements. The species in this figure are (from top to bottom): *Cicer arietinum* (top), *Medicago papillosa, Trifolium boissieri, Trifolium strictum, Trifolium glanduilferum, Lens culinaris, Vicia sativa* and *Pisum sativum* (bottom). Each row represents a different plastid genome and the LCBs are color-coded.

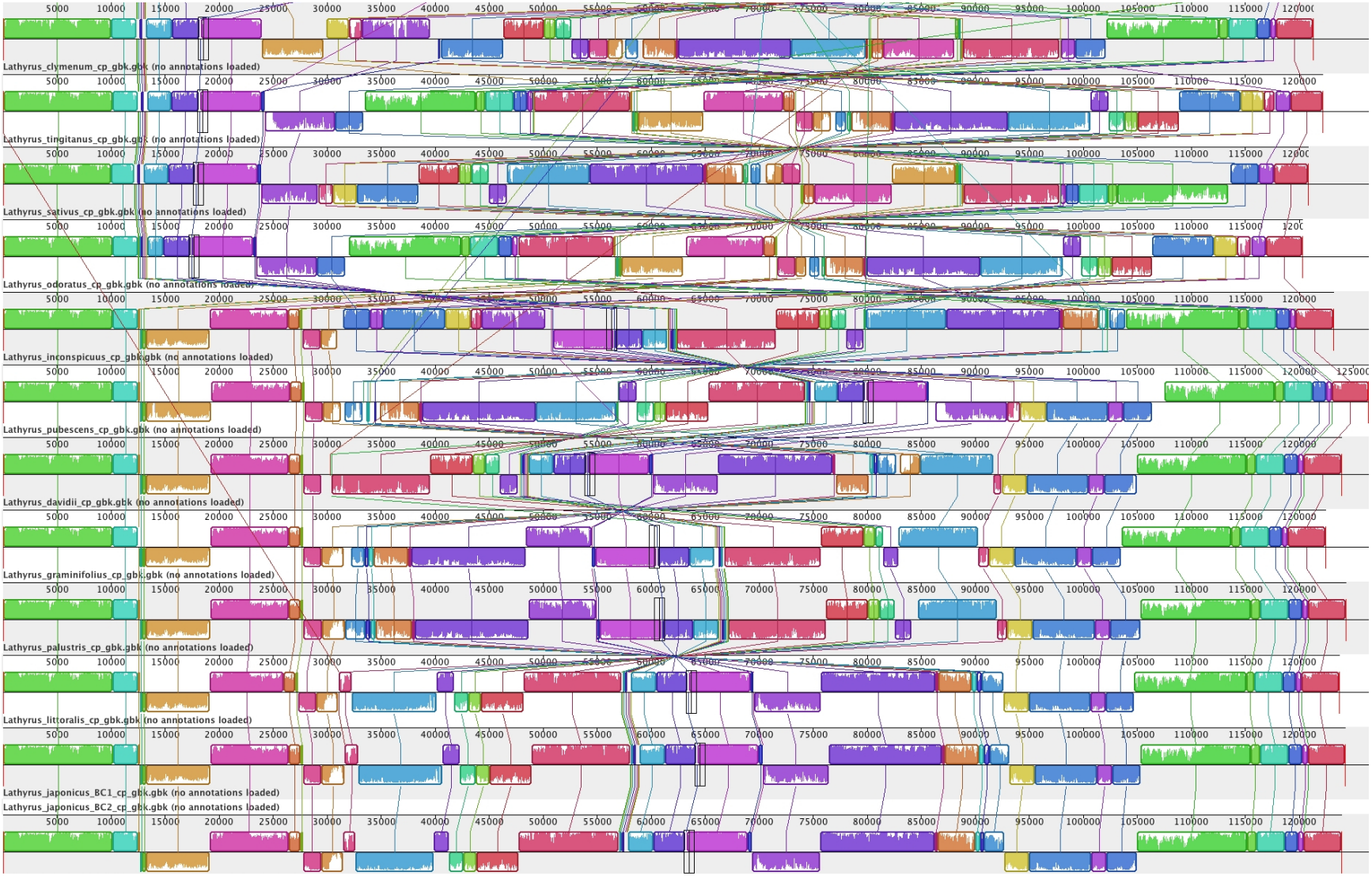

Lathyrus_clymenum_cp_gbk.gbk (no annotations loaded)
Lathyrus_tingitanus_cp_gbk.gbk (no annotations loaded)
Lathyrus_sativus_cp_gbk.gbk (no annotations loaded)
Lathyrus_odoratus_cp_gbk.gbk (no annotations loaded)
Lathyrus_inconspicuus_cp_gbk.gbk (no annotations loaded)
Lathyrus_pubescens_cp_gbk.gbk (no annotations loaded)
Lathyrus_davidii_cp_gbk.gbk (no annotations loaded)
Lathyrus_graminifolius_cp_gbk.gbk (no annotations loaded)
Lathyrus_palustris_cp_gbk.gbk (no annotations loaded)
Lathyrus_littoralis_cp_gbk.gbk (no annotations loaded)
Lathyrus_japonicus_BC1_cp_gbk.gbk (no annotations loaded)
Lathyrus_japonicus_BC2_cp_gbk.gbk (no annotations loaded)

Figure 6.3 A MAUVE alignment showing the boundaries of plastid identified locally collinear blocks (LCBs) and their rearrangements. The species in this figure are (from top to bottom): ***Cicer arietinum*** **(top),** ***Lathyrus clymenum, Lathyrus tingitanus, Lathyrus sativus, Lathyrus odoratus, Lathyrus inconspicuus, Lathyrus pubescens, Lathyrus davidii, Lathyrus graminifolius, Lathyrus palustris, Lathyrus littoralis, Lathyrus japonicus, Lathyrus ochroleucus*** **and** ***Lathyrus venosus*** **(bottom). Note that the plastomes of** ***L. palustris*** **and** ***L. graminifolius*** **are collinear as are the plastomes of:** ***Lathyrus japonicus, L. littoralis, L. ochroleucus, L. venosus.*** **Each column represents a different plastid genome and the LCBs are color-coded. White areas within each LCB represent a drop in nucleotide sequence similarity.**

Species **Seed accessions (germplasm)**	**Plastome length (nt)**	**NCBI accession**	**Herbarium voucher**
Cicer arietinum L. NA (NA)	125,319	NC_011163	SS14-13
Medicago truncatula Gaertn. NA (NA)	124,033	NC_003119	NA
Medicago papillosa Boiss. PI 631778 (USDA)	125,358	KJ850240*	SS14-24 (UBC)
Trifolium boissieri Guss. PI 369022 (USDA)	125,741	KJ788284*	SS14-01 (UBC)
Trifolium strictum L. PI 369147 (USDA)	125,835	KJ788292*	SS14-02 (UBC)
Trifolium glanduliferum Boiss. PI 296666 (USDA)	126,182	KJ788285*	SS14-03 (UBC)
Pisum sativum L. W6 32866 (USDA)	122,150	KJ806203*	SS14-11 (UBC)
Lens culinaris Medik. PI 592998 (USDA)	123,129	KJ850239*	NA
Vicia sativa L. PI 293436 (USDA)	122,610	KJ850242*	NA
Lathyrus clymenum L. NA (RP**)	121,263	KJ850235*	SS14-14 (UBC)
Lathyrus tingitanus L. NA (RP**)	122,323	KJ850238*	SS14-15 (UBC)
Lathyrus odoratus L. NA (RP**)	120,453	KJ850237*	SS14-17 (UBC)
Lathyrus inconspicuus L. W6 2817 (USDA)	123,314	KJ850236*	SS14-18 (UBC)

Species **Seed accessions (germplasm)**	**Plastome length (nt)**	**NCBI accession**	**Herbarium voucher**
Lathyrus pubescens Hook. & Arn. NA (RP**)	126,422	KJ806200*	SS14-19 (UBC)
Lathyrus davidii Hance NA (RP**)	123,896	KJ806192*	SS14-20 (UBC)
Lathyrus palustris L. NA (NA) [1]	124,288	KJ806198*	SS14-25 (UBC)
Lathyrus japonicus Willd. NA (NA) [1]	124,243	KJ806194*	NA
Lathyrus littoralis Endl. NA (NA) [1]	123,735	KJ806197*	NA
Lathyrus graminifolius (S.Watson) T.G.White DLP*** accession: 920239	122,439	KJ806193*	SS14-21 (UBC)
Lathyrus ochroleucus Hook. NA (NA)[1]	124,263	KJ806199*	SS14-22 (UBC)
Lathyrus venosus Muhl. ex Willd. NA (NA)[1]	125,460	KJ806202*	SS14-23 (UBC)

Table 6.1 Summary Illumina sequencing accessions, plastome assembly – and voucher information. An asterisk marks plastome sequences newly reported in this thesis. **RP: Roger Parsons Sweet Peas. ***DLP: Desert Legume Project. [1]See table B.1 in appendix B.

Gene cluster	Genes	NO. Cistrons	GC bondary	GC length
GC-1	*rpl32-ndhF-psbA-matK–rbcL-atpE-atpB*	7	121,929- 125,304 and 1 -10,451	13,825
GC-2	*ndhC-ndhK-ndhJ*	3	10,452 - 12,791	2,339
GC-3	*rps4*	1	13,662 - 15,494	1,832
GC-4	*ycf3*	1	15495 - 18,263	2,768
GC-5	*psaA-psaB-rps14*	3	18,414 - 24,580	6,166
GC-6	*psbD-psbC-psbZ-psbM*	4	24,619 - 31,395	6,776
GC-7	*petN*	1	31,396 - 32,546	1,150
GC-8	*rpoB-rpoC1*	2	33,206 - 40,301	7,095
GC-9	*atpH-atpI-atpF-atpA-rps2-rpoC2*	6	40,302 - 51,330	11,028
GC-10	*psbK-psbI*	2	51,726 - 55,352	3,626
GC-11	*accD*	1	56,160 - 57,657	1,497
GC-12	*cemA-psaI*	2	57,658 - 60,174	2,516
GC-13	*petA*	1	60,175 - 61,943	1,768
GC-14	*psbE-psbF-psbL-psbJ*	4	61,944 - 63,554	1,610
GC-15	*psaJ-rps18-petG-petL-rpl33-rpl20*	6	63,840 - 67,444	3,604
GC-16	*clpP-5'rps12*	2	67,445 - 68,893	1,448
GC-17	*psbB- psbH-petB-petD-psbT-psbN*	6	69,978 - 75,728	5,750
GC-18	*rpl36-rps11-rpoA*	3	75,729 - 77,929	2,200
GC-19	*rps8*	1	77,943 - 79,069	1,126
GC-20	*rpl23-rps19-rps3-rpl2-rpl16-rpl14*	6	79,070 - 84,793	5,723
GC-21	*ycf2*	1	85,625 - 91,513	5,888
GC-22	*ndhB*	1	92,131 - 94,758	2,627
GC-23	3'*rps12-rps7*	2	94,759 - 96,271	1,512
GC-24	*16SrDNA-23SrDNA-4.5SrDNA-5SrDNA*	4	96,310 - 104,698	8,388
GC-25	*ndhA-ndhI-ndhG-ndhH-rps15-ycf1*	6	10,6060 - 117,315	11,255

Gene cluster	Genes	NO. Cistrons	GC bondary	GC length
GC-26	*psaC-ndhE-ndhD*	3	117,316 - 119,927	2,611
GC-27	*ccsA*	1	119,928 - 121,207	1,279

Table 6.2 Gene cluster identified in the locally collinear blocks from the MAUVE alignment. The table lists the genes in each gene cluster in addition to the boundary and length of the gene cluster, in the *Cicer arietinum* plastome.

Chapter 7: Conclusions

The unifying theme of this thesis is the investigation of various aspects of plant genome evolution using massively parallel sequencing. The thesis focused on (i) the study of TE variation using WGSS, (ii) characterization of paleopolyploidy events using transcriptomics, (iii) determining the evolutionary origin of a polyploid species and (iv) plastid genome evolution within the IRLC legumes. The MPS method of choice was Illumina, and it was used in every chapter of the thesis. I furthermore used phylogenetics to frame the evolutionary questions that I put forward in each chapter. I developed two novel phylogenetic computational pipelines, T2Phy and Plast2Phy, which proved very useful in resolving phylogenetic relationships among my species of interest. I plan to make these pipelines publicly available and they could be very useful to other researchers.

7.1 Summary

The study of intra- and interspecific variation of TEs in *Theobroma cacao* and related species (chapter 2), demonstrates the usefulness of MPS sequencing in the study of repetitive elements within plant genomes. This observation has also been made in other plant systems, such as in *Asparagus* (Li et al., 2014), *Pisum* (Macas et al., 2007) and *Zea* (Tenaillon et al., 2011). I detected considerable differences in transposable element composition among - and within species, highlighting their dynamic role in plant genome evolution. Variation of transposable elements in plants is important especially given the great abundance of transposable elements in plant genomes and their potential impact on the genespace (Craig et al., 2002; Zhou et al., 2004; Martin et al., 2005). We used two different methods of looking at transposable element variation from Illumina short read data: reference-based mapping and graph-based clustering (Novak et

al., 2010). Both are effective at capturing variation, although each is appropriate at different levels of taxonomic comparison. Reference based mapping works well within a species while graph-based clustering is preferred for between species comparisons.

The discovery of a paleopolyploidy event with the blue-flowered *Linum* clade, described in chapter 3, will undoubtedly be very important for future flax (*Linum usitatissimum*) research. More broadly, these findings demonstrate the limitations of relying solely on the results of paralog age distribution plots from a single species in the inference of whole-genome duplication events in its evolutionary history. Especially since the paleopolyploidy event described here, was undetected in the analysis of the paralog age distribution plots in the flax genome paper (Wang et al., 2012). I argue that an analysis of Ks distributions from multiple species, when combined with phylogenetic reconstruction of paralogues and good taxon sampling, is a very powerful approach for discovering and characterizing paleopolyploidy events.

The polyploid origin of the North-American species *Lathyrus venosus* has been of considerable interest, where both hybrid (Gutiérrez et al., 1994) and non-hybrid (Khawaja et al., 1997) origin have been proposed. I found Illumina MPS very useful to reconstruct the complete plastomes and the ribosomal regions of several *Lathyrus* species. In addition I was able to retrieve considerable number of exomic SNPs, which I analyzed using genetic distance methods. Those regions were found highly informative for phylogenetic reconstruction of the studied *Lathyrus* species. *L. venosus* is closely related to *L. ochroleucus* on evidence from both nuclear and plastome sequences. There is no evidence of close relationship with *L. palustris* and no dual genomic signal that might indicate allopolyploidy. The evidence so far supports *L. venosus* as an autopolyploid species very closely related to the diploid *L. ochroleucus*. My results demonstrate

the usefulness of Illumina MPS data for reconstructing phylogenetic relationships among closely related species and for testing hypothesis regarding evolutionary origin of polyploid species.

In chapter 5 I showed that the *Trifolium subterraneum*-type plastomes, i.e. containing large amounts of repetitive DNA (described in Cai et al., 2008), is phylogenetically restricted to a single core clade of *Trifolium* ("the refractory clade"), comprising five of the eight recognized sections within the genus. Furthermore it is ubiquitously present in all members of this clade in our sample. It is thus reasonable to suppose that this may be the characteristic plastome of this core clade. *Trifolium* is a large genus of c. 300 species, and these five sections contain slightly over 200 species, or about 70% of the genus. An independent study supports these findings (Sabir et al., 2014). There are therefore likely to be abundant exemplars of this unusual plastome type, at varying degrees of evolutionary relatedness, available for future functional and evolutionary studies.

In chapter 6 I confirmed that the plastid genomes of *Trifolium* and the Fabeae (*Lens, Vicia, Pisum* and *Lathyrus*) species are highly rearranged, as a result from multiple rounds of translocations and/or inversions. Whatever the functional changes that result from this unprecedented genome instability, it is clear that it offers a unique opportunity for the study the organization of transcriptional units within the plastid genomes of flowering plants. It is reasonable to assume that any structural rearrangements that would break up the operon-like structures within the plastome would be very detrimental to the plastid and be selected against. My results are in agreement with this assumption, as many of the gene clusters that I observe are known plastid polycistronic operons (Table 6.2 and Table 2 in Sugita and Sugiura 1996). I also observed putative conserved polycistronic operons (GC-1 in Table 6.2), that previously were thought to be transcribed independently (i.e. as monocistronic units). These results demonstrate

that identification of conserved gene clusters in this clade of rapid structural evolution is a powerful way of validating previously described plastid operons and potentially to identify new ones.

7.2 Conclusions

On the whole, I am very optimistic about the future work in the field of plant genome evolution and evolutionary plant systematics. I feel that the two most important consequences of increased availability MPS are: (i) it enables researchers to ask important evolutionary questions in plant groups that do not have any prior genomic resources and (ii) without having highly specialized laboratory facilities available. The work presented in this thesis demonstrates these points in several ways. In chapter 2 I successfully analyzed TE composition of *Theobroma cacao* species, using simple Illumina WGSS techniques. In chapter 3 I discovered a paleopolyploidy event occurring within the flax genus (*Linum*), using straight up Illumina transcriptomics. It was somewhat helpful to have a fully sequenced genome of cultivated flax (*L. usitatissimum*), but it was not in any way essential. Much more important was the phylogenetically broad sampling of *Linum* species and careful data analysis. I was also able to obtain a well-supported phylogeny from the transcriptome assemblies. Similar type of analysis could be done in any genus, where a paleopolyploidy event was suspected or a high-resolution phylogeny was desired. In chapter 4 I put forward convincing evidence that *Lathyrus venosus* is of autopolyploid origin, based on phylogenetic analysis of plastid – and ribosomal sequences. I was also able to extract and analyze substantial numbers of exomic SNPs. These analyses were performed on genomic data that I acquired and analyzed during my thesis research and did not rely on external genomic resources. In chapter 5 and 6, I analyzed highly unusual plastomes in two tribes of the IRLC

legumes, Trifolieae and Fabeae. I was able to determine the evolutionary origin of repetitive plastomes within the clover genus (*Trifolium*) (chapter 5) as well providing evidence for the link between plastid operons and rearrangements within the plastid genomes of *Trifolium, Lens, Vicia, Pisum* and *Lathyrus* (chapter 6).

7.3 Future directions

The potential for use of MPS technology is vast. If sequencing costs continue to fall it may become possible to apply whole genome technology to complete biomes or floras. As an example I will use the flora of Iceland (where I am from and did my masters research). There are many unanswered questions regarding the origin of the current biodiversity of Iceland, especially plants (Aegisdottir, 2003; Aegisdottir and Thórhallsdóttir, 2004), which could be tackled by MPS approaches. The island itself is about 16 million years old, but Iceland has experienced multiple glaciation events and the most recent one ended somewhere around 11,000 years ago (Ingólfsson, 1991). There is no evidence for any plant species surviving these glaciation events (Aegisdóttir and Thórhallsdóttir, 2004), which means that the entire flora of Iceland is younger than 11,000 years old. Despite several biogeographical studies (Anamthawat-Jónsson et al., 1999; Aegisdóttir, 2003; Thórsson et al., 2010), it is still unclear where the plant species came from and what adaptive and genomic changes occurred after colonization. Studies using molecular markers have been applied in some plant genera (Aegisdóttir and Thórhallsdóttir, 2004; Aegisdóttir and Thórhallsdóttir, 2006), especially birch (*Betula*) (Anamthawat-Jónsson et al., 2010; Thórsson et al., 2010). These studies have revealed that *Betula* species in Iceland came from Northern Scandinavia (Thórsson et al., 2010). The methods I used in my thesis are in many ways ideal to determine the geographical origin of the Icelandic flora. The flora is relatively

species poor, comprised only about 480 indigenous species of vascular plants (Ægisdóttir and Thórhallsdóttir, 2004). That means that it would be relatively straightforward to get a comprehensive picture of the origin of Icelandic flora, by sequencing every species using any of the available MPS methods. For instance, even if plants were sequenced at a low depth, it would still be possible to assemble plastid genomes and compare their sequences to material collected outside of Iceland, allowing flora-scale phylogeography at an unprecedented data-richness.

Bibliography

Adachi Y, Kuroda H, Yukawa Y, Sugiura M. **2012**. Translation of partially overlapping *psbD-psbC* mRNAs in chloroplasts: the role of 5'-processing and translational coupling. *Nucleic Acids Research* **40**: 3152–8.

Adams KL, Wendel JF. **2005**. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**: 135–141.

Aegisdóttir HH. 2003. *Reproductive ecology, morphological- and genetic variation in* Campanula uniflora *in Iceland, Greenland and Svalbard)*. MSc. Thesis, Unviersity of Iceland in Reykjavik, Iceland.

Aegisdóttir HH, Thórhallsdóttir ÞE. 2004. Theories on migration and history of the North-Atlantic flora: a review. *Jökull* **54**: 1-16.

Aegisdóttir HH, Thórhallsdóttir TE. 2006. Breeding System Evolution in the Arctic: a Comparative Study of *Campanula uniflora* in Greenland and Iceland. *Arctic, Antarctic, and Alpine Research* 38: 305-312.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Anamthawat-Jónsson K, Bragason BT, Bödvarsdóttir SK, Koebner RM. **1999**. Molecular variation in *Leymus* species and populations. *Molecular Ecology* **8**: 309–315.

Anamthawat-Jónsson K, Thórsson *Æ*T, Temsch EM, Greilhuber J. **2010**. Icelandic Birch Polyploids—The Case of a Perfect Fit in Genome Size. *Journal of Botany* **2010**: 1–9.

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN. **2010**. The genome of *Theobroma cacao*. *Nature Genetics* **43**: 101–108.

Asmussen C, Liston A. **1998**. Chloroplast DNA characters, phylogeny, and classification of *Lathyrus* (Fabaceae). *American Journal of Botany* **85**: 387 – 401.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. **2008**. Multiple paleopolyploidizations during the evolution of the compositae reveal parallel Patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**: 2445–2455.

Barker MS, Vogel H, Schranz ME. **2009**. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* **1**: 391–399.

Bena G, Olivieri I, Lejeune B, Jubier M-F. **1998**. Ribosomal external and internal transcribed spacers: combined use in the phylogenetic analysis of *Medicago* (Leguminosae). *Journal of Molecular Evolution* **46**: 299–306.

Bennett MD, Smith JB. **1976**. Nuclear DNA amounts in angiosperms. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **274**: 227–274.

Bennett MD, Letich IJ. 2012. Plant DNA C-values database (release 6.0, Dec. 2012) http://www.kew.org/cvalues/.

Birney E, Clamp M, Durbin R. **2004**. GeneWise and Genomewise. *Genome Research* **14**: 988–995.

Blanc G, Wolfe KH. **2004**. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* **16**: 1667–1678.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. **2008**. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **200:** P10008.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. *Cell and Molecular Biology of Plastids*. Berlin, Heidelberg: Springer, 29–63.

Bock DG, Kane NC, Ebert DP, Rieseberg LH. **2014**. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *The New Phytologist* **201**: 1021–1030.

Boeke JD, Corces VG. **1989**. Transcription and reverse transcription of retrotransposons. *Annual Review of Microbiology* **43**: 403–434.

Bouckaert RR. **2010**. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**: 1372–1373.

Bouxin G. **2005**. Ginkgo, a multivariate analysis package. *Journal of Vegetation Science* **16**: 355–359.

Broich SL. **1989**. Chromosome numbers of North American *Lathyrus* (Fabaceae). *Madrono* **36**: 41–48.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. **2012**. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* **29**: 1917–1932.

Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl J V, Boore J, Jansen RK. **2008**. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *Journal of Molecular Evolution* **67**: 696–704.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Cerutti H, Osman M, Grandoni P, Jagendorf AT. **1992**. A homolog of *Escherichia coli* RecA protein in plastids of higher plants. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 8068–8072.

Cerutti H, Jagendorf AT. **1993**. DNA strand-transfer activity in pea (*Pisum sativum* L.) Chloroplasts. *Plant Physiology* **102**: 145–153.

Chalup L, Grabiele M, Solís Neffa V, Seijo G. **2012**. Structural karyotypic variability and polyploidy in natural populations of the South American *Lathyrus nervosus* Lam. (Fabaceae). *Plant Systematics and Evolution* **298**: 761–773.

Chaparro C, Sabot F. **2012**. Methods and software in NGS for TE analysis. *Methods in Molecular Biology* **859**: 105-114.

Chaudhuri P, Marron JS. **1999**. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**: 807–823.

Cheesman EE. **1944**. Notes on the nomenclature, classification and possible relationships of cocoa populations. *Tropical Agriculture* **21**: 144–159.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. **2006**. The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution* **23**: 2175–2190.

Conant GC, Wolfe KH. **2008**. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**: 861–862.

Cosner ME, Jansen RK, Palmer JD, Downie SR. **1997**. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Current Genetics* **31**: 419–429.

Craig NL, Craigie R, Gellert M, Lambowitz AM. **2002**. *Mobile DNA II*. Washington, DC: American Society for Microbiology.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW. **2006**. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**: 738–749.

Darling AE, Mau B, Perna NT. **2010**. Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**: e11147.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**: 772.

DeGiorgio M, Syring J, Eckert AJ, Liston A, Cronn R, Neale DB, Rosenberg NA. **2014**. An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. *BMC Evolutionary Biology* **14**: 67.

Dolezel J, Greilhuber J, Suda J. **2007**. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2**: 2233–2244.

Doyle JJ, Doyle JL. **1987**. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Doyle JJ, Egan AN. **2010**. Dating the origins of polyploidy events. *New Phytologist* **186**: 73–85.

Drew BT, Ruhfel BR, Smith S a, Moore MJ, Briggs BG, Gitzendanner M a, Soltis PS, Soltis DE. **2014**. Another Look at the Root of the Angiosperms Reveals a Familiar Tale. *Systematic Biology* **63**: 368–382.

Drummond AJ, Suchard M a, Xie D, Rambaut A. **2012**. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**: 1969–73.

Edgar RC. **2004**. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Ellinghaus D, Kurtz S, Willhoeft U. **2008**. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Ellis TH, Lee D, Thomas CM, Simpson PR, Cleary WG, Newman MA, Burcham KW. **1988**. 5S rRNA genes in *Pisum*: sequence, long range and chromosomal organization. *Molecular & General Genetics* **214**: 333–342.

Ellison NW, Liston A, Steiner JJ, Williams WM, Taylor NL. **2006**. Molecular phylogenetics of the clover genus (*Trifolium* - Leguminosae). *Molecular Phylogenetics and Evolution* **39**: 688–705.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. **2010**. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 16196–16200.

Eserman L a, Tiley GP, Jarret RL, Leebens-Mack JH, Miller RE. **2013**. Phylogenetics and diversification of morning glories (tribe Ipomoeeae, Convolvulaceae) based on whole plastome sequences. *American Journal of Botany* **101**: 92–103.

Felsenstein J. **1973**. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**: 240–249.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Feschotte C, Pritham EJ. **2007**. DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* **41**: 331–368.

Feulgren R, Rossenbeck H. **1924**. Mikroskopisch-chemischer Nachweis einer Nucleinsäure vom Typus der Thymonucleinsäure und die- darauf beruhende elektive Färbung von Zellkernen in mikroskopischen Präparaten. *Hoppe-Seyler´s Zeitschrift für physiologische Chemie* **135**: 203–248.

Figueira A, Janick J, Goldsbrough P. **1992**. Genome size and DNA polymorphism in *Theobroma cacao*. *Journal of the American Society for Horticultural Science* **117**: 673–677.

Flavell AJ, Smith DB, Kumar A. **1992**. Extreme heterogeneity of Ty1-copia group retrotransposons in plants. *Molecular and General Genetics* **231**: 233–242.

Gabriel A, Willems M, Mules EH, Boeke JD. **1996**. Replication infidelity during a single cycle of Ty1 retrotransposition. *Proceedings of the National Academy of Sciences* **93**: 7767–7771.

Galián JA, Rosato M, Rosselló JA. **2012**. Early evolutionary colocalization of the nuclear ribosomal 5S and 45S gene families in seed plants: evidence from the living fossil gymnosperm *Ginkgo biloba*. *Heredity* **108**: 640–646.

Ghulam MM, Courtois F, Lerbs-Mache S, Merendino L. **2013**. Complex processing patterns of mRNAs of the large ATP synthase operon in *Arabidopsis* chloroplasts. *PLoS ONE* **8**: e78265.

Goldman N, Yang ZH. **1994**. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Molecular Biology and Evolution* **11**: 725–736.

Goldman N, Anderson JP, Rodrigo AG. **2000**. Likelihood-based tests of topologies in phylogenetics. *Systematic biology* **49**: 652–670.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Green AG. **1986**. Genetic control of polyunsaturated fatty acid biosynthesis in flax (*Linum usitatissimum*) seed oil. *Theoretical and Applied Genetics* **72**: 654-661.

Guerra M. 2008. Chromosome numbers in plant cytotaxonomy: concepts and implications. *Cytogenetic and Genome Research* **120**: 339-350

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696–704.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. **2010**. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–21.

Guisinger MM, Kuehl J V, Boore JL, Jansen RK. **2011**. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* **28**: 583–600.

Gurdon C, Maliga P. **2014**. Two distinct plastid genome configurations and unprecedented intraspecies length variation in the *accD* coding region in *Medicago truncatula*. *DNA Research* 1–11. doi: 10.1093/dnares/dsu007

Gutiérrez JF, Vaquero F, Vences FJ. **1994**. Allopolyploid vs. autopolyploid origins in the genus *Lathyrus* (Leguminosae). *Heredity* **73**: 29–40.

Haberle RC, Fourcade HM, Boore JL, Jansen RK. **2008**. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *Journal of Molecular Evolution* **66**: 350–361.

Hamlin JAP, Arnold ML. **2014**. Determining population structure and hybridization for two iris species. *Ecology and Evolution* **4**: 743–755.

Hammett KRW. **1989**. *Lathyrus chloranthus* X *Lathyrus chrysanthus* a new interspecific hybrid. *Botanical Gazette* **150**: 469–476.

Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nature Biotechnology* **30**: 798–802.

Harris EH, Boynton JE, Gillham NW. **1994**. Chloroplast ribosomes and protein synthesis. *Microbiological Reviews* **58**: 700–754.

Harrison N, Kidner CA. **2011**. Next-generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you? *Taxon* **60**: 1552–1566.

Hedges SB, Blair JE, Venturi ML, Shoe JL. **2004**. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology* **4**: 2.

Heled J, Drummond AJ. **2010**. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**: 570–580.

Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J. **2010**. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biology* **10**: 204.

Huang X, Lu G, Zhao Q, Liu X, Han B. **2008**. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiology* **148**: 25–40.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Huson DH and Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molelcular Biology and Evolution* **23**: 254-267.

Huson DH, Scornavacca C. **2012**. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* **61**: 1061-1067.

Ingólfsson Ó. 1991. A review of the Late Weichselian and Early Holocene glacial and enviromental enviromental history of Iceland. In: Maizels JK, Caseldine C, eds. *Enviromental Cange in Iceland: Past and Present.* Netherlands: Kluwer, 13-29.

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. **2008**. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Molecular Phylogenetics and Evolution* **48**: 1204–1217.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. **2011**. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, Wu X, Zhang Yong, Wang J, Zhang Yeting, Carpenter EJ, Deyholos MK, Kutchan TM, Chanderbali AS, Soltis PS, Stevenson DW, McCombie R, Pires JC, Wong GK-S, Soltis DE, Depamphilis CW. **2012**. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13**: R3.

Johnson BR, Borowiec ML, Chiu JC, Lee EK, Atallah J, Ward PS. **2013**. Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology* **23**: 2058–2062.

Johnson MTJ, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, Chase MW, Clarke ND, Covshoff S, Depamphilis CW, Edger PP, Goh F, Graham S, Greiner S, Hibberd JM, Jordon-Thaden I, Kutchan TM, Leebens-Mack J, Melkonian M, Miles N, Myburg H, Patterson J, Pires JC, Ralph P, Rolf M, Sage RF, Soltis D, Soltis P, Stevenson D, Stewart CN, Surek B, Thomsen CJM, Villarreal JC, Wu X, Zhang Y, Deyholos MK, Wong GK-S. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PloS ONE* **7**: e50226.

Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. **2005**. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**: 462–467.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. **2000**. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceeding of the National Academy of Science* **97**: 6603–6607.

Kane NC, Cronk Q. **2008**. Botany without borders: barcoding in focus. *Molecular Ecology* **17**: 5175–5176.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. **2012**. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* **99**: 320–329.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kelly LJ, Leitch IJ. **2011**. Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research* **19**: 1–15.

Kenicer GJ, Kajita T, Pennington RT, Murata J. **2005**. Systematics and biogeography of *Lathyrus* (Leguminosae) based on internal transcribed spacer and cpDNA sequence data. *American Journal of Botany* **92**: 1199–1209.

Kenicer G. **2008**. An Introduction to the Genus *Lathyrus* L. *Curtis's Botanical Magazine* **25**: 286–295.

Khawaja HI, Ellis JR, Sybenga J. **1995**. Cytogenetics of *Lathyrus palustris*, a natural autohexaploid. *Genome* **38**: 827–831.

Khawaja HI, Sybenga J, Ellis JR. **1997**. Chromosome pairing and chiasma formation in autopolyploids of different *Lathyrus* species. *Genome* **40**: 937–944.

Koch MA, Haubold B, Mitchell-Olds T. **2000**. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* **17**: 1483–1498.

Kode V, Mudd EA, Iamtham S, Day A. **2005**. The tobacco plastid *accD* gene is essential and is required for leaf development. *The Plant Journal* **44**: 237–244.

Kozik A, Matvienko M, Kozik I, Van Leeuwen H, Van Deynze A, Michelmore R. **2008**. Eukaryotic ultra conserved orthologs and estimation of gene capture In EST libraries [abstract]. *Plant and Animal Genomes Conference* **16**: P6.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. **2009**. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**: 1639–1645.

Kumar A, Bennetzen JL. **1999**. Plant retrotransposons. *Annual Review of Genetics* **33**: 479–532.

Kumar A, Hirochika H. **2001**. Applications of retrotransposons as genetic tools in plant biology. *Trends in Plant Science* **6**: 127–134.

Kupicha FK. **1983**. The infrageneric structure of *Lathyrus*. *Notes from the Royal Botanic Garden Edinburgh* **41**: 209–244.

Kuzoff RK, Sweere JA, Soltis DE, Soltis PS, Zimmer EA. **1998**. The phylogenetic potential of entire 26S rDNA sequences in plants. *Molecular Biology and Evolution* **15**: 251–263.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. **2007**. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**: 3100–3108.

Lambert SM, Geneva AJ, Luke Mahler D, Glor RE. **2013**. Using genomic data to revisit an early example of reproductive character displacement in Haitian *Anolis* lizards. *Molecular Ecology* **22**: 3981–95.

Lavin M, Herendeen PS, Wojciechowski MF. **2005**. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology* **54**: 575–594.

Lemmon EM, Lemmon AR. **2013**. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **44**: 99–121.

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

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

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178–2189.

Li S-F, Gao W-J, Zhao X-P, Dong T-Y, Deng C-L, Lu L-D. **2014**. Analysis of Transposable Elements in the Genome of Asparagus officinalis from High Coverage Sequence Data. *PloS ONE* **9**: e97189.

Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**: 282–283.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Lilly JW, Havey MJ, Jackson SA, Jiang J. **2001**. Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants. *The Plant Cell* **13**: 245–254.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards S V. 2009a. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution* **53**: 320–328.

Liu L, Yu L, Pearl DK, Edwards SV. 2009b. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* **58**: 468–477.

Liu L, Yu L. 2010. Phybase: an R package for species tree analysis. *Bioinformatics* **26**: 962–963.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. **2012**. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* **40**: W622–627.

LPGW (The Legume Phylogeny Working Group). 2013. Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* **62**: 217–248.

Lynch M, Conery JS. **2000**. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Ma B, Tromp J, Li M. 2002. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**: 440–445.

Macas J, Neumann P, Navratilova A. **2007**. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* **8**: 427.

Mackay J, Dean JFD, Plomion C, Peterson DG, Cánovas FM, Pavy N, Ingvarsson PK, Savolainen O, Guevara MÁ, Fluch S, Vinceti B, Abarca D, Díaz-Sala C, Cervera M-T. **2012**. Towards decoding the conifer giga-genome. *Plant Molecular Biology* **80**: 555-569.

Mackenzie S, McIntosh L. **1999**. Higher plant mitochondria. *The Plant Cell Online* **11**: 571–585.

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, Gray JC, Kavanagh TA, Wolfe KH. **2010**. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* **20**: 1700–1710.

Maréchal A, Parent J-S, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. **2009**. Whirly proteins maintain plastid genome stability in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 14693–14698.

Maréchal A, Brisson N. **2010**. Recombination and the maintenance of plant organelle genome stability. *The New Phytologist* **186**: 299–317.

Marcais G and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764-770.

Margulis L. 1979. Origin of Eukaryotic Cells. New Haven, CT, USA: Yale University Press.

Marie D, Brown SC. **1993**. A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biology of the Cell* **78**: 41–51.

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A. **2009**. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **461**: 1135–1138.

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. **2002**. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *The Plant Cell* **14**: 2659–2679.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. **2013**. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**: 526–538.

McDill J, Repplinger M, Simpson BB, Kadereit JW. **2009**. The Phylogeny of *Linum* and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. *Systematic Botany* **34**: 386–405.

McDill J, Simpson BB. 2011. Molecular phylogenetics of Linaceae with complete generic sampling and data from two plastid genes. *Botanical Journal of the Linnean Society* **165**: 64-83.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. **2010**. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

McLachlan GJ, Peel D, Basford KE, Adams P. **1999**. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software* **4**: 1–14.

Metzker ML. **2010**. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**: 31–46.

Milne I, Stephen G, Bayer M, Cock PJ a, Pritchard L, Cardle L, Shaw PD, Marshall D. **2013**. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**: 193–202.

Mohanty AK, Misra M, Hinrichsen G. 2000. Biofibres, biodegradable polymers and biocomposites: An overview. *Macromolecular Materials and Engineering* **276**: 1–24.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. **2006**. Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology* **6**: 17.

Motamayor JC, Risterucci AM, Heath M, Lanaud C. **2003**. Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. *Heredity* **91**: 322–330.

Motamayor JC, Lachenaud P, Mota JWS, Loor R, Kuhn DN, Brown JS, Schnell RJ. **2008**. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS ONE* **3**: e3311.

Narayan RKJ. **1982**. Discontinuous DNA variation in the evolution of plant species. The genus *Lathyrus*. *Evolution* **36**: 877–891.

Novak P, Neumann P, Macas J. **2010**. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**: 378.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. **2013**. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* **29**: 792–793.

Novick PA, Smith JD, Floumanhaft M, Ray DA, Boissinot S. **2011**. The Evolution and Diversity of DNA Transposons in the Genome of the Lizard *Anolis carolinensis*. *Genome Biology and Evolution* **3**: 1-14.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson A, Rilakovic N, Ritland C, Rosselló J a, Sena J, Svensson T, Talavera-López C, Theißen G, Tuominen H, Vanneste K, Wu Z-Q, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland

K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S. **2013**. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–84.

Otto SP, Whitton J. **2000**. Polyploid incidence and evolution. *Annual Review of Genetics* **34**: 401–437.

Palmer JD, Thompson WF. **1982**. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* **29**: 537–550.

Palmer JD. 1991. Plastid chromosomes: structure and evolution. In: Bogorad L, editor. *Molecular Biology of Plastids*. Orlando,FL: Academic Press. 5–53.

Pearce SR, Knox M, Ellis TH, Flavell AJ, Kumar A. **2000**. Pea Ty1-copia group retrotransposons: transpositional activity and use as markers to study genetic diversity in Pisum. *Molecular and General Genetics* **263**: 898–907.

Potts AJ, Hedderson TA, Grimm GW. **2014**. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Systematic Biology* **63**: 1–16.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. **2007**. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.

Quinlan AR, Hall IM. **2010**. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. **2014.** R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria, 2014. http://www.R-project.org/.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**: 1645–1656.

Raubeson LA, Jansen RK. 2005. Chloroplast genomes of plants. In: Henry RJ, editor. *Plant diversity and evolution: genotypic and phenotypic variation in higher plants.* Cambridge, MA: CAB International. 45–68.

Rheindt FE, Fujita MK, Wilton PR, Edwards S V. **2014**. Introgression and phenotypic assimilation in *Zimmerius* Flycatchers (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. *Systematic Biology* **63**: 134–52.

Ruhfel BR, Gitzendanner M a, Soltis PS, Soltis DE, Burleigh JG. **2014**. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* **14**: 23.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Rousseau-Gueutin M, Huang X, Higginson E, Ayliffe M, Day A, Timmis JN. **2013**. Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (*accD*) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiology* **161**: 1918–1929.

Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, Jansen R, Ruhlman T. **2014**. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal*. doi: 10.1111/pbi.12179.

Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, Roulin A, Guiderdoni E, Delabastide M, McCombie R. **2011**. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *The Plant Journal* **66**: 241–246.

Sanderson MJ. **2002**. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* **19**: 101–109.

Sanderson MJ. **2003**. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**: 301–302.

Sang T, Crawford DJ, Stuessy TF. **1995**. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 6813–6817.

Schaefer H, Hechenleitner P, Santos-Guerra A, de Sequeira MM, Pennington RT, Kenicer G, Carine M. **2012**. Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evolutionary Biology* **12**: 250.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. **2004**. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**: 868–876.

Schubert I, Lysak M. 2011. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics* **27**: 207-216.

Schulman AH, Flavell AJ, Paux E, Ellis T. **2012**. The application of LTR retrotransposons as molecular markers in plants. *Methods in Molecular Biology* **859**: 115–153.

Senn HA. **1938**. Experimental Data for a Revision of the Genus *Lathyrus*. *American Journal of Botany* **25**: 67 – 78.

Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution* **25**: 960–971.

Shi T, Huang H, Barker MS. **2010**. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* **106**: 497–504.

Shimodaira H, Hasegawa M. **1999**. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**: 1114–1116.

Shimodaira H, Hasegawa M. **2001**. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**: 1246–1247.

Sigurbjarnarson G. 1983. The Quaternary alpen glaciation and marine erosion in Iceland. *Jökull* **33:** 87-93.

Singh KK, Mridula D, Rehal J, Barnwal P. 2011. Flaxseed: a potential source of food, feed and fiber. *Critical Reviews in Food Science and Nutrition* **51**: 210–222.

Smith SA, Dunn CW. **2008**. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**: 715–716.

Song S, Liu L, Edwards S V., Wu S. **2012**. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences* **109**: 14942–14947.

Soltis DE, Soltis PS, Schemske DW, Hancock JF, Thompson JN, Husband BC, Judd WS. **2007**. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* **56**: 13–30.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Depamphilis CW, Wall PK, Soltis PS. **2009**. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**: 336–348.

Soltis DE, Gitzendanner MA, Stull G, Chester M, Chanderbali A, Chamala S, Jordon-Thaden I, Soltis PS, Schnable PS, Barbazuk WB. **2013**. The potential of genomics in plant systematics. *Taxon* **62**: 886–898.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. **2009**. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research* **37**: 7002–7013.

Stern DB, Goldschmidt-Clermont M, Hanson MR. **2010**. Chloroplast RNA metabolism. *Annual Review of Plant Biology* **61**: 125–155.

Stoppel R, Meurer J. **2013**. Complex RNA metabolism in the chloroplast: an update on the *psbB* operon. *Planta* **237**: 441–449.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. **2012**. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99**: 349–364.

Stull GW, Moore MJ, Mandala VS, Douglas N a., Kates H-R, Qi X, Brockington SF, Soltis PS, Soltis DE, Gitzendanner M a. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* **1**: 1200497.

Sugita M, Sugiura M. **1996**. Regulation of gene expression in chloroplasts of higher plants. *Plant Molecular Biology* **32**: 315–326.

Sukumaran J, Holder MT. **2010**. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**: 1569–1571.

Sun C, Shepard DB, Chong RA, Arriaza JL, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. **2012**. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biology and Evolution* **4**: 168–183.

Sveinsson S, Gill N, Kane NC, Cronk Q. **2013**. Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (*Theobroma cacao* L.) and related species. *BMC Genomics* **14**: 502.

Swofford DL. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer Associates*, Sunderland, Massachusetts.

Syed N, Sureshsundar S, Wilkinson M, Bhau B, Cavalcanti J, Flavell A. **2005**. Ty1-copia retrotransposon-based SSAP marker development in cashew (*Anacardium occidentale* L.). *Theoretical and Applied Genetics* **110**: 1195–1202.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. **2013**. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**: 2725–2729.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. **2011**. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians* RID D-7782-2011. *Genome Biology and Evolution* **3**: 219–229.

Thórsson AET, Palsson S, Lascoux M, Anamthawat-Jonsson K. **2010**. Introgression and phylogeography of *Betula nana* (diploid), *B. pubescens* (tetraploid) and their triploid hybrids in Iceland inferred from cpDNA haplotype variation. *Journal of Biogeography* **37**: 2098–2110.

Vamosi JC, Dickinson TA. 2006. Polyploidy and diversification: a phylogenetic investigation in Rosaceae. *International Journal of Plant Sciences* **167**: 349–358.

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**: 177-190.

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH. **1999**. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell Online* **11**: 1769–1784.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. **2013**. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22**: 787–98.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, Lambert G, Galbraith DW, Grassa CJ, Geraldes A, Cronk QC, Cullis C, Dash PK, Kumar P, Cloutier S, Sharpe AG, Wong GK-S, Wang J, Deyholos MK. 2012. The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *The Plant Journal* **72**: 461–473.

Wasmuth JD, Blaxter ML. 2004. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* **5**: 187.

Weng M-L, Blazier JC, Govindu M, Jansen RK. **2013**. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. *Molecular Biology and Evolution*: 1–15.

Wernersson R, Pedersen AG. **2003**. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research* **31**: 3537–3539.

Wessler SR. **2006**. Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences* **103**: 17600–17601.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S. **2007**. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **35**: D5–D12.

Whittall J, Liston A, Gisler S, Meinke RJ. **2000**. Detecting nucleotide additivity from direct sequences is a SNAP: An example from *Sidalcea* (Malvaceae). *Plant Biology* **2**: 211–217.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. **2011**. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology* **76**: 273–297.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O. **2007**. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**: 973–982.

Wicker T, Narechania A, Sabot F, Stein J, Vu G, Graner A, Ware D, Stein N. **2008**. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.

Wojciechowski MF, Sanderson MJ, Steele KP, Liston A. 2000. Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach. In Herendeen PS, Bruneau A, eds. *Advances in Legume Systematics, part 9*. Royal Botanic Gardens Kew, UK: 277-298.

Wojciechowski MF, Lavin M, Sanderson MJ. **2004**. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* **91**: 1846–1862.

Wolfe KH, Li WH, Sharp PM. **1987**. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **84**: 9054–9058.

Wood GAR, Lass RA. **2001**. *Cocoa*. 4th edition. Blackwell, UK: Longman Group.

Wyman SK, Jansen RK, Boore JL. **2004**. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**: 3252–3255.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumarane M, Wurdack KJ, Endress PK. Matthew ML, Stevens PF, Mathews S, Davis CC. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 17519–17524

Xu Z, Wang H. **2007**. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**: W265–W268.

Yang Z. **1997**. PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**: 555–556.

Yoder JB, Briskine R, Mudge J, Farmer A, Paape T, Steele K, Weiblen GD, Bharti AK, Zhou P, May GD, Young ND, Tiffin P. **2013**. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). *Systematic Biology* **62**: 424–438.

Yunus AG, Jackson MT. **1991**. The Gene Pools of the Grasspea (*Lathyrus sativus* L.). *Plant Breeding* **106**: 319–328.

Zerbino DR, Birney E. **2008**. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**: 821–829.

Zhang Z, Schwartz S, Wagner L, Miller W. **2000**. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.

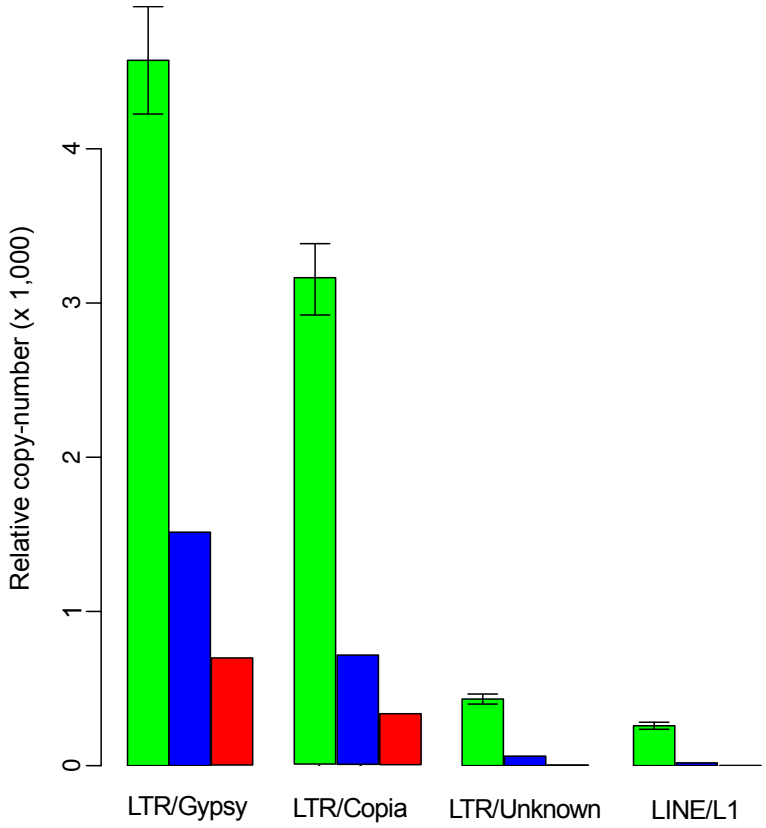
Zhou L, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL. **2004**. Transposition of hAT elements links transposable elements and V (D) J recombination. *Nature* **432**: 995–1001.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, Jong PJ de, Neale DB, Salzberg SL, Yorke JA, Langley CH. **2014**. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics* **196**: 875–890.

Zwickl DJ. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Ph.D. Thesis, University of Texas at Austin, USA.

Appendices

Appendix A - Supplementary material for chapter 2



A.1 Relative copy-number of transposable elements using reference based mapping to conserved regions of the class I LTR elements. Relative copy-numbers of the TE super-families in the three species represented with bar plots. Relative copy-number was calculated by dividing the total coverage of each super-family, within a sample, by the sample's mean UCOS coverage. The mapping was preformed with relaxed settings in the short read aligner and the reads were mapped to conserved regions of class I LTR elements.

Appendix B - Supplementary material from chapter 4

B.1 Source of plant material and voucher information

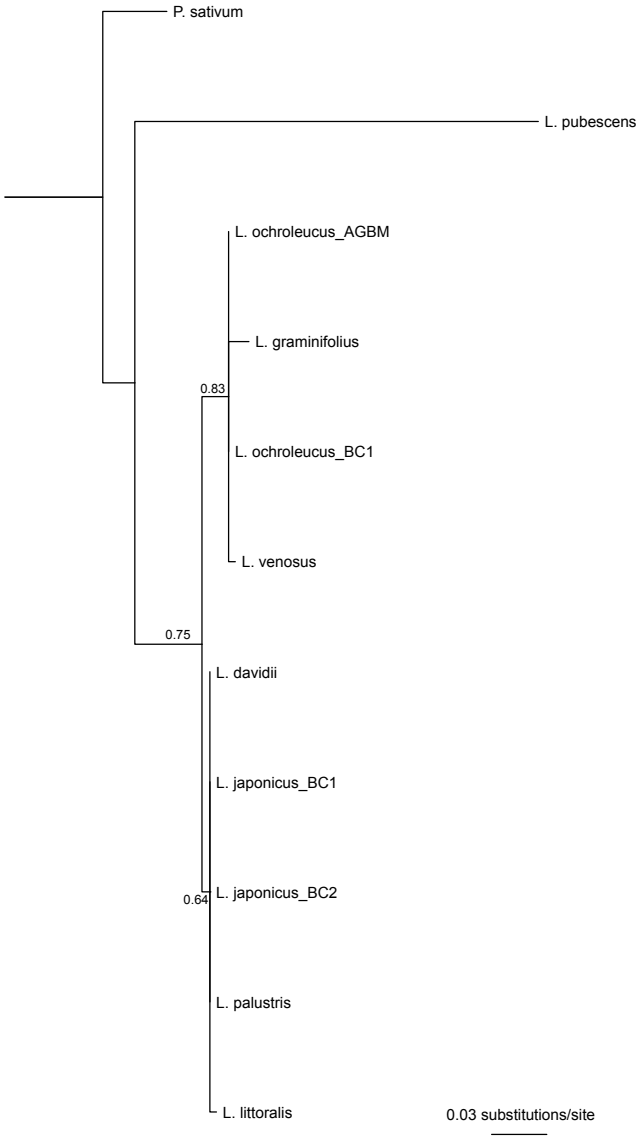
Species_population	Origin	Voucher***
Lathyrus davidii	RP*	UBC Herbarium
L. graminifolius	DLP** accession: 920239	UBC Herbarium
*L. japonicus*_BC1	Iona Beach, BC, CA	NA
L. japonicus _BC2	Iona Beach, BC, CA	NA
L. littoralis	Fort Worden State Park, WA, USA	NA
L. ochroleucus _ *AGBM*	Wonowon, BC, CA N56 43.275 / W121 48.424	UBC Herbarium
*L. ochroleucus*_BC1	Kamloops, BC, CA	UBC Herbarium
L. palustris	Iona Beach, BC, CA	UBC Herbarium
L. pubescens	RP*	UBC Herbarium
L. venosus	Moose Mountain Provincial Park, SK, CA	UBC Herbarium
L. sativus	RP*	UBC Herbarium
Pisum sativum cv. Carrera	USDA: W6-32866	UBC Herbarium

*Roger Parsons Sweet Peas

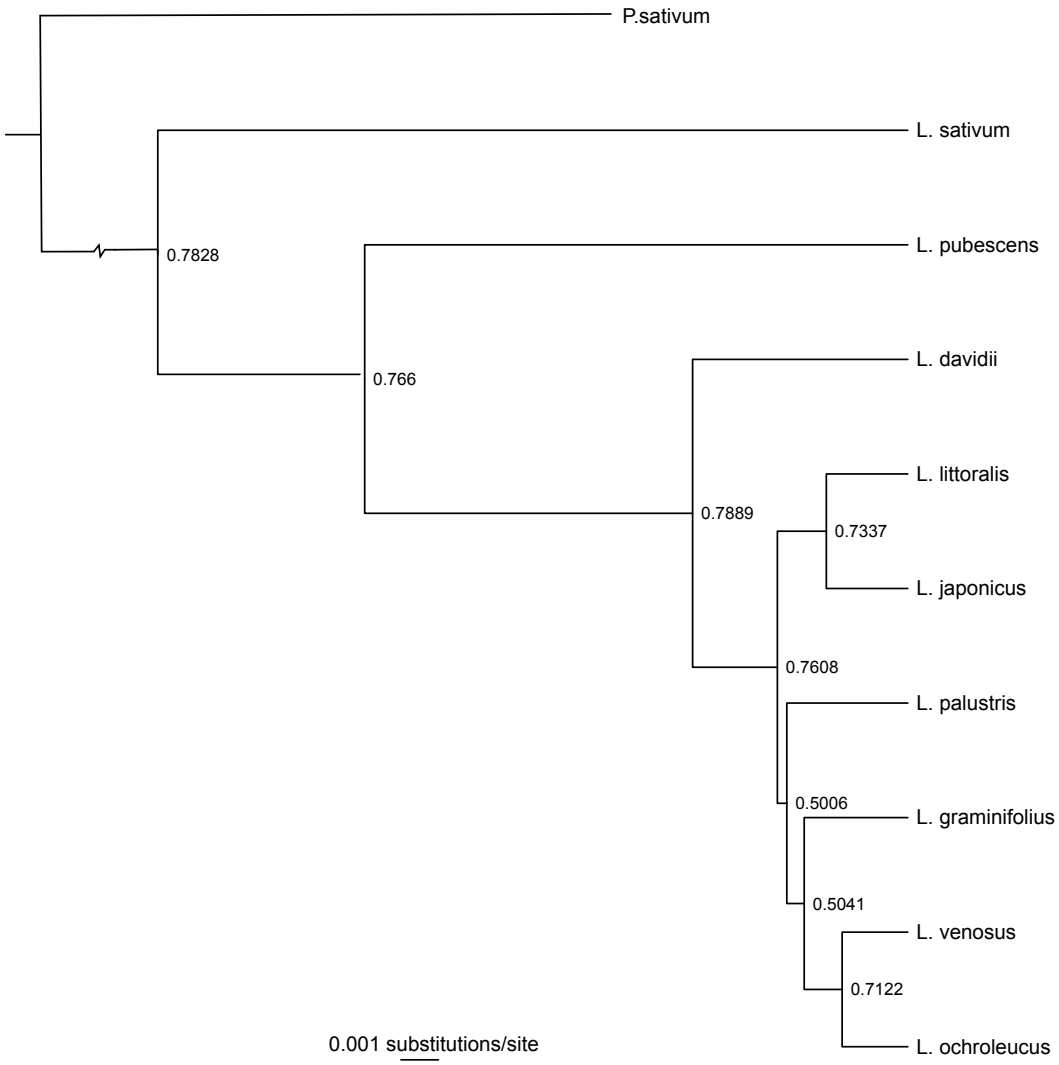
**Desert Legume Project (http://cals.arizona.edu/desertlegumeprogram/)

***Voucher samples will be deposited on acceptance for publication

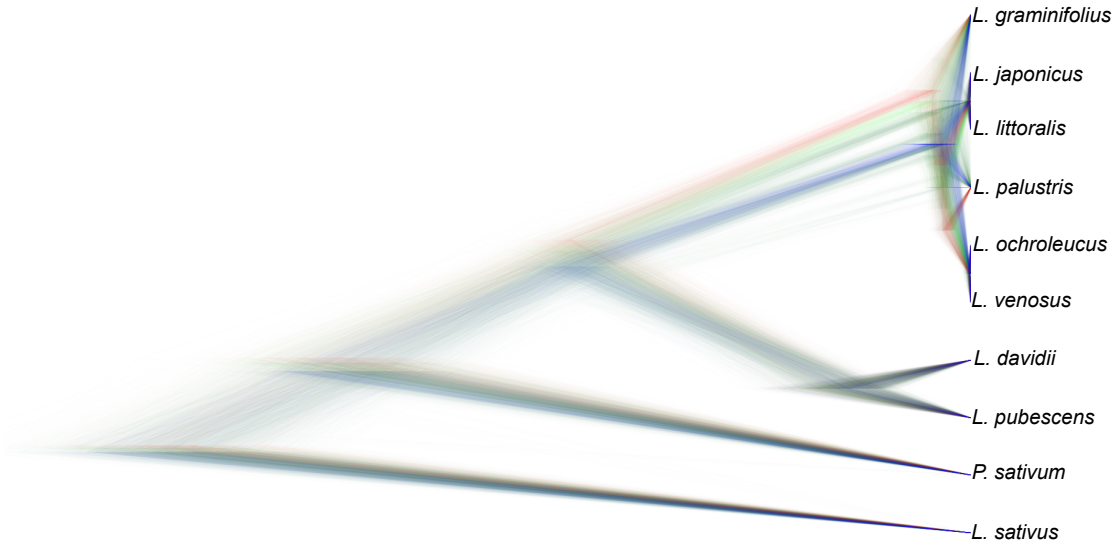
B.2 Phylogenetic relationships among the *Lathyrus* species based on the small ribosomal subunit (5S). The tree was inferred using maximum likelihood and node support values calculated from 100 bootstrap replicates.



B.3 Phylogenetic relationships among the *Lathyrus* species based on a *BEAST species tree reconstruction of the large ribosomal subunit (45S) and protein coding regions from the plastome. Posterior probability scores are indicated on the nodes of the tree.



B.4 Densitree generated using SNAPP A DensiTree visualization of the SNAPP analysis.



Appendix C - T2Phy

C.1 T2Phy: from transcriptomes to phylogeny

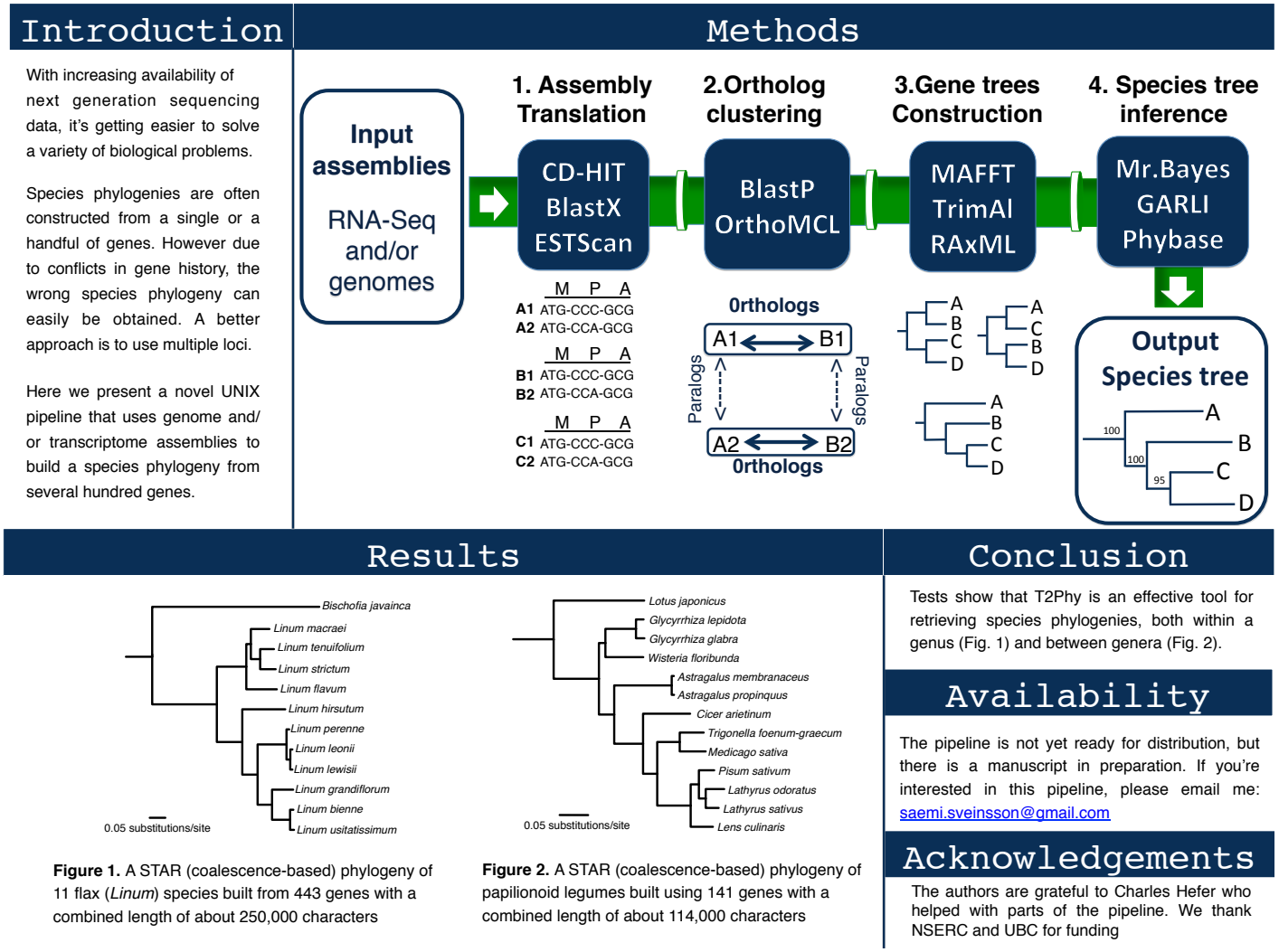


Figure 1. A STAR (coalescence-based) phylogeny of 11 flax (*Linum*) species built from 443 genes with a combined length of about 250,000 characters

Figure 2. A STAR (coalescence-based) phylogeny of papilionoid legumes built using 141 genes with a combined length of about 114,000 characters