Mating System Shifts and Transposable Element Evolution in Plants

by

J Arvid Ågren

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Ecology and Evolutionary Biology
University of Toronto

© Copyright by J Arvid Ågren, 2015
Mating System Shifts and Transposable Element Evolution in Plants

J Arvid Ågren

Doctor of Philosophy

Department of Ecology and Evolutionary Biology

University of Toronto

2015

Abstract

Transposable elements (TEs) are mobile genetic elements that can self-replicate and insert elsewhere in the genome. This movement often comes with a fitness cost and yet TEs are tremendously common. They contribute more than 80% to the largest plant genomes, but are much less abundant in the smallest plant genomes. Combined, these observations raise the two central questions of this thesis. First, why are TEs so common, and the genomes so large, in some species but not others? Second, what prevents TEs from completely taking over and causing genomes to fail to produce functional individual organisms? In this thesis, I have used population and comparative genomics approaches on whole genome data from closely related plant species and synthesising literature reviews to examine some of the evolutionary causes and consequences of TE proliferation in plants. In particular, I have investigated the role of mating system shifts, from outcrossing to self-fertilization and from sexual to asexual reproduction, in driving variation in TE abundance and genome size. I used the reference genome of Capsella rubella as an outgroup to show that the higher TE abundance and larger genome size of the
outcrossing *Arabidopsis lyrata* compared to selfing *A. thaliana* is due to TE driven expansion in the outcrosser rather than genome loss in the s切尔. I investigated three *Capsella* species of contrasting mating system and found that the evolution of self-fertilization may have different effects on TE evolution on short and on long timescales. I applied a phylogenetic comparative approach and whole genome sequencing to demonstrate that sex and TEs cannot explain the variation in genome size in evening primroses (*Oenothera*). Using transcriptome data I showed that the chromosomal distribution (sex chromosomes vs. autosomes) of nuclear gene with organellar origin in *Rumex hastatulus* is not consistent with either co-adaptation or sexual conflict hypotheses. Finally, I showed that studying the proliferation of TEs offers insights to the cardinal problem of social evolution of what prevents selfish behaviour at lower levels from destroying functionality of the group. Overall, my thesis is a contribution to understanding the evolutionary causes and consequences of TE proliferation.
Dedication

Till Frida
Acknowledgments

Throughout my PhD I have had the privilege to interact with a large number of people and their ideas, advice, and friendship have shaped both this thesis and my growth as a scientist.

First among them is my supervisor Stephen Wright. I thank him for his unwavering encouragement and advice, continued financial support, and for assembling such a stimulating group that is the Wright Lab. I am indebted to him for introducing me to transposon biology and letting me follow my interests as they developed, for always making time to chat and for constantly challenging me, and for setting such an example as a scientist.

I am also thankful to my Supervisory Committee members, Spencer Barrett and Asher Cutter, for discussions, their careful readings of early drafts, and insightful advice during the development of ideas and projects. Marc Johnson and Megan Frederickson generously took part in my Appraisal Examination. I am thankful to Marc for coming back to be on my Final Examination Committee together with Marla Sokolowski. Finally, I am also grateful to Maud Tenaillon (UMR de Génétique Végétale du Moulon) for serving as my External Examiner.

I have continuously benefited from interactions and discussions with members of the labs of Spencer Barrett and John Stinchombe, participants in the Evolutionary Genetics Journal Club, and with Locke Rowe. I am also grateful to Deborah Charlesworth (University of Edinburgh) for first introducing me to Stephen while I was still only a second year undergraduate student and for her continuing support over the years. In addition to having some excellent collaborators on many of my projects, I also received invaluable technical assistance from Wei Wang and
Niroshini Epitawalage. Jill Wheeler has been instrumental in developing my teaching skills and philosophy.

I joined the Wright Lab at the same time as Robert Williamson and Emily Josephs and I could not have asked for better companions during this journey. Their friendship is greatly treasured. Other Wright Lab members, past and present, have also contributed a great deal, both on personal and professional levels. Special thanks to Jesse Hollister, Young Wha Lee, Josh Hough, and Matthew Hartfield. The broader grad student and post doc community of EEB, no one mentioned no one forgotten, has also played a key role in making these five years so fun and stimulating.

Massey College has provided an intellectual environment and friendships beyond my wildest imagination, especially during my three years as a resident Junior Fellow. This experience has greatly shaped the kind of scientist I strive to be. I am grateful to Master Emeritus John Fraser and Elizabeth MacCullum, and the broader Massey Community, for making the College such a special place and a home away from home.

My parents, Jon Ågren and Maria Wikman, and my sister, Frida Ågren, have always provided never-ending love and encouragement in all my endeavours, and this thesis was no exception. For that they have my deepest gratitude. Finally, I am forever grateful to Utako Tanebe, for her support, partnership, and for always brightening up my day.

Throughout my PhD, I have been financially supported by a University of Toronto Fellowship and by a Massey College Junior Fellowship. The research in this thesis was funded through grants from the Natural Sciences and Engineering Research Council (NSERC) to Stephen Wright.
Table of Contents

Dedication .................................................................................................................................................. iv

Acknowledgments ........................................................................................................................................ v

Table of Contents ....................................................................................................................................... vii

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

List of Figures ........................................................................................................................................... xiii

List of Supplementary Tables .................................................................................................................. xv

List of Supplementary Figures ................................................................................................................ xvii

List of Boxes ............................................................................................................................................... xix

List of Appendices .................................................................................................................................... xx

CHAPTER 1 GENERAL INTRODUCTION ......................................................................................... 1

1.1 The biology of selfish genetic elements ......................................................................................... 1

1.2 Transposable elements ...................................................................................................................... 3

1.3 The role of sex and selfing in transposable element evolution ...................................................... 6

1.4 Conflict and cooperation in cyto-nuclear interactions .................................................................. 9

1.5 Transposable elements and evolutionary transitions in individuality ........................................... 12

1.6 Study systems .................................................................................................................................. 13

1.6.1 Arabidopsis .................................................................................................................................. 14

1.6.2 Capsella ...................................................................................................................................... 14

1.6.3 Oenothera .................................................................................................................................. 15

1.6.4 Rumex ........................................................................................................................................ 15

1.7 Research objectives .......................................................................................................................... 17

1.7.1 Chapter 2 Sizing up Arabidopsis genome evolution .................................................................. 17

1.7.2 Chapter 3 The Capsella rubella genome and the genomic consequences of rapid mating system evolution ........................................................................................................ 18

1.7.3 Chapter 4 Mating system shifts and transposable element evolution in the plant genus

Capsella ....................................................................................................................................................... 19
1.7.4 Chapter 5 Co-evolution between transposable elements and their host: a major factor in genome size evolution .......................................................... 20
1.7.5 Chapter 6 No evidence that sex and transposable elements drive genome size variation in evening primroses ............................................................... 21
1.7.6 Chapter 7 Chromosomal distribution of cytonuclear genes in a dioecious plant with sex chromosomes ............................................................... 21
1.7.7 Chapter 8 Evolutionary transitions in individuality: insights from transposable elements ......................................................................................... 22

CHAPTER 2 SIZING UP ARABIDOPSIS GENOME EVOLUTION ........................................ 23

CHAPTER 3 THE CAPSELLA RUBELLA GENOME AND THE GENOMIC CONSEQUENCES OF RAPID MATING SYSTEM EVOLUTION ................................................................. 28

3.1 Main ............................................................................................................. 29
3.2 Methods ..................................................................................................... 43
  3.2.1 Genome assembly and annotation ......................................................... 43
  3.2.2 FISH and comparative chromosome painting ....................................... 49
  3.2.3 URLs .................................................................................................... 50
  3.2.4 Accession codes ................................................................................... 50
3.3 Acknowledgements .................................................................................... 51
3.4 Contributions ............................................................................................ 52

CHAPTER 4 MATING SYSTEM SHIFTS AND TRANPOSEABLE ELEMENT EVOLUTION IN THE PLANT GENUS CAPSELLA ................................................................. 53

4.1 Background ............................................................................................... 54
4.2 Results ....................................................................................................... 58
  4.2.1 Distribution of TE insertions in Capsella ............................................... 58
  4.2.2 Abundance of TEs .............................................................................. 61
  4.2.3 TE insertions near genes ..................................................................... 63
  4.2.4 Frequency distributions of TEs ............................................................. 64
4.3 Discussion ................................................................................................. 66
4.4 Conclusions ............................................................................................... 68
4.5 Methods ................................................................................................... 69
  4.5.1 Sampling and sequencing ..................................................................... 69
  4.5.2 Genome assemblies ............................................................................. 70
4.5.3 Identification of unique TE insertions ................................................................. 70
4.5.4 Frequency distributions ....................................................................................... 72
4.5.5 Availability of supporting data ............................................................................... 73
4.6 Competing interests ................................................................................................. 73
4.7 Authors’ contributions ............................................................................................. 74
4.8 Acknowledgements .................................................................................................. 74
4.9 Supplementary Information ...................................................................................... 75

CHAPTER 5 CO-EVOLUTION BETWEEN TRANSPOSABLE ELEMENTS AND THEIR
HOSTS: A MAJOR FACTOR IN GENOME SIZE EVOLUTION? ............................... 80
5.1 Introduction ............................................................................................................... 81
5.2 Re-activation of TE activity ....................................................................................... 84
5.3 The genomic immune system ................................................................................... 86
5.4 The evolution of genomic virulence ......................................................................... 90
5.5 Determining the outcome of TE–host interactions ................................................... 93
5.6 Conclusions ............................................................................................................... 96

CHAPTER 6 NO EVIDENCE THAT SEX AND TRANSPOSABLE ELEMENTS DRIVE GENOME
SIZE VARIATION IN EVENING PRIMROSES ....................................................... 98
6.1 Introduction ............................................................................................................... 99
6.2 Material and Methods .............................................................................................. 104
6.2.1 Study system ......................................................................................................... 104
6.2.2 Genome size estimates ......................................................................................... 106
6.2.3 Phylogenetic analysis ........................................................................................... 106
6.2.4 DNA isolation and Illumina sequencing ................................................................ 108
6.2.5 Repetitive content analysis .................................................................................. 109
6.3 Results ..................................................................................................................... 110
6.3.1 Variation in genome size ....................................................................................... 110
6.3.2 Phylogenetic signal ............................................................................................. 111
6.3.3 Phylogenetic generalized least squares analysis .................................................. 111
6.3.4 Repetitive content ............................................................................................... 113
6.4 Discussion ............................................................................................................... 115
6.5 Acknowledgements .................................................................................................. 118
6.6 Supplementary Information ...................................................................................... 120
CHAPTER 7 CHROMOSOMAL DISTRIBUTION OF CYTO-NUCLEAR GENES IN A DIOECIOUS PLANT WITH SEX CHROMOSOMES ............................................................... 133

7.1 Introduction ................................................................................................................. 134

7.2 Methods .......................................................................................................................... 137

7.2.1 Gene identification and functional annotation .................................................... 137

7.2.2 Statistical analyses ...................................................................................................... 138

7.3 Results and Discussion ................................................................................................. 140

7.4 Acknowledgements ......................................................................................................... 144

7.5 Supplementary Material .............................................................................................. 144

7.5.1 Power Analysis .............................................................................................................. 144

CHAPTER 8 EVOLUTIONARY TRANSITIONS IN INDIVIDUALITY: INSIGHTS FROM TRANSPOSABLE ELEMENTS ................................................................. 146

8.1 Conflict and cooperation across the hierarchy of life .................................................... 147

8.1.1 Empirical studies of evolutionary transitions ....................................................... 152

8.2 The biology of transposable elements ......................................................................... 153

8.3 Mechanisms of social group maintenance .................................................................. 155

8.4 Self-restraint in the genome ......................................................................................... 156

8.4.1 Alternative strategies for transposons ..................................................................... 158

8.5 Policing the genome ....................................................................................................... 158

8.5.1 Empirical examples of transposon policing .............................................................. 159

8.5.2 Efficiency of transposon policing ........................................................................... 160

8.6 Social group maintenance in the genome ................................................................... 161

8.7 The future of social evolution is in the genome ........................................................... 162

8.8 Concluding remarks ...................................................................................................... 163

8.9 Acknowledgements ....................................................................................................... 164

CHAPTER 9 CONCLUDING DISCUSSION .......................................................................... 165

9.1 Summary of thesis chapters .......................................................................................... 165

9.1.1 Chapter 2 Sizing up Arabidopsis genome evolution .................................................. 165

9.1.2 Chapter 3 The Capsella rubella genome and the genomic consequences of rapid mating system evolution ................................................................. 166

9.1.3 Chapter 4 Mating system shifts and transposable element evolution in the plant genus Capsella .......................................................................................... 166
9.1.4 Chapter 5 Co-evolution between transposable elements and their host: a major factor in genome size evolution.......................................................................................................................... 167
9.1.5 Chapter 6 No evidence that sex and transposable elements drive genome size variation in evening primroses.......................................................................................................................... 167
9.1.6 Chapter 7 Chromosomal distribution of cytonuclear genes in a dioecious plant with sex chromosomes.......................................................................................................................... 168
9.1.7 Chapter 8 Evolutionary transitions in individuality: insights from transposable elements.......................................................................................................................... 168

9.2 Avenues for future work ........................................................................................................................................................................................................... 169
9.2.1 The rate and direction of TE evolution following a mating system shift................................................................................................. 169
9.2.2 The causes of genome expansion........................................................................................................................................................................ 170
9.2.3 Transposable elements and the general theory of conflict and cooperation......... 172

9.3 Conclusion.............................................................................................................................................................................................................. 173

References ............................................................................................................................................................................................................... 174

Appendix ........................................................................................................................................................................................................ 212
List of Tables

Table 3.1 *Capsella rubella* genome annotation results................................................................. 30

Table 6.1 Clustering statistics for RepeatExplorer ............................................................................. 110

Table 7.1 Number of genes in *Rumex* data set.................................................................................. 138

Table 8.1 The major transitions in evolution and individuality............................................................. 150
List of Figures

Figure 1.1 The biology of transposable elements ................................................................. 4

Figure 1.2 Inheritance in plant cells ..................................................................................... 11

Figure 1.3 Evolutionary transitions in individuality and the hierarchy of life .................... 13

Figure 1.4 Study systems ..................................................................................................... 16

Figure 3.1 Genomic structures, chromosome painting and comparative genomic mapping in C. rubella, A. lyrata and A. thaliana ........................................................................... 32

Figure 3.2 Evolution of gene expression in selfing and outcrossing Capsella and comparisons to Arabidopsis ..................................................................................................................... 35

Figure 3.3 Polymorphism comparisons in C. rubella and C. grandiflora ............................. 37

Figure 3.4 Evolution of genome structure and transposable element abundance in Capsella and Arabidopsis .......................................................................................................................... 40

Figure 3.5 Estimates of transposable element copy number and expression in C. rubella and C. grandiflora .......................................................................................................................... 42

Figure 4.1 Phylogenetic relationships within the Capsella genus ........................................ 57

Figure 4.2 Venn diagram with the number of unique and shared TE insertion sites in three Capsella species ............................................................................................................................... 60

Figure 4.3 Average TE copy number in the three Capsella species ....................................... 60

Figure 4.4 Average TE copy number in 100 bp bins near their closest gene in the three Capsella species .............................................................................................................................. 64

Figure 4.5 Histogram of population frequencies of TEs in the three Capsella species ........ 65
Figure 5.1 Factors influencing genome size via the evolution of transposition rates .................. 83

Figure 6.1 Molecular phylogeny of *Oenothera* used in the PGLS analysis ............................. 112

Figure 6.2 Repetitive content of three *Oenothera* species .................................................. 114

Figure 7.1 Representation of the chromosomal location of cyto-nuclear genes in *Rumex hastatulus* ........................................................................................................................................ 140

Figure 8.1 The evolutionary transition to the genome ................................................................. 152

Figure 8.2 Classes of transposables elements ............................................................................. 155

Figure 9.1 Effect of selfing on TE evolution ................................................................................ 170
List of Supplementary Tables

Supplementary Table 3.1 Genomic libraries included in the *Capsella rubella* genome assembly and their respective assembled sequence coverage levels in the final release........................................... 248

Supplementary Table 3.2 Summary statistics of the output of the whole genome shotgun assembly prior to screening, removal of organelles and contaminating scaffolds and chromosome-scale pseudomolecule construction................................................................. 249

Supplementary Table 3.3 Final summary assembly statistics for chromosome scale assembly. 250

Supplementary Table 3.4 Placement of the individual Fosmid clones and their contribution to the overall error rate............................................................................................................. 250

Supplementary Table 3.5 Placement of the individual BAC clones and their contribution to the overall error rate............................................................................................................. 251

Supplementary Table 3.6 Summary statistics of *de novo* assembly of *C. grandiflora*........ 252

Supplementary Table 3.7 Assembly statistics for *de novo* assembly of *N. paniculata*. ... 253

Supplementary Table 3.8 Mapping of genic regions in *C. rubella* (CR) onto *C. grandiflora* (CG) and *N. paniculata* (NP) assemblies using liftOver with >80% identity from best orthologous chains. ................................................................................................................................. 255

Supplementary Table 3.9 Percentage of the *C. rubella* (CR) genome for which alignments are reliably suggested in the *C. grandiflora* (CG) and *N. paniculata* (NP) assemblies........... 255

Supplementary Table 3.10 Description of the ten non-chromosomal scaffolds with largest inferred genomic content. ............................................................................................................. 256

Supplementary Table 3.11 Ontology categories identified as significantly differentiated in expression between *C. rubella* and *C. grandiflora* using the DAVID system. ......................... 257
Supplementary Table 3.12 Candidate differentially expressed genes falling in floral QTL regions that have been mapped in interspecific crosses between *C. grandiflora* and *C. rubella.*

Supplementary Table 3.13 Genes in the brassinosteroid pathway showing changes in gene expression

Supplementary Table 3.14 *Capsella* samples used in RNAseq population genetics analysis.

Supplementary Table 3.15 Proportions of new nonsynonymous mutations in bins with different negative fitness effects.

Supplementary Table 3.16 Parameter estimates for three demographic models of the split between *C. grandiflora* and *C. rubella.*

Supplementary Table 4.1 Origins of sequenced *Capsella* samples

Supplementary Table 4.2 Assembly statistics for *C. orientalis de novo* assembly

Supplementary Table 6.1 Seed curation information for *Oenothera* plants used in flow cytometry

Supplementary Table 6.2 Genome size estimates in *Oenothera*
List of Supplementary Figures

Supplementary Figure 3.1 Marker placements for the genetic map on Capsella ....................... 229

Supplementary Figure 3.2 Pairwise linkage plot from an initial de novo assembly of the C. rubella genome. .................................................................................................................. 234

Supplementary Figure 3.3 Dot plot of the final C. rubella assembly (x-axis) against the A. lyrata assembly (y-axis). ........................................................................................................ 236

Supplementary Figure 3.4 Dot plot of Fosmid clones on regions of C. rubella scaffolds ........ 237

Supplementary Figure 3.5 CoGe dotplot comparison.................................................................. 243

Supplementary Figure 3.6 Observed size of scaffolds on the x-axis vs. inferred size based on Illumina coverage on the y-axis. .......................................................................................... 244

Supplementary Figure 3.7 Classification of Illumina genomic reads into repetitive and nonrepetitive categories. ........................................................................................................ 244

Supplementary Figure 3.8 Chromosome painting analysis ...................................................... 245

Supplementary Figure 3.9 Dotplot comparisons of Arabidopsis lyrata, Capsella rubella, and Schrenkiella parvula................................................................. 246

Supplementary Figure 3.10 Principal components analysis and hierarchical clustering analysis of differentially expressed genes in C. rubella and C. grandiflora........................................... 246

Supplementary Figure 3.11 Frequency spectra from C. rubella and C. grandiflora ............... 247

Supplementary Figure 4.1 Total TE copy number in sampled individuals in the three Capsella species. .................................................................................................................. 77

Supplementary Figure 4.2 Average TE copy number in the three Capsella species .............. 78
Supplementary Figure 4.3 Inferred TE frequencies in highly selfed accessions and synthetic hybrid .......................................................... 79

Supplementary Figure 7.1 Power analysis of chromosomal distribution .............................................. 145
List of Boxes

Box 8.1 Major transitions in evolution and individuality ............................................................... 148
List of Appendices

Supplementary Note 3.1.......................................................................................................................... 212
CHAPTER 1 GENERAL INTRODUCTION

"How did it come about that most genes, most of the time, play fair, so that a gene's fitness depend only on the success of the individual that carries it?"

John Maynard Smith, 1985

1.1 The biology of selfish genetic elements

Historically, the predominant view of genomes is one of a highly coordinated network, with all parts playing fair, working together to produce individual organisms. The existence of genes playing unfair, promoting their own transmission at the expense of other genes in the genome (now usually referred to as selfish genetic elements) has been known since the 1920s, but such cases were typically considered to be genetic oddities (Burt and Trivers 2006; Werren 2011). In 1945, the Swedish botanist and cytogeneticist Gunnar Östergren provided the first clear articulation that selection on certain genes may be in conflict with the rest of the organism (Östergren 1945). Discussing supernumerary (i.e. non-vital) B chromosomes in various plant species, he pointed out that "in many cases these chromosomes have no useful function at all to the species carrying them, but that they often lead an exclusively parasitic existence’ and that "[B chromosomes] need not be useful for the plants. They need only be useful to themselves". However, while Östergren’s work coincided with several other similar empirical observations, many of those in plants (reviewed in Ågren and Wright 2015), it would take several decades until selfish genetic elements in general, and their evolutionary implications in particular, became widely appreciated.
Three parallel developments have been identified as central to bring about this conceptual shift (Werren 2011). First, inspired initially by Hamilton’s inclusive fitness models (1963; 1964; 1967) and more forcefully by the writings of Williams (1966) and Dawkins (1976), the gene’s eye view of evolution was rapidly gaining popularity. In particular, the emphasis of genes as immortal replicators manipulating organisms (‘vehicles’ by Dawkins’ nomenclature) to further their own transmission made it easier to accept that all genes may not necessarily share the same fitness interest. Second, this theoretical progress coincided with early sequencing efforts that reported that large chunks of eukaryotic genomes were made up of genetic material, such as repetitive DNA, with seemingly no connection to organismal fitness (e.g Britten and Kohne 1968; Britten 1969). Moreover, while it was clear that genome size varied dramatically across species (we now know it varies more than 100,000-fold), there was no correlation between the amount of DNA of a species (C-value) and its perceived complexity. For example, the genome of single-celled amoeba is about 100 times the size of that of humans. This lack of correlation was later termed the “C-value paradox” (Thomas 1971). These observations were central to two papers published back-to-back in Nature in 1980. Doolittle and Sapienza (1980) and Orgel and Crick (1980) independently argued that large parts of eukaryotic genomes can best be described as selfish, with negative or neutral effects on organismal fitness. Other key conceptual writings of the time were provided by Cosmides and Tooby (1981) and Dawkins (1982, Chapters 8–9). Third, empirical work in molecular genetics continued to provide new examples of selfish genetic elements. When Werren and colleagues published the first comprehensive review of all kinds selfish genetic elements discovered at the time, their discussion covered
examples ranging from meiotic drive and supernumerary B chromosomes to killer plasmids, selfish mitochondria, and transposable elements (Werren et al. 1988). We now know that selfish genetic elements are prominent features of the genomes of virtually all organisms (Hurst and Werren 2001; Burt and Trivers 2006; Werren 2011). Indeed, in light of the growing evidence of the central role played by selfish genetic elements in genome evolution, Rice (2013) recently argued that “nothing in genetics makes sense except in the light of genomic conflicts”. It is in that spirit that this thesis is written.

1.2 Transposable elements

Transposable elements (TEs) are among the most well studied selfish genetic elements. First described in maize by Barbara McClintock (1950), a discovery that earned her the 1983 Nobel Prize in Physiology or Medicine, TEs are mobile genetic elements that are typically categorized into two classes depending on their mode of self-replication (Finnegan 1989; Wicker et al. 2007; Piégu et al. 2015; Figure 1.1). Class 1, or retrotransposons, replicate via an RNA intermediate, a process sometimes called copy-and-paste transposition. Class 2 elements, however, employ a cut-and-paste mechanism that does not involve reverse transcription. Beyond class, TEs can be further categorized based on sequence similarity and structural similarity. Comparative work suggests that species vary greatly in both abundance and diversity of TEs (Kidwell 2002; 2003; Elliot and Gregory 2015).
Figure 1.1 The biology of transposable elements

TEs taxonomy is based on mode replication and sequence similarity and structural similarity. Class 1 elements, sometimes known as retrotransposons, apply a ‘copy-and-paste’ mechanism and produce an RNA intermediate that is then reverse-transcribed into DNA and integrated into a new location in the genome. However, Class 2 DNA elements replicate via a ‘cut-and-paste’ approach and the transposon is physically excised and integrated elsewhere in the genome.

While self-replication provides a way for TEs to increase in copy number, their movement usually comes with a fitness cost for the individual organism. Three sources of the deleterious effects of TEs have been suggested. First, TEs may land in protein-coding or regulatory regions of genes, and so disrupt their function (Finnegan 1992). Second, TEs may induce large-scale chromosomal rearrangements, insertions, and deletions through ectopic recombination between non-homologous sites (Montgomery et al. 1987). Third, the process of transposition itself may come at a metabolic cost (Brookfield 1991). These models are typically difficult to distinguish empirically (Dolgin and Charlesworth
2008). Observations in *Drosophila melanogaster*, an outcrossing species with a large effective population size, seem to be best described by the ectopic recombination model (Barron et al. 2014) whereas in predominately selfing species, where most TEs will be homozygous, such as *Arabidopsis thaliana* (Wright et al. 2003) and *Caenorhabditis elegans* (Dolgin et al. 2008), the deleterious effects of inserting near genes may be more important. Regardless of the source, the phenotypic effects of TEs are extensive, ranging from fruit shape to hybrid incompatibilities (Lisch 2013; Ågren 2013a), and there is abundant evidence from both natural and experimental populations of the fitness costs of TEs (Charlesworth et al. 1994; Pasyukova et al. 2004; Lynch 2007, chapter 7; Lisch 2013).

Given the evidence that TEs are predominately deleterious, it is striking that TEs are so common. By analyzing ten million protein-encoding genes from prokaryotic, eukaryotic, archaeal, and viral genomes, Aziz et al. (2010) argued that TEs are the most common genetic material in nature. Indeed, to paraphrase the great late JBS Haldane, the creator could be said to have an inordinate fondness for TEs (Ågren 2013b). TEs are particularly common in plant genomes (Fedoroff 2012) and there appears to be a positive correlation between the fraction of the genome comprised of TEs and genome size in angiosperms (Tenaillon et al. 2010; Michael 2014). Species with relatively large genomes, such as maize (Schnable et al. 2009) and barley (Wicker et al. 2005), have a TE content of over 80% whereas species with relatively small genomes, such as *Arabidopsis thaliana*, tend to have TE contents of around 20% (The *Arabidopsis* Genome Initiative 2000; Hu et al. 2011). This strong association means that there is much to be gained from studying TE and genome size evolution in conjunction (Grover and Wendel 2010; El
Baidouri and Panaud 2013; Bennetzen and Wang 2014; Lee and Kim 2014), and several chapters in this thesis consider the causes and consequences of TE-genome size co-evolution.

Taken together, the observations outlined above raise two questions that are at the heart of this thesis. First, why are TEs sometimes so common and at other times so rare? And, second, what prevents TEs from completely taking over genomes and causing the breakdown of organismal function? Surprisingly, despite extensive theoretical and empirical attention, there is little consensus about the answers to these questions (Nuzhdin 1999; Dolgin and Charlesworth 2008; Biémont 2010; Tenaillon et al. 2010; Grandbastien and Casacuberta 2012; Barron et al. 2014). The advancement of whole genome sequencing has opened up a gold mine of data, allowing theoretical models to be tested across genomes and across species at a scale not previously possible. The central goal of this thesis research is therefore to apply population and comparative genomics tools to closely related plant species to examine the accumulation of TEs and their role in genome size variation, and to use these insights to ask what prevents the accumulation of TEs from being lethal to the individual organism.

1.3 The role of sex and selfing in transposable element evolution

Stebbins (1950) famously called the shift from outcrossing to predominantly self-fertilizing the most common evolutionary transition in plants. Numerous genomic consequences of this shift have since been described (reviewed in Charlesworth and
Wright 2001; Wright et al. 2008; Barrett et al. 2014). Because mating system will affect both the mode of transmission of and selection against TEs, it provides an interesting window into TE evolution. In general, highly selfing/asexual genomes are expected to experience less conflict between selfish genetic elements and the rest of the host genome than outcrossing sexual genomes (Hickey 1982; Burt and Trivers 1998; Burt and Trivers 2006). This prediction stems from several factors and I will illustrate them using TEs, but the logic typically holds for selfish genetic elements more broadly. To start, selfing prevents the spread of TEs into new genetic backgrounds. In a highly selfing or asexual lineage, any TE is essentially stuck in that lineage, which should increase variation in fitness between individuals. The increased variation should result in stronger purifying selection in selfers, as a lineage without selfish genetic elements should out-compete a lineage with TEs (Cavalier-Smith 1980; Hickey 1982). Moreover, modeling has suggested that the greater linkage disequilibrium in selfing compared to outcrossing genomes may in some, albeit rather limited, cases cause selection for reduced transposition rates (Charlesworth and Langley 1986). Overall, this reasoning leads to the prediction fewer TEs should be observed in highly selfing or asexual taxa (Wright and Schoen 1999; Morgan 2001). This prediction runs opposite to the popular mutation-accumulation reasoning of Lynch and Conery (2003; Lynch 2007; 2011), which predicts more TEs in selfing compared to outcrossing genomes. This alternative hypothesis builds on the expectation that the evolution of selfing is associated with a reduction in the effective population size ($N_e$; Nordborg 2000). A reduction in $N_e$ should reduce the efficacy of selection and therefore lead to higher accumulation of TEs in selfers relative to outcrosser, all else being equal. This scenario is especially likely if selection is mainly
acting on fitness reductions due to chromosomal rearrangements caused by ectopic recombination (Charlesworth and Charlesworth 1995).

Similar to the shift from outcrossing to selfing, theory predicts that the evolution of asexuality from sexual reproduction will also affect TE proliferation (recently reviewed in Glémin and Galtier 2012 and Bast 2015). Just like the shift from outcrossing to selfing, the net outcome of the shift will be determined by the balance between effects on transmission and efficacy of selection against TEs. On the one hand, sexual reproduction allows TEs to spread to new lineages in a population, a route absent in asexual populations. On the other hand, asexual lineages may be driven to extinction through an indefinite accumulation of TEs by a ratchet like process (Hickey 1982; Nuzhdin and Petrov 2003; Crespi and Schwander 2012). Two key differences between the genomic consequences of selfing and asexuality may be important in TE evolution. First, occasional outcrossing will occur even in highly selfing populations (Vogler and Kalisz 2001; Igić and Kohn 2006), whereas between-individual transmission is typically completely absent in asexuals. Second, the evolution of selfing is associated with elevated levels of homozygosity across the genome. In the absence of gene conversion, asexuals will remain heterozygous permanently. Thus, if the degree of heterozygosity is important in determining the strength of selection against TEs, as under the ectopic recombination model described above, selection against TEs may differ between selfers and asexuals.

A key aim of this thesis is to investigate how these contrasting effects of mating on TE evolution play out in natural plant populations, and to explore their importance in governing variation in genome size.
1.4 Conflict and cooperation in cyto-nuclear interactions

As outlined above, a central theoretical motivation for the reasoning outlined above is that the evolutionary fate of mobile genetic elements, such as TEs, is likely to be affected by the mode of inheritance. Although TEs are probably the best example of mobile genetic elements, another striking instance of gene movement is that between the organellar (mitochondrial and chloroplast) genomes and the nuclear genomes in eukaryotes. Indeed, since the origin of the eukaryotic cell, most mitochondrial and chloroplast genes have migrated to the nuclear genome (Adams and Palmer 2003; Martin 2003; Rand et al. 2004; Burt and Trivers 2006; Bar-Yaacov et al. 2012). This movement has been essential to the evolution of complexity (Lane 2005; 2015) but may also result in evolutionary conflicts.

Both the mitochondrial and chloroplast genomes are of bacterial origin and entered a mutualistic partnership with plants cells some time 1.5–2 billion years ago (Dyall et al. 2004). Like in any mutualistic relationship, fitness interests between the organellar genomes and the nuclear genome sometimes overlap and sometimes they diverge (Cosmides and Tooby 1981; Werren and Beukeboom 1998; Burt and Trivers 2006). Just like for TEs, sexual outcrossing is key in driving the conflict. Ironically, the solution to one kind of genetic conflict is thought to have led to a different kind. Because a diploid zygote is the product of two different genomes, there is a potential conflict between nuclear genes over who makes it into the zygote. This potential conflict is avoided by the imposition of fair meiosis (Leigh 1977; Haig and Grafen 1991; Frank 2003; Brandvain and Coop 2015). The potential conflict between the organellar genomes
from each parent may be circumvented by enforcing uniparental (usually maternal) inheritance of organellar genomes (Eberhard 1980; Cosmides and Tooby 1981, but see Christie et al. 2015). Regardless of the origin, the uniparental inheritance of organelles is in contrast with the bi-parental inheritance of autosomal nuclear genes. As a consequence, there is a well-documented conflict between nuclear and organellear genes over sex determination and ratio (Cosmides and Tooby 1981; Werren and Beukeboom 1998), and several mitochondrial genes in plants are known to cause male sterility (Touzet and Meyer 2014).

The relationship between organellar and nuclear genes is further complicated by the presence of sex chromosomes (Rand et al. 2001). In species where males are the heterogametic sex, i.e. systems with XX females and XY males, any gene on the X chromosome spends 2/3 of its time in females and therefore shares an elevated co-transmission with organellar genes relative to genes on the Y or on autosomes. The extent of the alignment of fitness interests between maternally inherited organellar genes and nuclear gene may therefore depend on whether the nuclear genes are located on autosomes or on one of the sex chromosomes (Figure 1.2).

The asymmetry of co-transmission between organellar and certain nuclear genes has led several authors to examine the genomic signatures of the potential conflict and/or co-adaptation between organellar and nuclear genes (Charlesworth et al. 2010; Drown et al. 2012; Hill and Johnson 2013; Dean et al. 2014; Rogell et al. 2014; Hill 2015). One example of such genomic signature is the chromosomal distribution, i.e., the relative abundance on autosomes and sex chromosomes, of cytonuclear genes (nuclear genes whose protein product is active in an organelle and therefore thought to be of
mitochondrial or chloroplast origin). Studying the chromosomal distribution of cyto-
nuclear genes may therefore provide another window to a central aim of this thesis: how
differences in transmission shape the conflict between mobile genetic elements and the
rest of the genome.

Figure 1.2 Inheritance in plant cells

Plant cells carry genetical material in three compartments: the nuclear, mitochondrial and
chloroplast genomes. Nuclear and organellar genes differ in their mode of inheritance and
so do different chromosomes of the nuclear genome.
1.5 Transposable elements and evolutionary transitions in individuality

The activity of TEs constantly threatens the integrity of genomes. What, then, prevents TEs from completely taking over and causing genomes to fail to produce functional individual organisms? One way to approach this question is to take advantage of the rich theoretical framework developed to study the suppression of selfish behaviour at other biological levels. Throughout evolutionary history, evolutionary transitions in individuality have occurred when units that were previously able to reproduce independently now could only do so as part of a new level of individuality (Buss 1987; Maynard Smith and Szathmáry 1995; Michod 1999). This is what has given life its hierarchical structure: genes in genomes, genomes in cells, cells in multicellular organisms, and multicellular organism in (eu)social groups (Figure 1.3). One of the major achievements of the modern study of social evolution is the insight that whatever level in this hierarchy we are interested in, regardless whether we are studying the origin and maintenance of fair meiosis or the policing of worker eggs in social insects, we are faced with similar conceptual issues (Bourke 2011; Ågren 2013c). In particular, what prevents selfish behaviour at lower levels from disrupting the functionality of higher levels?

The empirical TE literature has a lot to offer this question. By incorporating empirical studies of TEs into the evolutionary transitions framework, we can begin to get a sense of the extent to which the same factors suppress selfish behaviour up and down the hierarchy of life. Reciprocally, conceptualizing TEs this way may also offer new insights into their evolutionary dynamics.
Evolutionary transitions in individuality have given life its hierarchical structure, with layers of cooperation: genes in genomes, genomes in cells, cells in multicellular organisms, and multicellular organism in eusocial groups.

1.6 Study systems

For the projects in this thesis I have used a variety of study systems (Figure 1.4). Below, I provide a brief overview of the plant genera used: *Arabidopsis* (Chapter 3), *Capsella* (Chapters 3-4), *Oenothera* (Chapter 6), and *Rumex* (Chapter 7). More specific information on the species studied and on the data used is provided in the relevant chapters.
1.6.1 *Arabidopsis*

The genus *Arabidopsis*, of the mustard family Brassicaceae, has been a model system in plant genetics ever since mutations began to be systematically scored in the 1940s (Meyerowitz 2001). In particular, the short lifespan, selfing lifestyle, and small genome (long thought to be the smallest of all plants) has made *A. thaliana* the subject of an intensive research programme in molecular and cell biology, and more recently evolutionary biology and ecology. Indeed, these traits were an important reason why *A. thaliana* became the first plant species to have its whole genome sequenced (The *Arabidopsis* Genome Initiative 2000). Since then, the number of *A. thaliana* relatives with whole genome sequences has rapidly grown. By now, *Arabidopsis*, and the Brassicaceae more broadly, has become a rich resource for comparative work on the genomic causes and consequences of a variety of traits, including mating system, polyploidy, and local adaptation (reviewed by Koenig and Weigel 2015). The outcrossing *A. lyrata* became the second *Arabidopsis* species with a reference genome (Hu et al. 2011) and therefore provides an opportunity to begin examining the effect of mating system on TE evolution at a genome-wide scale.

1.6.2 *Capsella*

*Capsella* is another Brassicaceae genus and offers an excellent opportunity to extend the mating system comparisons to more species and time scales. While *A. thaliana* has been selfing for millions of years, the predominately selfing *C. rubella* diverged from its self-incompatible progenitor *C. grandiflora* within the last 200,000 years (Guo et al. 2009; Foxe et al. 2009; Brandvain et al. 2013). Moreover, the selfing *C. orientalis* appears to
have diverged from the same ancestor some 900,000 years ago (Hurka et al. 2011; Douglas et al. 2015), allowing the role of selfing in TE evolution to be studied on different timescales.

1.6.3 Oenothera

The evening primroses, Oenothera (Onagraceae), were the system of choice of Hugo de Vries and provided the basis for his mutation theory of evolution as a rival to Darwinian evolution (Provine 1971). For the purposes of this thesis, the repeated evolution of functional asexuality and evidence that the reduction in the efficacy of selection associated with the shifts is a product of the age of the shift (Hollister et al. 2015) makes Oenothera an ideal system to examine the role of sex on genome size and TE evolution on a genus-wide scale.

1.6.4 Rumex

Rumex (Polygonaceae) is one of only a few plant genera with a chromosome-mediated sex-determination system (Charlesworth 2002; Ruiz Rejón 2004; Charleworth 2015). Rumex hastatulus is thought to have evolved sex chromosomes some 15-16 million years ago (NavajasPerez et al. 2005) but the Y chromosome already shows evidence of degeneration, and a large proportion of X-linked genes are hemizygous (Hough et al. 2014). Data from R. hastatulus were used to examine the chromosomal distribution of cytonuclear genes.
Figure 1.4 Study systems

Species from the study systems used in this thesis: (a) *Arabidopsis thaliana*, (b) *Capsella rubella*, (c) *Rumex hastatulus* (female on the left, male on the right), and (d) *Oenothera biennis*. Photo credit to Jon Ågren (*A. thaliana*), Young Wha Lee and Gavin Douglas (*C. rubella*), Spencer C.H. Barrett (*R. hastatulus*), and Marc T.J. Johnson (*O. biennis*).
1.7 Research objectives

In this thesis, I used a combination of population and comparative genomic approaches to whole genome data from a variety of plant species to make inferences about the role of mating system in TE proliferation. I also investigated how TEs and mating system shifts affect genome size evolution and how the presence of sex chromosomes affect the movement of genes from organellar to nuclear genome. Finally, I synthesized empirical work on TEs with the theoretical frameworks of evolutionary transitions in individuality and social evolution to position TEs within the broader study of conflict and cooperation.

Below, I outline the chapters of my thesis. All chapters have been written as self-contained papers and have been published in peer-reviewed journals (citations are provided). As a consequence, there is some inevitable overlap in the material covered in the introduction and discussion sections of some chapters.

1.7.1 Chapter 2 Sizing up *Arabidopsis* genome evolution

This perspective piece, which I co-wrote with Stephen I Wright, takes its starting point in the recently published *Arabidopsis lyrata* genome, the outcrossing relative of the highly selfing model organism *A. thaliana*, and discusses the role of mating system in causing the higher transposon abundance and larger genome in the outcrosser. It also suggests that the sequencing of the genome of the outgroup species *Capsella rubella* will help determine whether these differences are due to DNA loss in *A. thaliana* versus TE-driven expansion in *A. lyrata*. 

1.7.2 Chapter 3 The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution

In this chapter, I set out to describe abundance and distribution of transposable elements in the reference genome of the selfer *Capsella rubella*. This work was done as part of the *Capsella* genome project and I did some of this work in collaboration with Adrian Platts (McGill University, Canada), Daniel Koenig (Max Planck Institute for Developmental Biology, Tübingen, Germany), and Florian Maumus (URGI, Versailles, France). I used these results to address two questions. First, has the recent (< 200,000 years) shift to selfing in *C. rubella* relative to its closest relative the outcrosser *C. grandiflora* been associated with a shift in TE copy number and expression? Using short genomic and RNA Illumina reads, and a TE database generated together with Daniel Koenig and Florian Maumus, I found no clear evidence for consistent copy number differences between *C. rubella* and *C. grandiflora*. Similarly, although expression comparisons indicate a possible higher variance in TE expression in *C. rubella*, there was no evidence for consistently higher TE expression in *C. grandiflora*. In addition, TE density along the chromosomes is very similar in the two species. Overall, while other analyses found evidence that the evolution of selfing in *C. rubella* has been associated with a rapid reduction in genome-wide levels of purifying selection, as well as a distinct shift in expression of genes linked to flower morphology, we do not detect a parallel change in TE abundance or expression.
Second, is the difference in TE abundance and genome size between the two Arabidopsis species due to TE-driven expansion in the outcrosser A. lyrata, and/or genome loss in the selfer A. thaliana? I used the genome sequence of C. rubella as an outgroup to polarize derived genomic changes in the two species. Using three-species whole genome alignments (generated by Adrian Platts) I found that many regions showed evidence for derived genome expansion in the outcrossing species, and there was a strong correlation between parts of the A. lyrata genome that had expanded and transposon abundance.


1.7.3 Chapter 4 Mating system shifts and transposable element evolution in the plant genus Capsella

In this chapter, I expanded the Capsella mating system comparison to include sequence data from multiple individuals of C. rubella and C. grandiflora, as well as from the older selfer C. orientalis. The goal was to further explore whether evolution of self-fertilization from outcrossing will generally result in an increase or decrease in TE abundance and the
timescale over which this change occurs. I used *Capsella rubella* reference genome and short Illumina reads to quantify abundance, genome distributions, and population frequencies of TEs in all species. I found that while the older selfer has very low TE abundance, the younger selfing species has more TEs overall than the outcrossing progenitor. Moreover, the outcrosser appeared to have higher current TE activity than either of the two selfers. I used these observations to argue that selfing may have different effects on short and long time scales.


### 1.7.4 Chapter 5 Co-evolution between transposable elements and their host: a major factor in genome size evolution?

In this chapter, Stephen Wright and I review the extensive empirical and theoretical literature on TEs and genome size to argue that genome size changes may in part be driven by the dynamics of co-evolution between TEs and their hosts. We also outline scenarios when this perspective may lead to novel predictions about the evolutionary dynamics of genome size.

1.7.5 Chapter 6 No evidence that sex and transposable elements drive genome size variation in evening primroses

In this chapter, I set out to explore the role of sex and TE abundance in genome size evolution at a larger time scale, and if the effect of asexuality on genome size depends on the time since the shift to asexuality. I used a phylogenetic comparative approach and whole genome sequencing to investigate the contribution of sexual reproduction and TE content to genome size variation in the evening primrose (*Oenothera*) genus. I found that variation in sexual/asexual reproduction cannot explain the almost two-fold variation in genome size and that genome size was not associated with transposable element abundance; instead the larger genomes had a higher abundance of simple sequence repeats.


1.7.6 Chapter 7 Chromosomal distribution of cytonuclear genes in a dioecious plant with sex chromosomes

This chapter focuses on a different kind of gene movement: the transfer of genes from organellar (mitochondrial and chloroplast) genomes to the nuclear genome since the origin of the mitochondrial and chloroplast genomes. Together with Josh Hough (PhD Candidate, Stephen I. Wright and Spencer C.H. Barrett Labs, University of Toronto) I tested the hypotheses that the elevated co-transmission between the X chromosome and
organellar genes (2/3) compared with autosomal genes (1/2) may lead to either an
overrepresentation of nuclear-mitochondrial and nuclear-chloroplast (i.e., cytonuclear)
genes on the X chromosome relative to autosomes (the co-adaptation hypothesis), or an
underrepresentation due to selection for reduced male mutational load (the sexual-conflict
hypothesis). Using transcriptome sequences from *Rumex hastatulus* (Polygonaceae), we
found no evidence of under- or overrepresentation of either nuclear-mitochondrial or
nuclear-chloroplast genes on the X chromosome, and thus no support for either the co-
adaptation or the sexual-conflict hypothesis.

This study was published in *Genome Biology and Evolution* (Hough J†, Ågren
JA‡, Barrett SCH, and Wright SI. 2014. Chromosomal distribution of cyto-nuclear genes
in a dioecious plant with sex chromosomes. *Genome Biology and Evolution*. 6: 2439–
2443, †= equal contribution).

1.7.7 Chapter 8 Evolutionary transitions in individuality: insights from
transposable elements

In this chapter, I review data on TEs in the light of the theoretical framework of
evolutionary transitions in individuality and argue that the rapid influx of whole genome
data has created a rich empirical data source that can be used to examine general
questions related to evolutionary transitions in individuality and social evolution. This
chapter has been published in *Trends in Ecology and Evolution* (Ågren JA. 2014.
Evolutionary transitions in individuality: insights from transposable elements. *Trends in
CHAPTER 2 SIZING UP ARABIDOPSIS GENOME EVOLUTION

Plant genomes vary tremendously in size, even among close relatives (Bennett and Leitch 2010). In addition, there is considerable variation in related features such as chromosome number and size, number of genes and transposable element (TE) content. There have been several attempts to shed light on both the relative importance of these contributors, and the relative role of mutation versus natural selection, in driving genome size evolution. However, our ability to understand the evolution of genome size to date has been limited, due in part to the lack of large-scale genomic information from closely related species. With the recent publication of the complete genome of the primarily self-incompatible plant Arabidopsis lyrata (Hu et al. 2011), a close relative of the workhorse of plant genetics Arabidopsis thaliana, many of these questions can now be approached with increased strength under a comparative framework.

Several theoretical predictions have been made concerning genome size evolution. First, if genome expansion is governed by slightly deleterious mutations, species with larger effective population sizes (\(N_e\)) are expected to experience more efficient selection, and therefore maintain smaller genomes (Lynch and Conery 2003). Second, asexual and highly selfing species are expected to experience reduced activity of TEs, which could contribute to genome shrinkage in these organisms (Bestor 1999). This second prediction is somewhat at odds with the first, as the effective population size of selfing species should be reduced by at least half that of outcrossing species (Nordborg 2000). Third, organisms with relatively quick development times tend to have smaller genomes, and there may be stronger selection for a compact genome in faster growing
organisms (Pagel and Johnstone 1992). Finally, species may differ in their relative rates of insertion and deletion, thereby driving genome expansion versus contraction under a purely neutral process (Petrov et al. 2000).

It has long been known that the genome of *A. thaliana* genome was one of the smallest in angiosperms, and indeed this was a major reason for its adoption as a model system. Although the other species in the genus also have relatively small genomes, the *A. thaliana* genome is reduced by almost one half compared with the others, suggesting recent and rapid DNA loss. Moreover, compared with its closest relatives, the species also shows a number of relevant shifts in life history; it is highly selfing, in contrast with its predominantly self-incompatible and outcrossing relatives; it is annual rather than perennial, and patterns of nucleotide diversity suggest that it has a relatively small effectively population size (Clark et al. 2007; Ross-Ibarra et al. 2008), although diversity levels are still reasonably high. Thus, comparative genomics in this system provides an exciting framework for investigating the prime factors driving genome size evolution.

Despite the general similarity in gene order and a high level of sequence similarity in genes, Hu et al. (2011) report a difference of approximately 80 MB (over 200 MB compared with 125 MB) in genome size between *A. thaliana* and *A. lyrata*. What is perhaps most striking about the genome size shift is the consistency of genome loss; there are clear reductions in size due to chromosome rearrangements, TE copy number, small and large deletions, and even gene number. Furthermore, with the exception of single base pair deletions, the DNA size change is apparent for deletion/insertion events at all size ranges, although it is especially exaggerated at the larger size range.
What processes may be responsible for this genome loss? Clearly, the directionality is at odds with predictions based on differences in effective population size, as there is consistent evidence for a moderate loss of genetic variation in *A. thaliana*. In contrast, the global nature of genome size reduction suggests that directional shifts in mutation pressure and/or stronger selection favouring deletions over insertions could have driven genome shrinkage in *A. thaliana*. Population genetic analysis can help distinguish these possibilities. If selection were acting against insertions but favouring deletions, we would expect to detect higher population frequencies of deletions compared with insertions. The authors analyzed polymorphism data and found support for this. By using *A. lyrata* to polarize the directionality of polymorphic size changes in *A. thaliana*, they show that insertions segregate at much lower population frequencies than deletions, suggesting ongoing selection pressures against insertions and in favour of deletions.

However, analysis of TEs suggests a somewhat different interpretation. Consistent with the general patterns, TE copy number is consistently higher in *A. lyrata* regardless of element type, ruling out a major expansion of any particular class of TE. In total, at least 30% of the *A. lyrata* genome is comprised of TEs, in contrast with 24% in *A. thaliana*. Analyses of sequence divergence of long-terminal repeats of retrotransposons and TE phylogenies suggest an enrichment of younger TE insertions in *A. lyrata* compared with *A. thaliana*. This suggests higher rates of TE activity in the outcrossing species, consistent with theoretical predictions about reduced genomic conflict following the evolution of selfing (Bestor 1999). If a higher genome-wide rate of fixation of deletions is the only process governing genome loss in *A. thaliana*, we would expect the opposite pattern: fewer old insertions and a shift toward more young
insertions. Furthermore, previous work suggests that TE insertions segregate at lower population frequencies in *A. lyrata* than in *A. thaliana* (Wright et al. 2001; Lockton and Gaut 2010), consistent with either stronger selection against TE insertions in *A. lyrata* or a reduced rate of transposition in *A. thaliana* (Wright et al. 2001). Together with the genomic data and recent analysis of expression levels (Hollister et al. 2011), the transposon patterns imply reduced activity in *A. thaliana* relative to *A. lyrata*. As one possible explanation, transposon silencing has been suggested to be more efficient in *A. thaliana*, although whether this is a cause or an effect of having more young insertions in *A. lyrata* remains unclear (Hollister et al. 2011).

We are therefore left with evidence for two processes explaining differences in genome size, stronger selection for DNA removal and/or lower rates of TE activity. It is unclear whether the patterns could in fact be driven by a single mechanism. For example, higher TE activity via outcrossing in *A. lyrata* may generally lead to higher rates of gene duplication and a greater abundance of small pieces of neutral DNA. However, there are several future studies that could help distinguish whether either, or both, of these factors are contributing. If genome size evolution is mainly because of the differences in TE activity, we would expect polymorphism analysis in *A. lyrata* to reveal similar selection pressures on small insertions and deletions at homologous sequences, whereas basic transposition rates should differ between species. Alternatively, if selection favouring deletions in *A. thaliana* is the primary cause, we would expect similarly-aged TE insertions to show a higher rate of DNA loss in the selfing species. Finally, the sequencing of genome of the outgroup species *Capsella rubella* will be crucial for
determining what factors are due to DNA loss in *A. thaliana* versus expansion in *A. lyrata*.

The contribution of *A. thaliana* genome sequencing to our understanding of plant genetics can hardly be exaggerated. The sequencing of the *A. lyrata* genome opens up many new avenues of research. In particular, the presence of the genome of two closely related species will create many possibilities to examine genome consequences of recent evolutionary changes, such as change alterations in mating system, life history and effective population size.
CHAPTER 3 THE \textit{CAPSELLA RUBELLA} GENOME AND THE GENOMIC CONSEQUENCES OF RAPID MATING SYSTEM EVOLUTION

Abstract

The shift from outcrossing to selfing is common in flowering plants (Stebbins 1957; Barrett 2002) but the genomic consequences and the speed at which they emerge remain poorly understood. An excellent model for understanding the evolution of self-fertilization is provided by \textit{Capsella rubella}, which became self-compatible <200,000 years ago. We report a \textit{C. rubella} reference genome sequence and compare RNA expression and polymorphism patterns between \textit{C. rubella} and its outcrossing progenitor \textit{Capsella grandiflora}. We found a clear shift in the expression of genes associated with flowering phenotypes, similar to that seen in \textit{Arabidopsis}, in which self-fertilization evolved about 1 million years ago. Comparisons of the two \textit{Capsella} species showed evidence of rapid genome-wide relaxation of purifying selection in \textit{C. rubella} without a concomitant change in transposable element abundance. Overall we document that the transition to selfing may be typified by parallel shifts in gene expression, along with a measurable reduction of purifying selection.
3.1 Main

The switch from obligatory outcrossing to predominant self-fertilization in plants is one of the most striking and repeated examples of convergent evolution (Stebbins 1957; Barrett 2002). Selfing is thought to be favoured because of its inherent transmission advantage, as well as the advantage of assured reproduction when mates, pollinators or both are scarce. Selfing should evolve whenever these advantages outweigh the costs associated with inbreeding depression (Charlesworth and Willis 2009). In contrast to the immediate benefits of selfing, reduced effective recombination rates, greater population subdivision and more frequent genetic bottlenecks may incur longer-term costs as a result of reductions in effective population size and selective interference among linked sites (Charlesworth and Wright 2001), all of which are potential contributors to the high rates of extinction of selfing lineages (Lynch et al. 1995). A key problem in understanding the causes and consequences of the evolution of selfing has been partitioning the changes that occurred after the mating system evolution, as many species diverged before the evolution of selfing. As an example, Arabidopsis thaliana probably became selfing only several million years after it was established as a separate species (Tang et al. 2007; Ossowski et al. 2010).

A unique opportunity to understand the evolution of selfing is offered by the genus Capsella, which is from the same family as Arabidopsis. The highly selfing species C. rubella, found throughout much of southern and Western Europe, separated less than 200,000 years ago from the self-incompatible, obligate outcrosser C. grandiflora, which is restricted primarily to the northwest of Greece (Guo et al. 2009; Foxe et al. 2009). In
contrast to *Arabidopsis*, the breakdown of self-incompatibility in *Capsella* was concurrent with species divergence (Guo et al. 2009; Foxe et al. 2009; St Onge et al. 2011).

Table 3.1 *Capsella rubella* genome annotation results

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary protein coding loci (n)</td>
<td>26,521</td>
</tr>
<tr>
<td>Alternatively spliced gene models (n)</td>
<td>1,926</td>
</tr>
<tr>
<td>Alternatively spliced gene models (n)</td>
<td>1,926</td>
</tr>
<tr>
<td>Average number of exons</td>
<td>5.4</td>
</tr>
<tr>
<td>Median exon length (bp)</td>
<td>150</td>
</tr>
<tr>
<td>Median intron length (bp)</td>
<td>103</td>
</tr>
<tr>
<td>microRNA genes (n)</td>
<td>86</td>
</tr>
</tbody>
</table>

We shotgun sequenced the genome of the *C. rubella* reference line Monte Gargano (Italy) to 22× coverage using a combination of platforms (Supplementary Table 3.1 and Supplementary Note 3.1). For the final assembly of 134.8 Mb, covering all eight chromosomes, we used a genetic map with 768 markers11 (Supplementary Figs. 3.1 and 3.2, Supplementary Tables 3.2, 3.3, 3.4, 3.5 and Supplementary Note 3.1), *Arabidopsis lyrata* synteny (Supplementary Fig. 3.3) and BAC and fosmid paired-end link support (Supplementary Note 3.1 and Supplementary Fig. 3.4). We predicted 28,447 transcripts from 26,521 protein-coding genes and 86 microRNA loci (Table 3.1). We also conducted *de novo* genome assemblies from Illumina libraries for an outbred *C. grandiflora*.
accession and the close outgroup species *Neslia paniculata* (Supplementary Note 3.1, Supplementary Fig. 3.5 and Supplementary Tables 3.6, 3.7, 3.8, 3.9).

Although the *C. rubella* genome assembly is ~40% shorter than the nuclear DNA content estimated from flow cytometry, 219-Mb, k-mer analysis and remapping of Illumina reads indicated that the assembly encompasses most of the euchromatin (Supplementary Note 3.1, Supplementary Table 3.10 and Supplementary Fig. 3.6). Almost half of the 219-Mb genome seems to be repetitive, including centromeric satellite repeats (Supplementary Fig. 3.7). Apart from the centromeres, fluorescence in situ hybridization (FISH) identified the 45S ribosomal DNA (rDNA) arrays on chromosomes 2, 3 and 7, the 5S rDNA locus on chromosome 5 and an interstitial telomeric sequence in the pericentromeric region of chromosome 7 as notable genomic locations of repeats (Fig. 3.1a,b, Supplementary Note 3.1 and Supplementary Fig. 3.8).
Figure 3.1 Genomic structures, chromosome painting and comparative genomic mapping in *C. rubella*, *A. lyrata* and *A. thaliana*

(a) Comparative genome structure and major chromosome landmarks in *C. rubella* (CR). The 24 ancestral genomic blocks are indicated by uppercase letters (A–X) and are
colored according to their position on the eight chromosomes of the ancestral crucifer karyotype (ACK12). (b) Comparative chromosome painting of CR1 and CR2. Differentially labeled *A. thaliana* BAC contigs corresponding to the genomic blocks A, B, C, D and E were used as painting probes on the pachytene bivalents of CR1 and CR2. The true fluorescence signals were pseudocolored according to the color code used in a. The arrowheads indicate unpainted pericentromeric heterochromatin. Scale bars, 10 µm.

(c,d) Comparative genome mapping of *C. rubella* with *A. lyrata* (AL; c) and *A. thaliana* (AT; d). The outer ring shows the percentage of the genomic window that comprises transposable elements, with a maximum at 60% coverage, the second ring shows gene density, and the inner ring shows orthologous regions between species on the basis of whole-genome alignment and orthologous chaining. Note that the *A. lyrata*, but not the *C. rubella*, assembly includes gaps for inferred centromeric heterochromatin. From synteny analyses of the three species, the approximate gene intervals contained within each block include: A/B, AT1G01010–AT1G36980; C, AT1G41830–AT1G56200; D, AT1G56210–AT1G64720; E, AT1G6364790–AT1G680950; F, AT3G01070–AT3G25530; G, AT2G04039–AT2G07050; H, AT2G10870–AT2G20900; I, AT2G20920–AT2G26430; J, AT2G26670–AT2G48160; K, AT2G01060–AT2G04038; L, AT3G25545–AT3G32980; M/N, AT3G42170–AT3G63490; O, AT4G00026–AT4G05530; P, AT4G06534–AT4G12590; Q/R, AT5G01010–AT5G30510; S, AT5G32440–AT5G42110; T/U, AT4G12640–AT4G40100; V, AT5G42140–AT5G47760; W/X, AT5G47800–AT5G67640.

Consistent with previous findings (Schranz et al. 2006), we found that the large-scale synteny between *C. rubella* and *A. lyrata* is almost complete (Fig. 3.1c and Supplementary Note 3.1). Comparisons with *Schrenkiella parvula* (Supplementary Fig. 3.9) indicated that all three major differences between *C. rubella* and *A. lyrata* are either specific to the *A. lyrata* genome or errors in the *A. lyrata* assembly (Supplementary Note 3.1). Further comparisons delimited the breakpoints of major rearrangements in *A. thaliana* (Fig. 3.1d; Schranz et al. 2006). Overall, we conclude that *C. rubella*, despite having gone through an extreme genetic bottleneck, retains a largely ancestral genome structure.

To investigate the functional consequences of the mating system change, we compared flower transcriptomes from four *C. rubella* and four *C. grandiflora* accessions.
Many *C. rubella* alleles are found in *C. grandiflora* (Guo et al. 2009; Foxe et al. 2009); thus, DNA sequence variation should not confound RNA sequencing (RNA-seq) comparisons. RNA expression levels were more highly correlated within (average Pearson correlation coefficients, 0.95 for *C. grandiflora* and 0.94 for *C. rubella*) than between species (average Pearson correlation coefficient, 0.82; Supplementary Fig. 3.10). We identified 246 genes that were expressed more strongly in *C. rubella* relative to *C. grandiflora* and 373 that were expressed more weakly relative to *C. grandiflora*, with a minimum fold change of 1.5, false discovery and significance thresholds of 0.5% and a minimum normalized expression of 19 (Supplementary Note 3.1). The set was enriched for Gene Ontology terms related to floral development and growth functions, which is consistent with changes in reproductive organ size and development between species (Supplementary Table 3.11). One-hundred fifty-eight differentially expressed genes colocalized with interspecific quantitative trait loci that are responsible for differences in petal size and pollen number11 (Supplementary Table 3.12), with 17 found within 2 Mb of petal size quantitative trait loci peaks. Thus, some of the expression changes may be due to cis-regulatory changes that had a role in floral evolution. Pathway analyses (Supplementary Note 3.1) identified a reduction in brassinosteroid signaling, which is involved in hormone-triggered pollen maturation (Ye et al. 2010), in *C. rubella* (Supplementary Table 3.13).

To investigate whether the *C. rubella* and *C. grandiflora* pair is representative of the expression differences in flowers of closely related selfers and outcrossers, we compared the selfer *A. thaliana* and the predominantly outcrossing *A. lyrata*. We found that the overlap in expression changes between the two species pairs was much higher.
than that expected by chance. For example, of 373 genes that were expressed more strongly in *C. rubella*, 75 orthologs were also expressed more strongly in *A. thaliana* compared to *A. lyrata*, whereas only 16 showed higher expression in *A. lyrata*. In contrast, of 246 genes that were expressed more weakly in *C. rubella*, 46 orthologs were also expressed more weakly in *A. thaliana*, whereas only 12 showed lower expression in *A. lyrata* (Fisher's exact test $P < 1 \times 10^{-13}$; Fig. 3.2). These results suggest that parallel floral evolution in selfers may be associated with parallel changes in gene expression. A caveat is that some of these changes could reflect the altered abundance of specific tissue types because of changes in flower morphology.

![Figure 3.2 Evolution of gene expression in selfing and outcrossing *Capsella* and comparisons to *Arabidopsis*](image)

**Figure 3.2 Evolution of gene expression in selfing and outcrossing *Capsella* and comparisons to *Arabidopsis***

Distribution of fold changes in gene expression in *C. grandiflora* relative to *C. rubella* (x-axis) and *A. lyrata* relative to *A. thaliana* (y-axis) at genes showing significant down-regulation or significant up-regulation in *C. rubella*. 
Population genetic theory predicts that selfers should accumulate slightly deleterious mutations because of a reduced effective population size (Charlesworth and Willis 2009). To test this hypothesis, we characterized genome-wide patterns of coding sequence polymorphisms discovered in RNA-seq data from five outbred *C. grandiflora* (ten haploid chromosomes) and six *C. rubella* individuals (Supplementary Note 3.1 and Supplementary Table 3.14). We identified 48,518 high-quality SNPs in 4,225 genes. The vast majority (81%) segregated only in *C. grandiflora*. Of the remainder, 7% segregated in both species, 8% segregated only in *C. rubella* and only 4% were fixed between the two species. On average, diversity at synonymous sites was 0.02 in *C. grandiflora*, whereas it was sixfold lower (0.003) in *C. rubella* (Fig. 3.3a), which is similar to previously documented differences in smaller gene sets (Schranz et al. 2006; St Onge et al. 2011; Slotte et al. 2012).
Figure 3.3 Polymorphism comparisons in *C. rubella* and *C. grandiflora*

(a) Average pairwise differences (π) at nonsynonymous (nonsyn) and synonymous (syn) sites. Error bars indicate standard errors across all loci. Cr, *C. rubella*; Cg, *C. grandiflora*. (b) Ratio of nonsynonymous to synonymous polymorphisms at each derived frequency class using data subsampled to six chromosomes per species. *N. paniculata* was used as an outgroup to infer derived status. (c) Proportion of synonymous and nonsynonymous polymorphisms unique to each species, as well as shared and fixed differences. Simulated (sim) values are from forward computer simulations using the inferred demographic model and strength of selection on nonsynonymous sites (see Main text). Obs, observed.

To test for a change in the efficacy of selection, we examined the ratios of nonsynonymous to synonymous polymorphisms. The ratio was much higher for *C. rubella*–specific (0.68) than *C. grandiflora*–specific SNPs (0.35; two-tailed Fisher's exact test *P* < 0.0001). This was true for all site frequency classes when subsampling the data to equivalent haploid sample sizes (Fig. 3.3b), suggesting that relaxed selection is not simply due to rare variants in *C. rubella* having experienced less purifying selection after the recent genetic bottleneck that purged the majority of variation. In contrast, fixed SNPs behaved similarly to *C. grandiflora*–only SNPs, with a non-synonymous-to-synonymous ratio of 0.38, whereas SNPs segregating in both species had the lowest ratio
(0.22), consistent with these being the oldest on average and thus having experienced the most selection.

There are at least three potential explanations for the elevated non-synonymous-to-synonymous ratio in C. rubella. The first is experimental error: because diversity is generally lower in C. rubella, SNP errors relative to true polymorphisms may inflate the relative measure of non-synonymous variation. However, dideoxy sequencing indicated that the false positive rate was less than $2 \times 10^{-3}$ per SNP, which is much lower than the C. rubella polymorphism density (Supplementary Note 3.1). Another explanation could be that the distribution of selection coefficients is altered because the evolution of selfing in C. rubella is associated with changes in morphology, life history, habitat and range. However, extensive population genetic modelling and computer simulations indicated that such a scenario is not required to explain the data (Fig. 3.3c, Supplementary Note 3.1, Supplementary Fig. 3.11 and Supplementary Tables 3.15 and 3.16). The elevated non-synonymous-to-synonymous ratio in C. rubella could also be due to a reduced efficacy of purifying selection caused by the demographic and selective effects associated with the shift to selfing, as is predicted by theory (Charlesworth and Wright 2001). Our results seem to be the most consistent with this third hypothesis. Most previous work has found little or no evidence for relaxed selection on non-synonymous sites in selfing lineages (Wright et al. 2002; Haudry et al. 2008; Escobar et al. 2010); our use of genome-wide polymorphism data to quantify current selection pressures in a recently derived selfing lineage provides a more powerful test of this hypothesis, suggesting that selfing lineages may in fact experience considerable accumulation of deleterious mutations even over short timescales.
Apart from genome-wide changes in the non-synonymous-to-synonymous ratio, a potential consequence of the transition to selfing is a change in the abundance and distribution of transposable elements. Not only are self-replicating transposable elements expected to spread more efficiently through the genomes of highly outcrossing species, but self-regulated transposition is also more probable in selfers (Charlesworth and Langley 1986; Wright and Schoen 1999). In agreement, many transposable element classes have fewer members in the *A. thaliana* than in the *A. lyrata* genome (Hu et al. 2011). Despite its nuclear DNA content being in the same range as that of *A. lyrata*, the *C. rubella* genome is more similar to that of *A. thaliana* in several features related to transposable element frequency and density (Fig. 3.4). First, intergenic distances are more similar in *C. rubella* and *A. thaliana* (Fig. 3.4a). Across all chromosomal blocks, the *A. lyrata*–to–*C. rubella* ratios were positive, with a mean of 1.6, whereas the mean for *A. thaliana* compared to *C. rubella* was 0.95. Thus, intergenic space has either shrunk in *A. thaliana* and *C. rubella* or has expanded in *A. lyrata*. Similarly, transposable element density is low in the *C. rubella* assembly (Fig. 3.4b), particularly in gene-rich regions (Fig. 3.1), and is more comparable to that in *A. thaliana* (Fig. 3.4c). This suggests that genome expansion and contraction have occurred in different regions, with *C. rubella* having a compact euchromatic region comparable to *A. thaliana*, whereas *A. lyrata* has experienced greater recent transposable element activity near genes.
Figure 3.4 Evolution of genome structure and transposable element abundance in *Capsella* and *Arabidopsis*

(a) Slope of the physical positions in orthologous blocks between *C. rubella* and *A. thaliana* and *A. lyrata*. Ancestral orthologous blocks are labeled on the x axis (see Fig. 3.1). (b) Transposable element genomic coverage in the three species. LINE, long interspersed nucleotide repetitive elements; SINE, short interspersed nucleotide repetitive elements. (c) Distribution of the distances between transposable elements and their nearest protein-coding genes. (d) Age distribution of full-length LTR retrotransposons estimated using the rate of substitution between LTRs of individual insertions and assuming a substitution rate of $7 \times 10^{-9}$ per bp per generation. MYA, million years ago.
Given that the structure of gene-rich regions in the *C. rubella* genome is similar to that in *A. thaliana*, we tested whether the shift to selfing was associated with a rapid loss of transposable element abundance, activity or both. The age distribution of long terminal repeat (LTR) retrotransposons in *C. rubella* seemed to be similar to that in *A. thaliana*, with no evidence for the high rate of recent transposition seen in *A. lyrata* (Fig. 3.4d). Only 5% of the full-length LTR retrotransposons seemed to be younger than 100,000 years, which is close to the estimated speciation time (Supplementary Table 3.10), suggesting that the vast majority of transposition occurred before the shift to selfing. However, identification of transposable element insertions using paired-end genomic Illumina sequencing revealed no clear evidence for consistent copy number differences between *C. rubella* and *C. grandiflora* (Fig. 3.5a and Supplementary Note 3.1). Similarly, although expression comparisons indicate a possible higher variance in transposable element expression in *C. rubella*, there is no evidence for consistently higher transposable element expression in *C. grandiflora* (Wilcoxon rank sum test $P = 0.4857$; Fig. 3.5b). In addition, transposable element density along the chromosomes is very similar in the two species (Fig. 3.5c). Given that species divergence is low relative to the coalescent history of *C. grandiflora* (that is, most of the *C. rubella* alleles are shared with *C. grandiflora*), it is probable that longer timescales are required for transposable element copy number to diverge noticeably. Thus, our analyses suggest little evidence for large-scale changes in transposable element abundance, as the evolution of selfing in occurred about 100,000 years ago, and imply that transposable element activity may be specifically elevated in *A. lyrata* (Hollister et al. 2011).
Figure 3.5 Estimates of transposable element copy number and expression in *C. rubella* and *C. grandiflora*

(a) Numbers of insertion sites identified using read mapping of paired-end Illumina genomic data (using Popoolation TE (Kofler et al. 2012)) in two *C. rubella* accessions and two *C. grandiflora* accessions. SINE/LINE, SINE and LINE. (b) Mean and standard error of the proportion of RNA-seq transcripts mapping to transposable elements. (c) Distribution of transposable element insertions along chromosome 1 in two *C. rubella* accessions and two *C. grandiflora* accessions.
*Capsella rubella* is a young species with an origin that is probably associated with a severe founder event and a shift to a highly selfing mating system. These recent and correlated events have had genome-wide consequences that range from divergence in gene expression for a suite of reproductively related genes to a genome-wide decline in the efficacy of natural selection on amino acid polymorphisms. Moreover, our comparisons among three closely related species, *A. thaliana*, *A. lyrata* and *C. rubella*, highlight the fluidity of large-scale genome structure, typified by differential expansion of centromeric repeats and changes in transposable element activity. The factors driving such contrasting modes of genome expansion and shrinkage are far from resolved, and it will be important to broaden future comparisons to larger phylogenetic scales to better understand the processes driving genome structure evolution.

3.2 Methods

3.2.1 Genome assembly and annotation

3.2.1.1 Sequencing

Whole-genome shotgun sequencing of *C. rubella* was conducted using the Monte Gargano (Italy) reference strain. The majority of the sequencing reads were collected with standard Sanger sequencing protocols and Roche 454 XLR and Illumina GAIIx machines at the DoE JGI in Walnut Creek, California (JGI sequencing protocols; see URLs). One linear Roche 454 library (ten runs, 2.78 Gb), three 2.5-kb insert size paired libraries (three runs, 713.1 Mb), two 4-kb insert size paired libraries (four runs, 434.5
Mb), one 8-kb insert size library (three runs, 935.4 Mb) and one 10-kb insert size library (two runs, 515.4 Mb) were sequenced with standard XLR protocols.

3.2.1.2 Genome assembly and construction of pseudomolecule chromosomes

The sequence reads were assembled using our modified version of Arachne v.20071016 (Jaffe et al. 2003). This produced 1,859 scaffold sequences, with a scaffold L50 value of 6.7 Mb, 71 scaffolds larger than 100 kb and a total genome size of 137.1 Mb. Scaffolds were screened against bacterial proteins, organelle sequences and the GenBank nucleotide repository (nr) and removed if they were found to be contaminants. Additional scaffolds were removed if they (i) consisted of >95% 24-mers that occurred four other times in scaffolds larger than 50 kb, (ii) contained only unanchored RNA sequences or (iii) were less than 1 kb in length.

The combination of 768 markers in an Illumina-based genetic map (Supplementary Note 3.1), A. lyrata synteny derived from whole-genome alignments and BAC and fosmid paired-end link support was used to identify misjoins in the assembly. Scaffolds were broken if they contained syntenic and linkage group discontiguity coincident with an area of low BAC and fosmid coverage. To avoid bias, discontiguity on the basis of synteny alone was not deemed sufficient for breaking scaffolds. A total of 16 breaks were executed, and 45 broken scaffolds were oriented, ordered and joined using 37 joins to form the final assembly containing eight pseudomolecule chromosomes. The
final assembly contained 853 scaffolds (9,675 contigs) that covered 134.8 Mb of the genome with a contig L50 value of 134.1 kb and a scaffold L50 value of 15.1 Mb. Except for a single potential misplacement of a small chromosome 7 contig on chromosome 4, the assembly was supported by an independent map with 999 markers within *C. rubella*, with 194 F2 individuals from the cross of 1408 and Monte Gargano (Guo et al. 2012).

Completeness of the euchromatic portion of the genome assembly was assessed using 1,167 full-length cDNAs from *C. rubella* along with an 108-bp Illumina EST library from *C. grandiflora* leaf material (plants grown at the University of Toronto and library construction and sequencing conducted at the Genome Quebec Innovation Centre in Montreal). The aim of this analysis was to obtain a measure of the completeness of the assembly rather than a comprehensive examination of gene space. The screened alignments indicated that 1,166 of 1,167 (99.74%) of the full-length cDNAs aligned to the assembly, and 27,876 of 32,766 (86.29%) of the Illumina ESTs aligned. The ESTs that did not align were checked against the NCBI nr, and a large fraction was found to be prokaryotic rDNA.

### 3.2.1.3 Annotation

Protein-coding genes were predicted with a pipeline that combines ESTs, homology and *de novo* prediction methods (Goodstein et al. 2012). Three-hundred twenty-nine million 108-bp-long single-ended Illumina reads from ten libraries of *C. grandiflora* cDNAs were assembled with tophat 1.3.0 and cufflinks 1.0.3, aligned to the *Capsella* genome and assembled with PASA (Haas et al. 2003) (alignments required 95% identity and 50%
length), generating 28,322 EST assemblies with a median length of 1,426 bp. These were aligned to the *Capsella* genome (requiring 95% sequence identity and 50% coverage of the input sequence) and further assembled with PASA (Haas et al. 2003) to generate 27,399 EST assemblies with a median length of 1,432 bp. Predicted protein sequences from *Arabidopsis* (v. TAIR10), papaya (ASGPB v0.4 Dec2) and grapevine (Genoscope 12 × 05/10/10) to the softmasked *Capsella* v1.0 assembly with gapped blastx (Altschul et al. 1990) were aligned to generate putative protein-coding gene loci from regions with EST assemblies, protein homology or both, extending to include overlap where necessary. Gene predictions were generated from putative loci with FGeneS+ (Solovyev et al. 2006), exonerate (Slater and Birney 2005; with setting –model protein2genome) and GenomeScan (Yeh et al. 2001). The gene prediction at each locus with the highest amount of support from EST assemblies and protein homology was chosen to be improved using evidence from the EST assemblies with a second round of PASA. Gene models with homology to repeats were removed. This produced an annotation at each of 26,521 protein-coding loci, with 1,926 alternative splice forms predicted to produce a total of 28,447 transcripts.

3.2.1.4 Whole-genome alignment

Two approaches were used to conduct whole-genome alignments and identify orthologous gene positions. First, CoGe (Lyons et al. 2008) was used, using the quota align algorithm in Synmap, to identify orthologous gene positions across the genomes of *C. rubella, A. thaliana, A. lyrata* and *S. parvula*. 
Additionally, to gain more comprehensive synteny information for gene-poor regions, whole-genome orthologous alignments were generated for *C. rubella*, *A. thaliana* and *A. lyrata*. To generate three-species whole-genome alignments, LASTZ (Harris 2007) alignments and orthologous chaining (Kent et al. 2003) were conducted using *Capsella* as the reference, retaining only a single chain per species per region. Version 0.62-1 of the circos software (Krywinski et al. 2009) was used to create the circular plots. For comparative mapping in the circos plots, orthologous chains were filtered to a minimum length of 100 kb.

3.2.1.5  RNA extraction, sequencing and read mapping

Total RNA was harvested from mixed flower buds flash frozen in liquid nitrogen from five *C. grandiflora* genotypes from Greece and six *C. rubella* genotypes sampled from different geographic locations (Supplementary Table 3.2) using the RNAeasy plant mini kit (Qiagen) with minor modifications to obtain the required yield (~5 µg) for RNA sequencing. After extraction, mRNA isolation, library preparation and paired-end 38-bp read sequencing were conducted on three flow cells of an Illumina Genome Analyzer (GAIi) at the Centre for Analysis of Genome Evolution and Function (CAGEF) at the University of Toronto.

To compare our expression patterns in *Capsella* with those in *Arabidopsis*, we prepared duplicate RNA from stage-12 floral buds of *A. thaliana* (Col-0 reference accession) and *A. lyrata* (accession MN47; Hu et al. 2011)). Barcoded RNA-seq libraries were prepared from each sample with an adaptation of the standard Illumina mRNA-seq
method, and single-end sequencing (one flowcell lane worth) was performed to generate 78-bp reads on an Illumina Genome Analyzer (GAII). Methods for plant growth, floral bud collection, RNA extraction, library construction and sequencing for the Arabidopsis samples are as reported in Gan et al. (2011).

We used TopHat (Trapnell et al. 2012) to map sequence reads to the reference genome of C. rubella, A. thaliana and A. lyrata RNA sequence reads were mapped to their respective reference genomes using the same parameters described above.

3.2.1.6 Genomic resequencing and analysis

Genomic extraction of leaf material was conducted for two C. grandiflora accessions and three C. rubella accessions, as well as one accession of N. paniculata, using a modified CTAB protocol; 108-bp paired-end genomic sequencing was conducted at the Genome Quebec Innovation Centre, and 150-bp sequencing was performed at the Max Planck Institute for Developmental Biology. To infer the ancestral state of segregating polymorphisms in Capsella, we mapped Neslia Illumina reads onto the Capsella reference genome using Stampy (Lunter and Goodson 2011), which is optimized for mapping divergent sequences. Additionally, we conducted 5 kb–insert mate-pair sequencing of one of the C. grandiflora accessions and the N. paniculata accession and produced Illumina-only de novo assemblies as described in the Supplementary Note 3.1.
3.2.1.7 Transposable element annotation

The TEdenovo pipeline from the REPET v2.0 package (Flure et al. 2011) was used for de novo identification of repeated sequences in the following genomes: A. thaliana (ecotypes Col, Ler, Kro, Bur and C24 from the 1001 Arabidopsis genomes project; see URLs), A. lyrata, A. alpina (courtesy of G. Coupland), Brassica rapa, C. rubella, Eutrema halophila and S. parvula. The selected sequences from all species were combined into the Brassicaceae repeat library, to which we also appended the A. thaliana repeat library from the Repbase database. TEannot from the REPET v2.0 package was run against the C. rubella, A. thaliana and A. lyrata genomes using the Brassicaceae library.

LTR retrotransposons were identified de novo for each genome using LTRharvest from the Genome Tools v1.3.9 package (Ellinghaus et al. 2008) with default parameters. MUSCLE v3.8.31 (Edgar 2004) was used to align LTRs from annotated elements, alignment ends were trimmed at each end to have three consecutive matching nucleotides, and the Kimura two-parameter distance was calculated for each alignment using the EMBOSS v6.4.0 dismat function (Rice et al. 200). Insertion time was then calculated using the methods described in Hu et al. (2011).

3.2.2 FISH and comparative chromosome painting

Cytogenetic analyses were conducted using meiotic chromosomes at the stage of pachytene or diakinesis prepared from young flower buds. See Supplementary Figure 3.8
for details of rDNA and telomere repeat probes and *A. thaliana* BAC contigs used as chromosome-specific probes for comparative chromosome painting in *C. rubella*. All DNA probes were labeled with biotin–deoxyuridine triphosphate (dUTP), digoxigenin-dUTP or Cy3-dUTP by nick translation and hybridized to suitable chromosome spreads. Fluorescence signals were analyzed with an Olympus BX-61 epifluorescence microscope and a CoolCube camera (MetaSystems) and pseudocolored using Adobe Photoshop CS2 software (Adobe Systems). See Supplementary Figure 3.8 and Lysak et al (2013) for a detailed description of all protocols used.

### 3.2.3 URLs

JGI sequencing protocols,


### 3.2.4 Accession codes

The assembly and annotation (Entrez BioProject ID PRJNA13878) are available from GenBank (accession number ANNY00000000) and from the JGI PHYTOZOME portal (see URLs). RNA-seq data sets are available from GenBank (GEO SuperSeries GSE45687 and SRA accession PRJNA194469). Seeds from the reference *C. rubella* strain Monte Gargano are available from the *Arabidopsis* Biological Resource Center.
(ABRC) under accession number CS22697 and the Nottingham *Arabidopsis* Stock Centre (NASC) under accession number N9609.

### 3.3 Acknowledgements

The work conducted by the DoE JGI is supported by the Office of Science of the DoE under contract number DE-AC02-05CH11231. We thank J. Bergelson, J. Borevitz, A. Hall, C. Langley, K. Mayer, J. Nasrallah, B. Neuffer, Y. Van de Peer and O. Savolainen for contributing to the initial sequencing proposal submitted to the Community Sequencing Program at JGI. We also thank T. Bureau, D. Schoen, P. Harrison, J. Stinchcombe, A. Moses and E. Harmsen for their contributions to the Value-directed Evolutionary Genomics Initiative (VEGI) grant (Genome Quebec/Genome Canada), which funded *C. grandiflora* genomic and mRNA sequencing, and G. Coupland (Max Planck Institute for Plant Breeding Research) and colleagues for information on *Arabis alpina* repeats. The work was supported by the Max Planck Society (D.W.), the Genome Quebec and Genome Canada VEGI grant (S.I.W. and M.B.), the Natural Sciences and Engineering Research Council of Canada (NSERC) (S.I.W.), the Swedish Research Council (T.S.), the Carl Trygger and Erik Philip-Sörensen foundations (T.S.), National Science Foundation (NSF) grant 0929262 (J. Steffen and R.M.C.), the French National Research Agency (ANR-08-KBBE-012-02 to H.Q.), the Czech Science Foundation (excellence cluster P501/12/G090 to M.A.L.) and the European Regional Development Fund (CZ.1.05/1.1.00/02.0068 to M.A.L.). D.K. was supported by a Human Frontiers in Science Program Long-Term Fellowship, and G.C. was supported by the Alfred P. Sloan
Population genetics analyses were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project b2012122. L.M.S. was supported by a European Community FP7 Marie Curie Fellowship (PIEF-GA-2008-221553) and an EMBO Long-Term fellowship.

3.4 Contributions

CHAPTER 4 MATING SYSTEM SHIFTS AND TRANSPOSABLE ELEMENT EVOLUTION IN THE PLANT GENUS CAPSELLA

Abstract

Background

Despite having predominately deleterious fitness effects, transposable elements (TEs) are major constituents of eukaryote genomes in general and of plant genomes in particular. Although the proportion of the genome made up of TEs varies at least four-fold across plants, the relative importance of the evolutionary forces shaping variation in TE abundance and distributions across taxa remains unclear. Under several theoretical models, mating system plays an important role in governing the evolutionary dynamics of TEs. Here, we use the recently sequenced Capsella rubella reference genome and short-read whole genome sequencing of multiple individuals to quantify abundance, genome distributions, and population frequencies of TEs in three recently diverged species of differing mating system, two self-compatible species (C. rubella and C. orientalis) and their self-incompatible outcrossing relative, C. grandiflora.

Results

We detect different dynamics of TE evolution in our two self-compatible species; C. rubella shows a small increase in transposon copy number, while C. orientalis shows a substantial decrease relative to C. grandiflora. The direction of this change in copy number is genome wide and consistent across transposon classes. For insertions near
genes, however, we detect the highest abundances in *C. grandiflora*. Finally, we also find differences in the population frequency distributions across the three species.

**Conclusion**

Overall, our results suggest that the evolution of selfing may have different effects on TE evolution on a short and on a long timescale. Moreover, cross-species comparisons of transposon abundance are sensitive to reference genome bias, and efforts to control for this bias are key when making comparisons across species.

### 4.1 Background

In plants, transposable element (TE) abundance ranges from around 20% in the compact *Arabidopsis thaliana* genome (The *Arabidopsis* Genome Initiative 2000) to over 80% in the maize genome (Schnable et al. 2009). Although it has long been clear that TE content varies enormously across taxa, the extent of, and evolutionary reasons for, TE variation among closely related species is less clear (Tenaillon et al. 2010; Wright and Ågren 2011). Where whole genome sequences are available, comparisons have been limited to two species (Hu et al. 2011; Tenaillon et al. 2011; Chia et al. 2012; Chen et al. 2013) and where more species were compared, analyses were typically restricted to one or a few TE families (Young et al. 1994; Charlesworth and Charlesworth 1995; Francis et al. 1995; Wright et al. 2001; Tam et al. 2007; Kawakami et al. 2011; Estep et al. 2013). Until recently, the lack of large-scale genomic data for closely related species has precluded comprehensive tests. This problem is rapidly diminishing with the increase in available
whole genome sequences, allowing theoretical models to be tested across genomes and across species at a scale not previously possible (Tenaillon et al. 2011; Renny-Byfield et al. 2011; Piednoël et al. 2012; Muños-Diez et al. 2012).

According to several models, mating system is expected to play an important role in driving the evolutionary dynamics of TEs (reviewed by Wright et al. 2008; Ågren 2014). There are two main reasons for this. First, the spread of TEs may be inhibited by a lack of outcrossing (Cavalier-Smith 1980; Hickey 1982; Morgan 2001; Boutin et al. 2012). Second, self-regulation of transposition is more likely to evolve in selfers than in outcrossers and host-silencing mechanisms are more likely to spread to fixation with greater linkage to the active element (Charlesworth and Langley 1986). Therefore, all else being equal, outcrossing species are predicted to maintain a higher abundance of TEs than selfing species. Alternatively, the expected reduction in the effective population size \( (N_e) \) in selfers relative to outcrossers (Nordborg 2000; Charlesworth and Wright 2001) and the associated reduction in the efficacy of selection may lead to fewer TEs in the genomes of outcrosser (Charlesworth and Charlesworth 1995; Lynch 2007). Furthermore, if selection is mainly due to chromosomal rearrangements caused by between-element (ectopic) recombination, lower heterozygosity in selfers may lead to relaxed selection and as a consequence TE accumulation (Charlesworth and Charlesworth 1995; Wright and Schoen 1999; Morgan 2001). Empirical evidence to date, while limited, provides some support in favour of TE loss following the evolution of selfing (Wright et al. 2008). In particular, the *Arabidopsis thaliana* genome has consistently fewer insertions in comparison with its outcrossing congener *A. lyrata* (Lockton and Gaut 2010; Hu et al. 2011; de la Chaux et al. 2012), which may partly be driven by more efficient host
silencing via small RNAs (Hollister et al. 2011). Moreover, there is some evidence suggesting smaller genome sizes in selfers compared with related outcrosers (Albach and Greilhuber 2004; Wright et al. 2008; but see Whitney et al. 2010). However, whether the evolution of selfing will generally result in an increase or decrease in TE copy number, and the timescale over which TE abundance evolves, remains unclear.

The plant genus *Capsella* provides a promising system to study interspecific variation in TE abundance and distribution. *Capsella*, which diverged from *Arabidopsis* somewhere between 6 and 20 million years ago (Acarkan et al. 2000; Koch and Kiefer 2005; Beilstein et al. 2010), is a relatively small genus within the mustard family (Brassicaceae). Furthermore, the members of the genus vary in mating system (Hurka et al. 2012; Figure 4.1). The sequencing of the genome of the self-compatible *Capsella rubella* and comparisons with its self-incompatible closest relative *C. grandiflora* suggested that *C. rubella* has experienced a global reduction in the efficacy of natural selection on non-synonymous polymorphisms, but without evidence for major shifts in transposable element abundance during the less than 200,000 years since divergence (Slotte et al. 2013).
Figure 4.1 Phylogenetic relationships within the *Capsella* genus

For a comprehensive review of the evolutionary history of the genus, see (Hurka et al. 2012).

Here, we expand the mating system comparisons to population samples from three characterized *Capsella* species, two self-compatible species and one self-incompatible outcrosser. Self-compatible species have evolved at least twice through the divergence from a self-incompatible outcrossing ancestor similar to *C. grandiflora* (2n = 2x = 16), resulting in the recent *C. rubella* (2n = 2x = 16; divergence time 50–200,000 years Foxe et al. 2009; Guo et al. 2009) and *C. orientalis* (2n = 2x = 16; divergence time unknown, but believed to be older than *C. rubella* (Hurka et al. 2012)). *C. grandiflora* is geographically restricted to a glacial refugium in northern Greece and has a stable effective population size (N_e ~ 600,000), with relatively little population structure.
Gossmann et al. 2010; St Onge et al. 2011). *C. rubella* spans the Mediterranean region, while *C. orientalis* stretches from the far eastern parts of Europe, through the South Urals and western Mongolia to northwestern China (Hurka et al. 2012). The effective selfing rate in *C. rubella* has been estimated to be 0.90–0.97 (St Onge et al. 2011). Although the selfing rate in *C. orientalis* has not been quantified, very low allozyme variability suggests that the species is predominately selfing (Hurka et al. 2012). For an extensive review of the evolutionary history of the genus, see (Hurka et al. 2012). We quantify abundance, population frequencies, and genome wide distributions of TEs across the three species and use the results to examine whether the variation is consistent with the effects of mating system outlined above. We also discuss the residual uncertainty of using the reference genome of one species in a comparative study of TE abundance and distribution, and steps that may be taken to address the issue.

### 4.2 Results

#### 4.2.1 Distribution of TE insertions in *Capsella*

We quantified TE abundance using the paired-end read mapping approach of Kofler et al. (2012). Paired-end Illumina reads from multiple individuals from all three species (8 *C. grandiflora*, 10 *C. orientalis*, and 24 *C. rubella* individuals) were mapped to a repeat-masked *C. rubella* reference genome (Slotte et al. 2013), and a TE database (Slotte et al. 2013) with repeats from seven Brassicaceae species (*A. thaliana* (reference accession Col-0 and accessions Ler, Kro-0, Bur-0, and C24 from the 1001 *Arabidopsis* genomes
project), *Arabidopsis lyrata*, *Arabis alpina*, *Brassica rapa*, *Capsella rubella*, *Eutrema halophila*, *Schrenkiella parvulum*). Individual TE insertions were identified by cases where one read maps to a TE and a second to a unique genomic location (Kofler et al. 2012). The TE database comprised 4,261 different TE sequences. While the Kofler et al. (2012) method was originally implemented for pooled population samples, we sequenced individual samples, and used estimates of insertion frequencies to call insertions as heterozygous or homozygous. We performed several tests to assess the suitability of this approach, all of which confirm our general conclusions (see Methods for details).

We identified 21,716 unique insertions across the three species. Of all insertions considered, the majority (approximately 80%) are unique to one species (Figure 4.2). There is a strong consistency in the distribution of copies among TE families. LTR elements are the most common type, making up roughly 59% of all TEs; DNA elements comprise 19%; *Helitron* and non-LTR elements approximately 11% each (Figure 4.3a). The closely related *A. thaliana* and *A. lyrata* also show no difference in the relative abundance of families (de la Chaux et al. 2012). In both genera, non-LTRs are the smallest contributors to the TE load in the genomes. However, in *Arabidopsis* the DNA elements dominate (including Helitrons) making up over 55% of all TEs, consistent with the reported expansion of the Basho Helitrons in *A. thaliana* (Hollister et al. 2007).
Figure 4.2 Venn diagram with the number of unique and shared TE insertion sites in three *Capsella* species

![Venn diagram with numbers](image)

Figure 4.3 Average TE copy number in the three *Capsella* species

(a) genome wide, (b) on chromosome arms (b), and (c) in centromeric regions. Error bars are ±1 standard error.
Since using the *C. rubella* assembly as the reference genome may bias our analysis we used several approaches to assess whether our results were robust to the effect of reference genome bias. To begin assessing this issue we compared the proportion of reads from all species that mapped to the non-pericentromeric regions of the main chromosome scaffolds of the *C. rubella* reference genome. Reassuringly, these proportions do not differ dramatically between the three species (*C. rubella*, 30%; *C. grandiflora*, 27%; *C. orientalis*, 25%). We discuss additional approaches to control for reference genome bias in more detail in connection with the relevant results below.

### 4.2.2 Abundance of TEs

The estimated mean number of TE insertions varied between the three species (Kruskal-Wallis chi-squared = 62.0227, df = 2, *P* < 0.00001; Figure 4.3). Estimated TE copy number is lowest in self-compatible *C. orientalis*, highest in the outcrosser *C. grandiflora* and self-compatible *C. rubella*. This pattern holds true for all four classes of TE (LTRs, non-LTRs, DNA, and Helitrons). The mean copy number is slightly higher in *C. rubella* compared to *C. grandiflora* (Wilcoxon signed rank test with continuity correction, \(V = 528450463, P = 0.004385\)). This difference is due to a higher number of DNA elements (\(V = 17017897, P = 0.004798\)) and Helitrons (\(V = 9236731, P < 0.00001\)).

Moreover, within-species variation in TE abundance is highest in *C. rubella* (Supplementary Figure 4.1).

To better characterize between-species differences, we separately examined copy numbers in gene-rich chromosome arms, where we expect selection against TE insertions to be strong, and those in centromeric regions, where selection should be weak (Wright et
al. 2003). Indeed, all families show significantly higher densities in pericentromeric regions (in *C. orientalis* only LTR and *Helitrons*) (Mann–Whitney test, *P* < 0.05). When comparing species, significant lower TE numbers are still apparent on the chromosome arms for *C. orientalis*, but we see no significant difference between *C. rubella* and *C. grandiflora* (Kruskal-Wallis rank sum test, chi-squared = 15.4656 df = 2, *P* < 0.0001).

A central concern with our results is the bias that might arise from using *C. rubella* as the reference genome in the analysis. Reference genome bias can come in two forms: 1) a greater ability to find insertion sites due to higher mapping of flanking regions, and 2) a greater representation of TEs from the reference species. Although all of our species are closely related, and species differences in the percentage of reads mapping to the genic regions are small, our general patterns of TE abundance follow the phylogenetic pattern expected if reference genome bias is playing a role with species closest to *C. rubella* showing the greatest TE abundance. To further address the first concern we first took advantage of the previously generated Illumina-based de novo assemblies of *C. grandiflora* (Slotte et al. 2013) and the close outgroup *Neslia paniculata* (Slotte et al. 2013), and also performed a *de novo* assembly of *C. orientalis* (see Methods for details). To confirm the lower numbers of TEs in *C. orientalis*, we redid the analysis for the *C. orientalis* and *C. grandiflora de novo* assemblies with a TE database including only insertions identified in the *C. orientalis* assembly. The assemblies of the two species should be of similar quality and using a *C. orientalis* biased TE database should reverse the bias that might arise from using our larger TE database, as *C. grandiflora* may be expected to share more insertions with *C. rubella*. Taking this approach, we still detect significantly lower TE numbers in *C. orientalis* (Supplementary Figure 4.2a, generalized
linear model with Poisson distribution, \( z = -20.6, P < 0.00001 \). A caveat is that the de novo assembly of \textit{C. orientalis} is very TE-poor to begin with.

Second, to address the two concerns while removing any reference bias, we mapped all species against the de novo assembly of \textit{N. paniculata}, using a TE database based only on insertions identified in \textit{A. thaliana} and \textit{A. lyrata}. Although the number of insertions identified was dramatically reduced, we saw the same pattern of differential TE abundance, in particular fewer TEs in \textit{C. orientalis} (Supplementary Figure 4.2b, generalized linear model with Poisson distribution, \( z = -9.394, P < 0.00001 \)). Thus, while reference genome bias likely plays some role in the estimated magnitude of the between-species differences in TE abundance, on balance the data generally support the inferred low TE numbers in \textit{C. orientalis}.

4.2.3 TE insertions near genes

Under several population genetic models, TEs are expected to be rapidly removed from gene-rich regions (Dolgin and Charlesworth 2008). Comparing TE copy number in regions near genes may therefore provide insights about the number of recent insertions in a given genome. Using the gene annotation from \textit{C. rubella} (Slotte et al. 2013), we calculated the distance to the closest gene from the beginning (or end, whichever was closest) of each insertion. We find a trend to more insertions within 1,000 bp of genes in \textit{C. grandiflora} compared to the other species (Figure 4.4), with \textit{C. orientalis} again having the lowest TE density near genes. \textit{Capsella orientalis} is significantly different from the other two species (Kruskal-Wallis chi-squared = 1021.348, df = 2, \( P < 0.00001 \)). \textit{Capsella}
grandiflora shows significantly higher abundance than C. rubella in regions within 200 bp of the nearest gene (generalized linear model with Poisson distribution, $z = -4.789$, $P < 0.00001$). Focusing the analysis on regions close to genes, where problems with read mapping should be minimized, also allows us to address the second way in which reference genome bias may occur: a greater ability to find insertion sites due to higher mapping of flanking regions. Again, the lower number of TEs near genes in C. orientalis is consistent with our general conclusions.

![Figure 4.4](image)

**Figure 4.4 Average TE copy number in 100 bp bins near their closest gene in the three Capsella species**

Error bars are ±1 standard error.

### 4.2.4 Frequency distributions of TEs

We used the presence or absence of all identified insertions across individuals to calculate population frequency distributions for all three species. The TE distribution of
*C. orientalis* is distinct from that in *C. rubella* and *C. grandiflora* (Kruskal-Wallis chi-squared = 399.93, df = 2, \( P < 0.00001 \); Figure 4.5). This is true also when ignoring fixed insertions (Kruskal-Wallis chi-squared = 481.33, df = 2, \( P < 0.00001 \)). *C. orientalis* has the highest number of fixed insertions and the lowest proportion of rare insertions, consistent with a low or no contemporary accumulation of TEs and a general genome wide reduction in diversity across the genome. Consistent with the hypothesis of relaxed selection near centromeres, both *C. rubella* and *C. grandiflora*, but not *C. orientalis* differ in their frequency distributions between chromosome arms and centromeric regions, with a significant excess of common insertions in the centromeric regions (Wilcoxon Rank sum test, both \( P < 0.05 \)).

![Histogram of population frequencies of TEs in the three *Capsella* species](image)

**Figure 4.5 Histogram of population frequencies of TEs in the three *Capsella* species**

(a) genome wide, (b) on chromosome arms (b), and (c) in centromeric regions. 95% confidence intervals based on 200 bootstraps are plotted but too small to be seen.
4.3 Discussion

Here, we report results from a whole genome study of TE abundance and distributions in multiple individuals in three species from the plant genus *Capsella*. Comparing population samples from the outcrosser *C. grandiflora* to two of its self-compatible relatives allows us to begin to empirically dissect the population and genome wide effects of a mating system shift in driving TE evolution.

The evolution of selfing does not appear to have had the same effect in the two self-compatible *Capsella* species. Perhaps the most striking result of this study is the consistently lower TE copy numbers in the self-compatible *C. orientalis*. This reduction is apparent for all families and in both centromeric regions and along chromosome arms (Figure 4.3), as well as for recent insertions near genes (Figure 4.4). Furthermore, *C. orientalis* shows a detectable absence of rare and an excess of common TE insertions (Figure 4.5). The present transposition rate thus appears to be very low in *C. orientalis*. TE accumulation is known to be a key driver of genome size evolution in plants (Tenaillon et al. 2010; Ågren and Wright 2011) and this reduction in transposition rate may in part explain why *C. orientalis* has the smallest genome in the genus (Hurka et al. 2012).

In contrast to *C. orientalis*, *C. rubella* does not appear to be TE poorer than *C. grandiflora*. Instead, there is a trend of an increase in copy number, which seems to be due to higher accumulation in centromeric regions (Figure 4.3), although this observation may also be due to poorer mapping of the other species in these regions. However, when we consider only insertions near genes, which are where recent insertions tend to reside,
C. grandiflora has a higher abundance than C. rubella (Figure 4.4). Capsella rubella also has the highest excess of rare insertions, although this trend is most pronounced along the chromosome arms (Figure 4.5). This may reflect an increase in transposition rate or be a product of the recent population bottleneck C. rubella experienced in conjunction with the evolution of selfing (Foxe et al. 2009; Brandvain et al. 2013).

What would determine whether selfing leads to a net accumulation or loss of TEs? One important factor is likely to be the age of the selfing lineage (Wright et al. 2008). As outlined in the Introduction, selfing will reduce the effective population size and this reduction following the shift to selfing may initially result in an increase in fixation rates compared to the outcrossing relative. However, over time, the lack of outcrossing means that any new (deleterious) insertion that arises in either lineage will have a harder time spreading in the selfing lineage. As a consequence, we may observe different effects on selfing in a young and an old lineage. Here, we detected fewer TEs in C. orientalis, but a slight increase in C. rubella. Capsella orientalis diverged from C. grandiflora before C. rubella did, suggesting that it may have been self-fertilizing for longer. It is important to note, however, that the speciation event and the evolution of selfing may not have occurred simultaneously. This, for example, is the case in A. thaliana (Bechsgaard et al. 2006; Tang et al. 2007; Tsuchimatsu et al. 2010; Shimizu et al. 2011), where the evolution of selfing apparently occurred a long time after the speciation event. Although the shift to self-fertilization can occur both within a lineage and in conjunction with a speciation event, recent work by Goldberg and Igić (2012) indicate that the shift is ten times more likely to be associated with a speciation event than to occur within a lineage. While it is not clear whether the evolution of selfing in C.
orientalis coincided with the speciation event as it did in *C. rubella* (Foxe et al. 2009), the very recent origin of *C. rubella* and the very low species wide allozyme variability in *C. orientalis* (Hurka et al. 2012) suggest that *C. orientalis* may have been selfing longer than *C. rubella*. Proper dating of the origin selfing in *C. orientalis* should be the focus of future work.

We undertook several approaches to control for reference genome bias in copy number estimation. There is a clear effect of such bias in that the relative copy-number difference estimated depends strongly on which reference genome is being used for mapping. On the one hand, the *C. rubella* genome is by far the highest-quality reference genome, and in most cases we detect the highest copy numbers using this genome as the reference (compare Figure 4.3 with Supplementary Figure 4.2). However, taking this approach may also maximize the bias, causing an exaggerated assessment of copy number differences between species. Nevertheless, the patterns observed, particularly with our bias-free mapping to the *N. paniculata* genome, do suggest that our general conclusions may be robust to assembly and mapping differences. Ultimately, long-read data integrated with higher-quality assemblies of all *Capsella* species will be important for validating the results reported here.

4.4 Conclusions

Taken together our results suggest that the effects of mating system on transposon evolution may vary from case to case. A candidate factor determining the direction of the
effect may the age of the selfing lineage. Finally, cross-species comparisons of
transposon abundance are sensitive to reference genome bias and caution must be applied
when using re-sequencing approaches.

4.5 Methods

4.5.1 Sampling and sequencing

Samples from all species come from a large range of their species distributions
(Supplementary Table 4.1). *C. grandiflora* samples come from 12 populations, with one
individual sampled per population, spanning the native range in Greece. The thirteenth
sample was cross between two other populations. For *C. orientalis* we obtained samples
from five previously described populations (Hurka et al. 2012), with two individuals per
population sampled. After growth in the University of Toronto greenhouse for several
months, DNA from leaf tissue from all samples was extracted using a modified CTAB
protocol (Doyle et al. 1987). Sequencing was done at Genome Quebec Innovation Centre
using the Illumina Genome Analyzer platform 121 (Illumina, San Diego, California,
USA). *Capsella rubella* samples came from across its geographical range and were
grown and sequenced at the Max Planck Institute for Developmental Biology, Germany.
The median average coverage was 20x for *C. orientalis*, 39x for *C. grandiflora*, and 22x
for *C. rubella*. Sequences are available on the Sequence Read Archive
SRP041585), *C. grandiflora* (Accession number SRP044121), and *C. rubella* (Accession number PRJEB6689).

4.5.2 Genome assemblies

For *C. rubella*, we used the recently completed reference genome (Slotte et al. 2013). We also took advantage of the Illumina-based *de novo* assemblies of *C. grandiflora* and *N. paniculata* prepared for that analysis (for details see Slotte et al. 2013). The *C. orientalis* assembly was prepared from 17.6 Gb of 108 bp Illumina paired-end reads in ten libraries. Reads were assembled into contigs using the Ray (v 2.1.0) assembler (Boisvert 2010) with a Kmer of 31 under 20 multiple cores (N50 ~ 25 kb). Contigs shorter than 500 bp were discarded after scaffolding. For further assembly details see Supplementary Table 4.2. *De novo* assemblies are available on CoGe (https://genomevolution.org/CoGe/): *C. orientalis* (Genome ID 24033), *C. grandiflora* (Genome ID 24068), and *N. paniculata* (Genome ID 24067).

4.5.3 Identification of unique TE insertions

To detect TE insertions across the re-sequenced genomes we used PoPoolationTE (Kofler et al. 2012). The method requires three things: (1) annotated reference genome, (2) a library of TE sequences, and (3) paired-end sequence data. The strength of this approach is that it allows the identification of insertions not present in the reference genome. Here, we used the recently completed *C. rubella* genome (Slotte et al. 2013), as well as the
Brassicaceae TE database generated as part of the genome annotation (for details of TE annotations in the *C. rubella* genome see Slotte et al. 2013). We ran the pipeline using default settings on 108 bp paired-end Illumina samples from 8 *C. grandiflora*, 10 *C. orientalis*, and 24 *C. rubella* individuals.

PoPoolation typically requires DNA from pooled samples from multiple individuals; it then uses the read mapping results to estimate population wide frequencies. Here, instead of using pooled samples, we applied the pipeline to DNA samples from single individuals. We used the frequency output to infer whether a given insertion was present in the genome considered. Insertions were considered identical if their estimated location was within 200 bp of each other (Kofler et al. 2012). Any insertion with an estimated frequency higher than 0.8 was treated as a homozygous; insertions with an estimated frequency of < 0.2 were considered errors and insertions with an intermediate frequency were called as heterozygous. To test PoPoolationTE’s ability to correctly distinguish heterozygote and homozygous insertions we ran the pipeline on two *C. rubella* accessions (*cr1gr1* and JGI) that have been selfed for multiple generations in the greenhouse, as well as on a hybrid sample created by merging the sequences of both samples. If the programme can correctly infer homozygous and heterozygous insertions, we expect almost all insertions in the pure samples to be fixed and so have an inferred frequency of 1 and the hybrid to show an increase of calls around 0.5. Indeed, this is what we observe (Supplementary Figure 4.3). In the two highly selfed samples 88% and 82% of all insertions had an inferred frequency of 1, which was reduced in the hybrid to 61%. The shape of the count distribution also provides justification for using 0.8 as a cut-off. In the two highly selfed samples, the counts remain very low until around 0.8 where there is
an increase (although this is less clear in the JGI sample). Moreover, to further assess whether our conclusions were robust to our homo- heterozygous individual-based calling approach we redid the copy number comparison across species using the raw frequency estimates from all individuals and we find that patterns do not change (Kruskal-Wallis chi-squared = 29.6582, df = 2, P < 0.00001). Finally, to test how our individual-based approach compared with the pooled approach, we constructed pooled sequence samples by merging the sequences from all individuals of a species into one pooled sample. Again, we find that are conclusions about TE abundance in the three species do not change (Kruskal-Wallis chi-squared = 24.0303, df = 2, P < 0.00001). Venn diagrams of unique and shared insertions were generated using VENNY (Oliveiros 2014). Using the gene annotation from the C. rubella reference genome (Slotte et al. 2013), we calculated the distance to the closest gene for all insertions. We then calculated the number of insertions in 100 bp bins.

4.5.4 Frequency distributions

We used the presence or absence of all identified insertions to calculate frequency distributions for all three species. For the self-compatible species C. rubella and C. orientalis all insertions were assumed to be homozygous. As mentioned in the Introduction, the effective selfing rate in C. rubella is around 0.90–0.97 (St Onge 2011) and whereas selfing rate in C. orientalis has not been quantified but very low allozyme variability suggests that the species is predominately selfing (Hurka et al. 2012). In this case, the frequency for each insertion is the number of sampled individuals in which the
insertion was detected, divided by the total number individuals. For outcrossing C. grandiflora, we treated each haplotype as an independent sample. Frequency calculations were restricted to 8 randomly chosen individuals for each selfing species, and half of that (4) for C. grandiflora.

4.5.5 Availability of supporting data

The datasets supporting the results of this article are available for download. De novo assemblies are available on CoGe (https://genomevolution.org/CoGe/) for C. orientalis (Genome ID 24033), C. grandiflora (Genome ID 24068), and N. paniculata (Genome ID 24067). Paired-end sequences are available on the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) for C. orientalis (Accession number SRP041585), C. grandiflora (Accession number SRP044121), and C. rubella (Accession number PRJEB6689).

4.6 Competing interests

The authors declare that they have no competing interests.
4.7 Authors’ contributions

JAÅ and SIW conceived the study. BN, DK, and DW contributed seeds and/or sequencing data. JAÅ, WW, and SIW analysed the data. JAÅ and SIW wrote the manuscript, with help from all authors. All authors read and approved the final manuscript.

4.8 Acknowledgements

We thank A.D. Cutter, S.C.H Barrett, J.D. Hollister, and A.E. Platts for comments and discussion. Comments from two anonymous reviewers greatly improved the manuscript. J.A.Å is supported by a Junior Fellowship from Massey College, S.I.W. by a Natural Sciences and Engineering Research Council (NSERC) Accelerator Award, D.K by a Human Frontier Science Program Long-term Fellowship, and D.W by a Gottfried Wilhelm Leibniz Award from the German Research Foundation and the Max Planck Society.
4.9 Supplementary Information

Supplementary Table 4.1 Origins of sequenced *Capsella* samples

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession</th>
<th>Latitude (ºN)</th>
<th>Longitude (ºE)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. grandiflora</em></td>
<td>83.17</td>
<td>38.4379</td>
<td>21.42431667</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>85.33</td>
<td>39.55521667</td>
<td>20.91641667</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>AxE</td>
<td>n/a</td>
<td>n/a</td>
<td>Cross between individuals918/8 and Cg2e, two Corfu populations</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>103.17</td>
<td>39.51838333</td>
<td>21.56091667</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>5a</td>
<td>39.705025</td>
<td>19.7573444</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>91.2</td>
<td>39.86715</td>
<td>20.70708333</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>95.15</td>
<td>39.14541667</td>
<td>20.05816667</td>
<td></td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>86.8</td>
<td>39.01723333</td>
<td>20.13193333</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>690</td>
<td>36.15</td>
<td>-5.58</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>697</td>
<td>44.1336</td>
<td>10.2144</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>698</td>
<td>44.1336</td>
<td>10.2144</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>762</td>
<td>37.966667</td>
<td>23.716667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>844</td>
<td>35.202554</td>
<td>24.233091</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>879</td>
<td>35.29</td>
<td>24.42</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>907</td>
<td>39.666667</td>
<td>19.8</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>925</td>
<td>39.666667</td>
<td>20.85</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>984</td>
<td>39.5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1207</td>
<td>28.316667</td>
<td>-16.566667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1208</td>
<td>28.316667</td>
<td>-16.566667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1209</td>
<td>28.316667</td>
<td>-16.566667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1215</td>
<td>28.316667</td>
<td>-16.566667</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>ID</td>
<td>Value 1</td>
<td>Value 2</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1311</td>
<td>42.88333333</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1377</td>
<td>-34.6666667</td>
<td>-58.5</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1453</td>
<td>43.4666667</td>
<td>11.033333</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>774</td>
<td>41.8333333</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>86IT1</td>
<td>40.62</td>
<td>14.37</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1GR1</td>
<td>37.742485</td>
<td>26.81958</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1574-1</td>
<td>41.6</td>
<td>8.983333</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1407-8</td>
<td>35.183333</td>
<td>24.233333</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1504-11</td>
<td>28.666667</td>
<td>-17.866667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1575-1</td>
<td>41.383333</td>
<td>9.166667</td>
<td></td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>1249-11</td>
<td>38.783333</td>
<td>-9.383333</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1719-3</td>
<td>48.35</td>
<td>88.26</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1719-4</td>
<td>48.35</td>
<td>88.26</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1979-1</td>
<td>51.07</td>
<td>81.48</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1979-7</td>
<td>51.07</td>
<td>81.48</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1981-10</td>
<td>51.08</td>
<td>81.36</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1981-6</td>
<td>51.08</td>
<td>81.36</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1985-1</td>
<td>53.21</td>
<td>83.44</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>1985-5</td>
<td>53.21</td>
<td>83.44</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>2008-1</td>
<td>46.37</td>
<td>90.52</td>
<td></td>
</tr>
<tr>
<td><em>C. orientalis</em></td>
<td>2008-2</td>
<td>46.37</td>
<td>90.52</td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 4.2 Assembly statistics for *C. orientalis de novo* assembly

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of contigs</td>
<td>23,659</td>
</tr>
<tr>
<td>Total number of base pairs (bp)</td>
<td>143,780,143</td>
</tr>
<tr>
<td>Shortest contig (bp)</td>
<td>500</td>
</tr>
<tr>
<td>Longest contig (bp)</td>
<td>647,746</td>
</tr>
<tr>
<td>Average GC%</td>
<td>35.0</td>
</tr>
<tr>
<td>Non-ACGT bases (bp)</td>
<td>2,655,457</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>25,186</td>
</tr>
</tbody>
</table>

![Graph](image)

**Supplementary Figure 4.1** Total TE copy number in sampled individuals in the three *Capsella* species.
Supplementary Figure 4.2 Average TE copy number in the three *Capsella* species

Each species was mapped to its own genome (a) and to the *Neslia paniculata* assembly (b) using a TE database based on *Arabidopsis thaliana* and *Arabidopsis lyrata*. The difference between *C. rubella* and the other species in (a) is exaggerated by the higher quality of the *C. rubella* reference genome compared with the Illumina-only *de novo* assemblies of the other species. Error bars are ±1 standard error.
Supplementary Figure 4.3 Inferred TE frequencies in highly selfed accessions and synthetic hybrid

Counts of inferred TE frequency for two highly selfed accessions (cr1gr1 and JGI) of *C. rubella*, as well as on a hybrid sample created by merging the sequences of both samples. The y-axis is cut at 300 to highlight the increase in the number of insertions of intermediate frequencies inferred in the hybrid sample.
Abstract

Most models of genome size evolution emphasize changes in relative rates of and/or the efficacy of selection on insertions and deletions. However, transposable elements (TEs) are a major contributor to genome size evolution, and since they experience their own selective pressures for expansion, genome size changes may in part be driven by the dynamics of co-evolution between TEs and their hosts. Under this perspective, predictions about the conditions that allow for genome expansion may be altered. In this review, we outline the evidence for TE–host co-evolution, discuss the conditions under which these dynamics can change, and explore the possible contribution to the evolution of genome size. Aided partly by advances in our understanding of the mechanisms of TE silencing via small RNAs, there is growing evidence that the evolution of transposition rates can be important in driving genome expansion and contraction. Shifts in genome size and transposon abundance associated with interspecific hybridization and changes in mating system are consistent with an important role for transposition rate evolution, although other possible explanations persist. More understanding of the potential for the breakdown of host silencing mechanisms and/or the potential for TEs to evade host immune responses will improve our understanding of the importance of changes in TE activity in driving genome size evolution.
5.1 Introduction

Despite an extensive body of research, there is still little consensus on the predominant factors driving the evolution of genome size in natural populations (Petrov 2001; Lynch and Conery 2003; Gregory 2005; Whitney and Garland 2010). When considering the population genetic processes that could drive genome size evolution, several non-mutually exclusive possibilities arise (Petrov 2001). To start, mutational explanations predict that differences among species in the underlying rates of insertion relative to deletion drive the process of genome expansion and contraction (Petrov et al. 2000; Gregory 2004; Pettersson et al. 2009). Alternatively, changes in the strength or efficacy of natural selection, driven by differences in effective population size, could allow for genome expansion (Lynch and Conery 2003; Lynch 2007, 2011; but see Whitney et al. 2010 and Whitney and Garland 2010; Whitney et al. 2011). Finally, positive natural selection could fine-tune genome sizes for their species-specific optima (Bennet and Leitch 2005) or species may differ in their constraints on genome size depending on their life history (Pagel and Johnstone 1992; Charlesworth and Barton 2004).

Variation in genome size among closely related species has often been shown to be in large part due to differential accumulation of transposable elements (TEs; Vitte and Panaud 2005; Hawkins et al. 2006; Piegu et al. 2006). For example, there is a strong correlation between the fraction of the genome comprised of TEs and genome size in angiosperms (Tenaillon et al. 2010). Plant species with relatively small genomes, such as Arabidopsis thaliana, tend to have TE contents of around 20–30% (The Arabidopsis Genome Initiative 2000), whereas relatively large genomes such as maize (Schnable et al.
2009) and barley (Wicker et al. 2005) have much higher proportions (>85%). Given the predominant role of TE abundance in the evolution of genome size, models of genome size should consider the specific dynamics that can arise through transposon–host co-evolution. The incorporation of TEs in models of genome size evolution is especially important because the ability of TEs to self-replicate and evolve as genomic parasites make them distinct from being simply a particular class of insertion and factors determining their abundance can be unique (Charlesworth and Langley 1989).

Considering the contribution of TE abundance to genome size evolution, there are two main classes of explanation for genome size dynamics. First, TEs may play a passive role. Under this model, average transposition rates are constant, and species differences in TE abundance result from differences in deletion rates or in the efficacy of selection. Many of the most cited models for genome size change focus on this class of explanation, including Lynch’s mutation–drift models (Lynch and Conery 2003) and Petrov’s deletion bias explanation (Petrov et al. 2000; Petrov 2001). However, given the role of TEs as genomic parasites, another possibility is that species differences in TE activity may result in changes in TE accumulation. In this case, changes in host silencing mechanisms, and/or TE virulence (evasion of silencing and the degree of self-regulation), could lead to differential TE proliferation (Figure 5.1). Indeed, several recent studies have reported patterns consistent with the second explanation, which we discuss further below. As our understanding of small RNA-mediated TE silencing and TE transposition mechanisms improves, we are gaining increasing evidence for the contribution of this latter type of mechanism in genome size evolution. Furthermore, there is growing evidence that processes known to affect genomic architecture, such as hybridisation events and mating
system shifts, also affect the outcome of the evolutionary arms race between hosts and TEs.

Figure 5.1 Factors influencing genome size via the evolution of transposition rates

Hybridization can introduce active TEs into a host lacking effective silencing mechanisms and thereby reduce the efficacy of silencing leading to increased activity. High levels of outcrossing and sexual reproduction increase rates of recombination, and this reduces the chance of fixing host silencing alleles as well as reducing the selective advantage of host regulation. Larger effective population sizes increase the efficacy of selection on transposition rates, but this can enhance the efficacy of selection on both host silencing and on TE virulence, so the net effect on transposition rate is not clear. Note that increases in transposition rate alone are expected to increase TE abundance and thus increase genome size, but other factors not shown here (strength of selection against new insertions, rate, and strength of selection on DNA removal) can reduce or eliminate this association.

Here, we review the evidence to date that changes in TE activity contribute significantly to TE abundance and genome size evolution, and suggest opportunities for
further tests of this possibility. While our primary focus is on the plant literature, we also briefly discuss new insights into host silencing and TE co-evolution in animals, particularly *Drosophila*.

5.2 Re-activation of TE activity

Under models of host–parasite co-evolution, there is often an ongoing dynamic of parasite evasion of host silencing and the evolution of host resistance. Does such a dynamic occur in TE–host interactions? Some of the best evidence for such co-evolution comes from analyses of TE activity and accumulation in hybrids. The situation could in many ways be seen as analogous to conventional host–parasite relationships, where an arms race makes the host well prepared to deal with a specific parasite, but is very sensitive to new parasites due to a lack of evolved resistance (King et al. 2009). A number of studies have demonstrated increased TE activity and accumulation following hybridization. For example, experimental crosses between *A. thaliana* and *Arabidopsis arenosa* lead to elevated expression of *A. arenosa ATHILA* elements in the hybrid (Josefsson et al. 2006). Crosses between two species of Australian wallaby (*Macropus eugenii* and *Wallabia bicolor*) led to a large increase in the size of centromeres due to proliferation of retroelements (Waugh O'Neill et al. 1998). Although transcript levels may not necessarily correlate with elevated rates of transposition, for example if post-transcriptional silencing is in place (Nuzhdin et al. 1998; Vu and Nuzhdin 2011), there are examples where hybridization can also be linked to increased accumulation. Examples include the Sunfish transposon in *A. thaliana × A. arenosa* synthetic
allotetraploid (Madlung et al. 2005) and several DNA transposons in rice (*Oryza sativa*) in response to introgression from the wild relative *Zizania latifolia* (Shan et al. 2005; Wang et al. 2010).

While many of these results are based on experimental crosses, there are also examples from the wild. The sunflower genus *Helianthus* provides a particularly good example of hybridization leading to TE proliferation and genome expansion in natural populations. The three hybrid taxa *Helianthus anolmalus*, *Helianthus deserticola*, and *Helianthus paradoxxus* are all products of a hybridisation event between *Helianthus annuus* and *Helianthus petiolaris* that occurred around 60–200,000 years ago (Rieseberg et al. 1991; Rieseberg 1997). The three species show several novel karyotypic rearrangements compared to the parent species (Lai et al. 2005a). In addition, each daughter species has a genome about 50% larger than either parental lineage (Baack et al. 2005). This size difference is independent of ploidy; all species, hybrids and parental, are diploid and have the same chromosome number (n = 17). The proliferation of Ty3/gypsy retrotransposons has been shown to be responsible for around 73%, 79% and 62% of this difference genome size in *H. anolmalus*, *H. deserticola*, and *H. paradoxxus*, respectively (Ungerer et al. 2006; Staton et al. 2009). Another retrotransposon, Ty1/copia, also proliferated upon hybridisation, but to a lesser degree (Kawakami et al. 2010). While increased TE copy numbers following hybridization are consistent with derepression leading to increased TE activity, it is indirect evidence and alternative explanations persist. For example, a recent origin of the hybrid species may be associated with a reduction in effective population size (*N_e*). A reduction in *N_e* may reduce the efficacy of
selection against TE accumulation and/or increase homozygosity resulting in relaxed selection against ectopic recombination events (Charlesworth and Langley 1986).

5.3 The genomic immune system

The results from hybridization studies imply that host-silencing mechanisms are under continual evolution to recognize and control TEs. This genomic immune system is mainly administrated by a variety of small RNA mechanisms (Slotkin and Martienssen 2007; Malone and Hannon 2009). Small RNAs have been shown to play an important role in host–parasite interactions, including genomic parasites such as TEs (Obbard et al. 2009). The genomic immune system includes a broad class of processes, unified by the use of short (23–30 nucleotides) RNAs to recognize and manipulate complementary nucleic acids. Gene expression, epigenetic modification, and regulation of heterochromatin are all examples of pathways associated with small RNAs. Small interfering RNAs (siRNAs) in plants and animals and the piwi RNA (piRNA) pathway in animals are both examples of small RNAs involved in regulating TE activity. Below, we discuss the possible contributions of small RNAs to TE–host co-evolution.

piRNA is a maternally inherited TE silencing pathway, which has been shown to silence a variety of common Drosophila TE families. Examples include the TAHRE retroelement (Shpiz et al. 2007) and P transposons (Simmons et al. 2007). piRNAs work by interacting with silencing proteins such as Piwi, Aubergine, and Argonaut to silence transcription or initiate RNA decay (Hutvágner and Zamore 2002; Grivna et al. 2006).
The genomic loci from which the majority of piRNAs are derived are known as piRNA clusters (Brennecke et al. 2007). These clusters can stretch from a several to hundreds of kilobases in length and mostly lack protein coding genes, but contain multiple defective TEs. The location of essentially all Drosophila piRNA clusters in pericentromeric or telomeric heterochromatin suggests that chromatin may play a key role in the activity of piRNA (Aravin et al. 2007). One example of a piRNA cluster is the flamenco locus, which is a major piRNA cluster spanning at least 180 kb involved in repressing the retrotransposons gypsy, ZAM and Idefix (Prud'homme et al. 1995; Desset et al. 2003). The cluster contains several TEs and encodes a silencing program that functions through small piRNAs exerting their effect on different locations throughout the genome.

Mutations in flamenco inhibit the production of the majority of piRNAs derived from the cluster and the gypsy retrotransposons have been demonstrated to become re-activated as a consequence (Brennecke et al. 2007).

In plants, TE activity is suppressed by siRNAs through both post-transcriptional mechanisms and chromatin modifications (Lisch 2009). siRNAs rely on sequence identity (Almeida and Allshire 2005) and combine with proteins, such as Argonaut, and RNA transcripts to methylate the targeted TE sequence (Zhang 2008; Matzke et al. 2009). Mutants of A. thaliana and maize have played an important role in our understanding of the siRNA pathway. For example, in A. thaliana mutations in genes involved in methylation, such as MET1 and DDM1 have been reported to reduce methylation levels and increase TE activity (Zilberman and Henikoff 2007; Lister et al. 2008). In maize the Mu killer (Muk) is a silencer of the MuDR element (Slotkin et al. 2003). MuDR produces two transcripts: mudrA (the transposase) and mudrB (crucial for
the integration of the element). When *Muk* is present, both *mudrA* and *mudrB* transcripts are absent (Slotkin et al. 2005). The silencing effect is heritable and results in significantly lower transcription of *MuDR* elements.

With the large number of successful silencing mechanisms of TEs (see Aravin et al. 2007; Slotkin and Martienssen 2007 and Lisch 2009 for reviews), why are TEs still active? To start, some TEs play important roles in regulating gene expression and will be under positive selection to maintain their activity (see, e.g., Muotri et al. 2007). Second, TEs may not be silenced because they have just arrived in the genome. This could occur either by a mutational divergence from ancestral elements or by horizontal gene transfer from another species (Gilbert et al. 2010). Thirdly, silencing TEs could reflect a trade-off. In *A. thaliana*, for example, the density of methylated TEs is inversely correlated with gene expression (Hollister and Gaut 2009). This suggests that silenced TEs will affect neighboring genes by reducing their activity. For the host, silencing can therefore be seen as an evolutionary trade-off; methylation results in reduced TE activity (benefit) and also disruption of expression of neighboring genes (cost).

Recent studies support the hypothesis that activation of transposable elements in interspecific hybrids may be driven in part by changes in RNA-mediated silencing. In hybrid crosses between *A. arenosa* and *A. thaliana*, a uniparental pattern of TE activation was identified. *A. arenosa*-derived *ATHILA* retrotransposons were activated specifically when *A. arenosa* was the pollen (i.e., paternal) parent (Josefsson et al. 2006). This is consistent with recent evidence that siRNAs in the embryo are produced by the maternally derived tissue. Thus, if the maternal genome lacks the corresponding siRNA loci or fails to produce the corresponding siRNAs in sufficient quantities, the TEs from
the paternal parent will be overexpressed, potentially allowing for a burst of TE activity (Slotkin et al. 2009). Similar effects have been observed in *Drosophila virilis* crosses between the divergent strains 160 and 9. In the cross, at least five unrelated TE families increased in abundance (Lozovskaya et al. 1990; Petrov et al. 1995). The element *Penelope*, which is abundant in strain 160 but, with the exception of ancient and degraded copies practically absent from strain 9, seems to be the major contributor to the mobilization (Lyozin et al. 2001). The locus (or loci) involved in silencing *Penelope* has been shown to be X-linked and siRNAs necessary for silencing are therefore maternally inherited (Blumenstiel and Hartl 2005). Again, this is consistent with a scenario where lack of the appropriate small RNAs results in increased TE expression.

While interspecific hybridization provides one clear mechanism for the breakdown of host silencing, it remains to be seen whether species generally differ in their host-silencing abilities and if this contributes to differences in TE activity and genome size evolution. Evidence for intraspecific variation in transposon activity and their connection to differences in host-silencing ability is consistent with the potential for silencing changes to drive TE activity and genome size evolution (Nuzhdin et al. 1998; Vu and Nuzhdin 2011; Blumenstiel 2011). Furthermore, recent comparative studies in *Arabidopsis* further suggest species differences in host silencing may be a plausible contributor to genome size evolution. The model plant *A. thaliana* has one of the smallest genomes in the mustard family and is more than 1.5-fold smaller than its closely related congener *Arabidopsis lyrata*. Recent genome comparisons reveal a striking consistency in the reduction of TE copy number in *A. thaliana*; all major families show consistently fewer TE insertions (Lockton and Gaut 2010; Hu et al. 2011). Comparisons of the siRNA
complement of TEs in both species confirmed that TE insertions with higher levels of siRNA targeting show lower levels of expression, consistent with the notion that siRNAs are key regulators of TE activity (Hollister et al. 2011). Furthermore, the analysis revealed fewer uniquely matching siRNAs to TE insertions in *A. lyrata*, possibly suggesting that a less efficient host silencing system in this species may be contributing to the higher levels of TE accumulation and thus larger genome size. While the pattern does not necessarily reflect a causal connection from differences in silencing to higher TE expression and transposition rates (see below), it does provide important evidence that differences in host silencing across species could contribute to genome size evolution.

Beyond hybridization, what factors might drive species-specific differences in the efficacy of host silencing? Understanding the evolution of silencing efficiency needs to consider evolutionary dynamics of both transposons and their hosts, which we now consider in the next section.

### 5.4 The evolution of genomic virulence

Further reasons why not all TEs are silenced come from the fact that, much like other pathogens, TEs have evolved a variety of countermeasures to avoid host silencing. A common way for viruses to avoid host defence mechanisms is to actively interfere with the silencing process (Roth et al. 2004). Similar approaches have been reported for TEs. Recent evidence in *Drosophila* suggests that genes involved in small RNA-mediated
silencing are under recurrent positive selection (Kolaczkowski et al. 2010; Obbard et al. 2010) consistent with the notion of an ongoing co-evolutionary arms race between hosts and transposons. One example for TE evasion of the host immune system is the *MuDR* and *Spm* families in maize. Both elements are capable of demethylating (removing the silencing caused by methylation) family members and so returning to normal transposition rates (Lister et al. 2008; Cui and Fedoroff 2002). Another silencing-avoiding mechanism that has been suggested is transduplication, a process that involves capturing host gene sequence inside the TE (Juretic et al. 2005; Lisch 2005). In maize, there are around 3,000 portions of genes captured inside the terminal inverted repeats of *Mu* elements. Similar observations have been described in the *Helitron* (Lai et al. 2005b; Morgante et al. 2005), *CACTA* (Wicker et al. 2003; Paterson et al. 2009), and *Harbinger* superfamilies (International Brachypodium Initiative 2010). TEs may then avoid methylation by confusing the host defence system, as silencing a TE may silence the captured gene (Hollister et al. 2011). Further support for transduplication as a potent countermeasure comes from the observation that the expression of *pack-MULEs* is correlated with the number of host gene fragments they contain (Hanada et al. 2009). However, the causal link between gene capture and host avoidance has yet to be established in most of the above examples. An alternative explanation to the observation could be that these TEs have been co-opted for host gene function (Jian et al. 2004).

Studying the conditions under which self-regulation, the ability of TEs to control their own transposition rate, has evolved provides further evidence for selection on both host and elements. Recombination rate plays an important role in determining whether self-regulation will be favoured (Charlesworth and Langley 1986). High rates of
recombination will typically result in the break-up of the association of the parental and daughter copies of the TEs, which as long as the insertion is not dominant lethal or sterility inducing, will reduce the advantage of self-regulation. By the same reasoning, selection for self-regulation can be expected to be higher in asexual or highly selfing species than in sexual or outcrossing species, as the association between copies and negative fitness effects will be stronger in asexual and selfing species. Thus, an asexual lineage with TEs exercising self-regulation should out-compete an asexual lineage where this is not the case. In general, in situations where the fitness effects of host and element overlap, self-regulation is more likely.

The co-evolutionary arms race that characterizes TE evolution is not restricted to the interaction between elements and hosts but also takes place between elements. Orgel and Crick (1980) famously called selfish DNA the ultimate parasite. González and Petrov (2009a) later narrowed this down to non-autonomous TEs (TEs that cannot transpose without the help of another TE) since they even find ways to parasitize TEs themselves. The relationship between autonomous and non-autonomous miniature inverted repeat TEs (MITEs) has been examined in rice by looking at 119 different combinations of mariner-like autonomous and non-autonomous stowaway MITEs (Yang et al. 2009). The autonomous element transposes the non-autonomous element in only 3% of cases. However, when non-autonomous elements do transpose, they may transpose at a 3,000-fold higher rate than the autonomous equivalent. This happens when the non-autonomous element lacks a sequence motif that caused suppression of the autonomous element transposition. Non-autonomous TEs are targeted less frequently by siRNAs and can therefore be considered a means of avoidance of host silencing (Tenaillon et al. 2010).
This could counter the disadvantage that is associated with being unable to transpose autonomously.

5.5 Determining the outcome of TE–host interactions

What factors are likely to influence whether TE evasion or host silencing predominate, and thus the expected outcome for genome size evolution? Mating system should be a strong candidate. Highly selfing and asexual species are more likely to evolve self-regulated transposition, and additionally host-silencing mechanisms are more likely to spread to fixation with greater linkage to the active element (Charlesworth and Langley 1986). The evidence to date does suggest that asexual and highly selfing lineages have smaller genomes than outcrossing lineages (Govindaraju and Cullis 1991; Wright and Finnegan 2001; Albach and Greilhuber 2004; Wright et al. 2008, but see Whitney et al. 2010) and that TE abundance may be lower in selfing species (Morgan 2001). This is consistent with the prediction that self-regulation and host regulation will be more efficient with greater linkage. However, theoretical models also suggest that TE abundance may decrease in asexual and selfing lineages without the need for changes in TE activity (Dolgin and Charlesworth 2006; Morgan 2001; Wright and Schoen 1999). Thus, once again, it remains difficult to distinguish effects of host removal from changes in transposition rate. Furthermore, a recent study using phylogenetic independent contrasts showed only a weak effect of selfing rate on genome size (Whitney et al. 2010). However, mating system shifts may occur rapidly in response to environmental change (reviewed in Pannell 2009). Thus, rather than examining patterns across large
phylogenies, comparisons of species pairs may offer a more powerful test of the impact of mating system on genome size.

The consistently lower TE abundance in the highly selfing *A. thaliana* compared with the outcrossing *A. lyrata* conforms to the expected mating system effect on TE activity (Lockton and Gaut 2010; Hu et al. 2011). In addition, the pattern is also in contrast to predictions based on the efficacy of natural selection due to differences in effective population size and ectopic recombination rates. Analysis of the divergence of LTR sequences and the phylogenetic patterns of TE diversification in *Arabidopsis* suggest that there are more recent transposition events in the outcrossing *A. lyrata* compared with *A. thaliana*; there is a particular deficiency of young TEs relative to old insertions in *A. thaliana*. Furthermore, analyses of TE insertion polymorphism patterns (Wright et al. 2001; Lockton et al. 2008) do not suggest that selection against individual TE insertions themselves is more efficient in *A. thaliana*, suggesting instead that more efficient host silencing and/or greater self-regulation may be occurring. While it remains to be seen whether these differences reflect changes in host silencing (Hollister and Gaut 2009) or TE self-regulation, these patterns are consistent with the notion that TE activity evolution, rather than host removal, may be driving changes in genome size. On the other hand, it is worth remembering that *A. thaliana* generally has fewer genes, and shows a consistent pattern of DNA loss, so it is still unclear whether changes in TE activity compared to changes in rates and selection on DNA removal predominate (Hu et al. 2011). Nevertheless, recent evidence suggesting species differences in siRNA targeting (Hollister and Gaut 2009; Hollister et al. 2011) does imply the evolution of TE activity as potentially playing an important role.
The selective advantage of new host silencing mutations is likely to be weak, given that selection coefficients will be on the order of the transposition rates (Charlesworth and Langley 1989). Because of this, host-silencing ability may be more efficient in species with larger effective population sizes, possibly contributing to an association between effective population size and genome size. Interestingly, however, the efficacy of selection on TEs for higher virulence should also be a positive function of the effective population size; in larger populations, selection favouring TE variants that are more able to evade host silencing should be more efficient. Thus, this is a factor that operates against the common prediction (Lynch and Conery 2003; Lynch 2007) that species with larger effective population sizes will be better able to eliminate insertions and will have smaller genomes. Furthermore, when considering silencing mechanisms via RNA targeting, one might expect that divergent TEs that show reduced similarity to corresponding RNAs may be better able to evade the host immune system. As a result, TE diversification and thus host evasion can be expected to be higher in host species with a larger effective population size. While this may be countered in part by more efficient selection, against new insertions and for host silencing, higher transposition rates might be a general prediction in species with larger population size and could drive genome size evolution in surprising directions.

Clearly, more theoretical and empirical work is needed in this area. It is possible that species with intermediate effective population sizes are expected to have the greatest TE accumulation as a function of both transposition rates and selection from the host and this might explain in part recent work (Whitney et al. 2010; Whitney and Garland 2010).
showing little relationship between effective population size and genome size once phylogeny is controlled for.

5.6 Conclusions

Obviously, differences in TE activity are not the only contributors to genome size evolution. The wide taxonomic contrast of humans and *Drosophila* highlights this well: TE activity appears to be considerably lower in humans (International Human Genome Sequencing Consortium 2001), yet genome size is over an order of magnitude larger. Nevertheless, the results reviewed here do highlight the potential importance of changes in transposition rate in driving the evolution of TE abundance and genome size. Evidence from studies of hybridization and mating system shifts are consistent with theoretical models of TE activity changes, although it remains difficult to rule out alternative explanations. In the *Arabidopsis* contrast, for example, any factor that leads to lower accumulation of new insertions in *A. thaliana* compared with *A. lyrata* might be able to explain the patterns, so distinguishing effects of host removal from changes in transposition rate are not trivial (Hu et al. 2011).

As our understanding of the mechanisms of host silencing and TE evasion increase, more direct comparisons of TE activity in distinct species should enable powerful tests of the role of changes in transposition rate compared to host removal on TE abundance and genome size evolution. Furthermore, a greater understanding of these mechanisms combined with more theoretical work integrating the mechanistic details (Lu
and Clark 2010) will help make predictions about other factors influencing transposition rate beyond those discussed here. With emerging sequencing technologies, we gain an increasing ability to quantify changes in transposition rate, determine whether these changes are due to host factors or the evolution of TE activity itself, and finally to establish the extent to which such changes contribute to genome size evolution.
CHAPTER 6 NO EVIDENCE THAT SEX AND TRANSPOSABLE ELEMENTS DRIVE GENOME SIZE VARIATION IN EVENING PRIMROSES

Abstract

Genome size varies dramatically across species, but despite an abundance of attention there is little agreement on the relative contributions of selective and neutral processes in governing this variation. The rate of sex can potentially play an important role in genome size evolution because of its effect on the efficacy of selection and transmission of transposable elements. Here, we used a phylogenetic comparative approach and whole genome sequencing to investigate the contribution of sex and transposable element content to genome size variation in the evening primrose (*Oenothera*) genus. We determined genome size using flow cytometry for 30 species that vary in genetic system and find that variation in sexual/asexual reproduction cannot explain the almost two-fold variation in genome size. Moreover, using whole genome sequences of three species of varying genome sizes and reproductive system, we found that genome size was not associated with transposable element abundance; instead the larger genomes had a higher abundance of simple sequence repeats. Although it has long been clear that sexual reproduction may affect various aspects of genome evolution in general and transposable element evolution in particular, it does not appear to have played a major role in genome size evolution in the evening primroses.
6.1 Introduction

Variation in genome size is one of the most striking examples of biodiversity (Bennett and Leitch 2012; Gregory 2013). Genomes may be as small as 112 kb, as in the obligate endosymbiotic proteobacterium *Nasuia deltocephalinicola* (Bennet and Moran 2013), or as large as 150 Gb in the polyploid plant *Paris japonica* (Pellicer et al. 2010). Variation is not restricted to differences between species, as extensive genome size variation also exists within species (Bioment 2008; Diez et al. 2013; Long et al. 2013). Understanding the evolutionary processes underlying this variation has received much attention (e.g. Petrov 2001; Gregory 2005; Lynch 2007; Gaut and Ross-Ibarra 2008; Ågren and Wright 2011), yet little consensus exists on the relative contributions of these processes.

Variation in genome size may be influenced by both neutral and selective evolutionary processes. Several studies have shown that genome size may evolve neutrally, with increases and decreases mainly due to biases in insertion and deletion rates and/or recombination rates (Petrov 2001; Oliver et al. 2007; Nam and Ellegren, 2012). Genome size may also correlate with various ecologically relevant traits, with examples ranging from flowering time in plants (Meagher and Vassiliadis, 2005) to song attractiveness in grasshoppers (Schielzeth et al. 2014), and selection may thus act adaptively on genome size. Much recent debate has focused on the hypothesis that variation in the efficacy of selection, usually due to differences in effective population size (*N_e*), governs most variation in genome size across distantly related taxa (see for example exchanges by Lynch and Conery 2003; Charlesworth and Barton 2004; Daubin and Moran 2004; Lynch and Conery 2004; Whitney et al. 2010, Whitney and Garland
Finally, several studies show that differential accumulation of transposable elements (TEs) can explain differences in genome size (reviewed in e.g. Tenaillon et al. 2010; Ågren and Wright 2011; Michael 2014). For example, genome size variation among closely related species, including species of rice (Piegu et al. 2006), cotton (Hawkins et al. 2006), and Arabidopsis (Hu et al. 2011), can largely be attributed differences in TE abundance. Understanding what factors allow TEs to accumulate in some species, but not in others, may therefore be central to understanding variation in genome size in general, and among plants in particular.

Sex could be a key factor affecting genome size through its (potentially contrasting) effects on efficacy of selection and TE transmission (Hickey 1982; Wright and Schoen 1999; Morgan 2001; Docking et al. 2006; Dolgin and Charlesworth 2006; Glemin and Galtier 2012; Ågren 2014). However, whether the loss of sex will cause an increase or decrease in genome size remains unclear. On the one hand, sexual reproduction provides a way for TEs to spread to new lineages in a population, and despite the deleterious effects on host fitness, theory predicts that TEs in sexual populations should evolve maximum transposition rates (Charlesworth and Langley 1986). By contrast, within-lineage transmission of genes in asexual populations is expected to limit TE spread, and in the absence of horizontal gene transfer, reduction in transposition rate due to self-regulation is also more likely to evolve in highly selfing and asexual species (Charlesworth and Langley 1986). This scenario predicts higher TE abundances and larger genome sizes in sexually reproducing species compared to asexually reproducing relatives. On the other hand, asexual lineages typically evolve from sexual progenitors and will thus inherit their TE load. The reduced effective
population size in asexuals may then allow genetic drift to fix TEs by a Muller’s ratchet-like process at a higher rate in asexuals than in sexuals (Muller 1964; Gabriel et al. 1993; Dolgin and Charlesworth 2006; Charlesworth 2012). Thus, if genetic drift is more important than transmission in the accumulation of TEs, we expect to see smaller genomes in sexual lineages. Ultimately, how the effects of transmission and genetic drift will balance out, and thus whether sex will cause an increase or decrease in genome size will depend in important ways on effective population sizes, transposition and excision rates in natural populations (Dolgin and Charlesworth 2006). Thus the fate of transposable elements in sexual vs. asexual systems represents an important empirical problem for understanding the forces that govern the evolution of genome architecture.

Our understanding of the net effect of loss of sex on genome size evolution in natural populations has been limited by the lack of large-scale replicated comparisons. Arguably the most important study to date of the role of sex on TE abundance comes from the ancient asexually reproducing bdelloid rotifers, where the few TEs that exists seem to have been recently acquired through horizontal gene transfer (Arkhipova and Meselson 2000; 2005; Flot et al. 2013). Sexual *Daphnia pulex* harbour higher TE levels than asexuals (2010). Similarly, Zeyl et al. (1996) introduced *Ty3* retrotransposons into experimental sexual and asexual yeast lines and showed that the new TEs spread faster and reached higher abundances in the sexual compared to the asexual lines. Finally, Docking et al. (2006) found no evidence for a signal of relaxed selection on TE encoded proteins in a set of asexual plant species compared with sexual relatives. However, the extent to which transitions to asexuality can drive recurrent changes in genome size and TE abundance in natural populations is poorly understood.
Unlike with asexuals, considerable work has been done comparing genome size and transposable element abundance in selfing and outcrossing species (e.g. Lockton and Gaut 2010; de la Chaux et al. 2012; Ågren et al. 2014). Several studies have shown an association between outcrossing rate and genome expansion (Albach and Greilhuber 2004; Trivers et al. 2004; Wright et al. 2008; Hu et al. 2011), but it is difficult to distinguish whether these differences are due to TE accumulation in the outcrossers or loss of TEs in the selfers (Wright and Ågren 2011). After correcting for phylogenetic non-independence, Whitney et al. (2010) failed to detect a significant effect of outcrossing rate on genome size in a broad scale comparison of 205 species. Nevertheless, the effect of selfing rate approached significance ($P = 0.066$), and there are reasons to believe large-scale comparisons of a phylogenetically diverse sample of species may fail to capture the effect of mating system. Because both mating system and genome size often evolve rapidly, the short timescale of an intra-genus comparison can provide details not available from a broader multi-family approach with sparse taxon sampling, where transitions in mating systems may be harder to link directly to corresponding changes in genome size (Ågren and Wright 2011; Glémin and Galtier 2012). Furthermore, mating system is known to correlate with various life history traits, which may confound inferences about the effect of mating system on genome size. Testing the association between mating systems on smaller time scales and among species with similar life-history reduces the risk of confounding factors.

Predictions of the effects of asexuality on genome size are largely similar to those of selfing, but the effects should be magnified in asexuals and may therefore offer a more powerful test than selfing vs. outcrossing (Glémin and Galtier 2012). First, obligate
asexuals have the complete absence of outcrossing, whereas highly selfing species still experience occasional outcrossing (Vogler and Kalisz 2001; Igić and Kohn 2006), leading to low but significant levels of between-individual transmission. Furthermore, diploid asexuals will be associated with permanent heterozygosity, whereas selfers experience increased homozygosity. If selection against TEs is predominately due to chromosomal rearrangements caused by ectopic recombination between non-homologous sites, which will increase with increased heterozygosity (i.e. underdominance) (Charlesworth and Charlesworth 1983; Langley et al. 1988) selection against TEs may be stronger in asexuals compared to sexuals. Thus, stronger selection against ectopic recombination combined with a complete absence of genetic exchange through outcrossing may drive a faster and more extreme genome reduction in asexual lineages.

Genus-wide studies of the effect of sexual reproduction on plant genome size evolution have so far been lacking and the evening primrose plant genus *Oenothera* (Onagraceae) provides a promising system for addressing this problem. Functional asexuality (i.e. absence of recombination and segregation) in the family has evolved more than 20 times independently due to a genetic system called Permanent Translocation Heterozygosity (PTH), which is characterized by suppression of meiotic recombination and segregation (Cleland 1972; Harte 1994; Rauwolf et al. 2008; Johnson et al. 2009). Functional asexuality due to PTH has been described in eight plant families (Cleland 1972; Holsinger and Ellstrand, 1984; Harte 1994), and differs from apomixis in that individuals go through all stages of meiosis, and successful zygote formation still requires fertilization (see Whitton et al. 2008 for a review of other forms of plant asexuality). Thus, whereas many empirical systems used to study the genomic
consequences of loss of sex are associated with only one or a few transitions, the repeated independent transitions to functional asexuality makes *Oenothera* a powerful naturally replicated experiment (Dietrich et al. 1997; Wagner et al. 2007; Johnson et al. 2011). Sexual and PTH species often live in sympatry with overlapping ranges, and show little difference in life history and non-floral morphological traits. The consequences of a transition to PTH can therefore be investigated independently of ecological requirements, which is often difficult in other systems (Glémin and Galtier 2012). Lastly, PTH species tend to share the same ploidy level with their sexual relatives, allowing the effect of reproductive system to be decoupled from the effect of ploidy.

In this study we took a phylogenetic comparative approach to examine whether the repeated transitions from sex to functionally asexual PTH reproduction in *Oenothera* have been associated with a shift in genome size. We also used whole genome sequencing of three species of varying genome size and reproductive system to assess the contribution of transposable elements to genome size variation in the genus.

### 6.2 Material and Methods

#### 6.2.1 Study system

In this study we focus on the *Oenothera* genus of the evening primrose family Onagraceae. The genus is monophyletic (Levin et al. 2004; Wagner et al. 2007) and includes the largest number of PTH species and their sexual relatives (Raven 1979; Holsinger and Ellstrand 1984; Johnson et al. 2011). PTH in *Oenothera* results from three
mechanisms (Cleland 1972; Harte, 1994; Rauwolf et al. 2008). First, during metaphase I of meiosis, all 14 chromosomes form a ring \((x = n = 7)\) instead of bivalents, which restricts synopsis and recombination to the highly homozygous telomeres, such that genetically detectable recombination is effectively 0. Second, in anaphase I, alternate disjunction results in one haploid set of chromosomes segregating as a unit, and the other haploid set as another unit. Unless one of each unit is present in the zygote, the zygote will not survive. This balanced mortality of gametes prevents segregation, which leads to permanent heterozygosity. Finally, > 99.5\% of seeds are self-fertilized (R Godfrey and MTJ Johnson, unpublished results) because receptive stigmas accept pollen before flowers open. In short, the genetics of PTH reproduction in *Oenothera* can be likened with splitting the genome in half; only to later fuse the two halves back together, without recombination or segregation (Cleland 1972; Harte, 1994; Rauwolf et al. 2008). Recent evidence shows that recombination is also suppressed in sexual bivalent forming species, suggesting that recombination rates may not be dramatically different between sexual and PTH species. However, sexual species still undergo free segregation of homologous chromosomes, which should effectively eliminate genetic linkage disequilibria between chromosomes and allow the formation of homozygous loci from heterozygous parents as per Hardy-Weinberg expectations (Rauwolf et al. 2011; Golczyk et al. 2014). By contrast, these processes are completely lost in PTH species, leading to the perpetual propagation of single genotypes (Stebbins 1950; Cleland 1972).

As mentioned above, PTH has evolved more than 20 times independently in the Onagraceae. Although divergence times for the whole *Oenothera* clade have yet to be determined, estimates from population transcriptome samples suggest that the divergence
times for species in this study range from 5,500 years (\textit{O. biennis} and \textit{O. nutans}), 12,000 years (\textit{O. laciniata} and \textit{O. heterophylla}), to 500,000 (\textit{O. rosea} and \textit{O. tetraptera}) (Hollister et al. 2015).

### 6.2.2 Genome size estimates

In this study we examined genome size in 30 \textit{Oenothera} species, including 17 PTH and 13 sexual species (Figure 6.1). Sterilized seeds were stratified by sowing them on agar and kept in the dark at 4°C for three weeks. Seedlings were transferred to pots and grown in controlled glasshouse conditions between October 2013 and January 2014 at the University of Toronto. Leaf tissue, sampled at the same level of maturity (~ 5 weeks), from each species was sent to Plant Cytometry Services (PO Box 299, 5480 AG Schijndel, The Netherlands), who determined DNA content in picograms (pg) using flow cytometry with Propidium Iodide fluorescent dye and \textit{Pachysandra terminalis} as a standard (1C = 1.73; Zonneveld et al. 2005). This method has previously been shown to successfully detect small differences in genome size, including within species variation (Diez et al. 2013). Three replicates were performed per species.

### 6.2.3 Phylogenetic analysis

To control for statistical non-independence due to shared evolutionary history (Felsenstein 1985), we accounted for phylogeny in our statistical analysis. We inferred the phylogeny of the 30 species using the previously generated phylogeny of \textit{Oenothera}
in Johnson et al. (2009). Briefly, that analysis used sequences from two plastid (trnL-trnF and rps16) and three nuclear gene regions (PgiC, ITS, and ETS) from 121 species and created a maximum likelihood phylogeny using RAxML 7.0.4 (Stamatakis 2006). The tree was made ultrametric using non-parametric rate smoothing in TreeEdit (http://tree.bio.ed.ac.uk/software/treeedit). We pruned the tree to include only the 30 species studied here. For further details of the phylogeny see Johnson et al. (2009).

To start, we tested whether the data exhibited significant phylogenetic signal using Pagel’s $\lambda$ (Pagel 1999) implemented in the phylosig function in the phytools package (Revell 2012) of R version 3.0.2 (R Development Core Team 2013). This package assesses the significance of phylogenetic signal by performing a likelihood ratio test against the null hypothesis that $\lambda = 0$. Next, we performed phylogenetic generalized least squares (PGLS; Butler and King 2004) regression between reproductive system (sex was coded as 0 and PTH as 1) and genome size using the ape (Paradis et al. 2004) and geiger (Harmon et al. 2008) packages in R. We performed the PGLS tests under both neutral (Brownian motion) and stabilizing selection (Ornstein-Uhlenbeck) models of trait evolution. Akaike's Information Criterion (AIC) was used to determine which model best described the data. Prior to analysis, all genome size estimates were log10 transformed to ensure that the trait was modelled as having an equal probability of increasing or decreasing (c.f. O'Meara et al. 2006; Oliver et al. 2007; Lysak et al. 2009).
6.2.4 DNA isolation and Illumina sequencing

Seeds of *O. elata* (sex) and *O. biennis* (PTH), two species with relatively large genome sizes, and of *O. villaricae* (PTH), one of relatively small genome size (Supplementary Table 6.1), were imbibed at 4°C overnight in the dark with 3% Plant Preservative Mixture (Plant Cell Technology Store, Washington, DC, USA). Seeds were then sown on the substrate surface and grown until the early rosette stage (for details see Greiner and Köhl, 2014). For DNA isolation, 50 mg of leaf material was frozen in liquid N2 and ground using a mixer mill. After adding 775 µl of IGEPAL-buffer [1.9% IGEPAL, 1.9% CTAB, 130 mM Tris/HCl (pH 8.0), 130 mM EDTA (pH 8.0), 1.3% PVP-40, 1.8 M NaCl, 130 mM B(OH)3], 125 µl 1-thioglycerol, and 100 µl RNase A (50 mg/ml), samples were incubated at 60°C for 30 minutes under medium shaking. Cell debris was removed by centrifugation. The supernatant was treated twice with chloroform/isoamylalcohol (24:1) and subsequently with phenol/chloroform/isoamylalcohol (25:24:1, pH 7.5). The phenol/chloroform/isoamylalcohol treatment was repeated until both aqueous- and interphase, were clear. DNA was precipitated with 1/10 volume of 5 M NH4-acetate and 1 volume of isopropanol (-20°C), incubated at -20°C for 15 min, collected by centrifugation, and washed with 70% and/100% EtOH (-20°C). Resolved DNA was further purified using the Genomic DNA Clean & ConcentratorTM kit (Zymo Research Cooperation, Irvine, CA, USA). DNA was sequenced at the Max Planck-Genome-Centre Cologne (Germany) on an Illumina HiSeq2500 platform, 100 bp paired-end, utilizing a TruSeq DNA library (375 bp insert size). All sequences are available at the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra; Accession: PRJNA273500, ID: 273500).
6.2.5 Repetitive content analysis

To assess whether variation in genome size could be attributed to differential accumulation of repetitive elements such as TEs, we determined the repetitive content in three species of varying genome size and reproductive system. To characterize repeats we ran RepeatExplorer (Novak et al. 2013), a de novo graph-clustering pipeline for repeat characterization (Novak et al. 2010), which is implemented in the Galaxy platform (http://galaxyproject.org). As an alternative, read mapping approaches might have been used, but in the absence of an assembled reference genome for Oenothera or a close relative, such approaches are currently unfeasible. In addition, a comparison of TE-detection using read mapping and RepeatExplorer revealed an underestimation of TE abundance in interspecific comparisons of Cacao relatives, when read mapping was employed (Sveinsson et al. 2013). Hence, we are confident that graph based clustering methods, such as RepeatExplorer, are more suitable for interspecific comparisons of transposons than read mapping approaches (Sveinsson et al. 2013; but see Tenaillon et al. 2011 for a successful attempt at estimating interspecific differences in TE abundances using the well characterized maize reference genome).

For RepeatExplorer we filtered reads for quality, keeping only reads with a Phred quality score of at least 20 over 90% of their length. RepeatExplorer joins reads together in clusters based on sequence similarity, recording reads with similarity hits greater than 80% across 55% of the sequence length, and then matches these clusters against RepBase (Jurka et al. 2005) to identify repeats. We ran the pipeline on one sample per species under default settings (Table 6.1).
Table 6.1 Clustering statistics for RepeatExplorer

<table>
<thead>
<tr>
<th></th>
<th>O. villaricae</th>
<th>O. elata</th>
<th>O. biennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of reads used in analysis</td>
<td>1001571</td>
<td>1012981</td>
<td>729978</td>
</tr>
<tr>
<td>Number of reads in clusters</td>
<td>813712</td>
<td>857862</td>
<td>639924</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>24496</td>
<td>19474</td>
<td>12197</td>
</tr>
</tbody>
</table>

6.3 Results

6.3.1 Variation in genome size

We detected almost two-fold variation in genome size among the diploid species surveyed (Figure 6.1; Supplementary Table 6.2). Estimated genome size ranged from 0.64 pg in *O. mendocinensis* (sex) and *O. sandiana* (PTH), to 1.16 pg in *O. nutans* (PTH). The South American clade (Subsection Munzia: *O. mendocinensis*, *O. nana*, etc.) consistently had the smallest genomes, whereas the recently radiated North American *O. biennis* clade (Subsection *Oenothera*) consistently had the largest genome sizes. The so-called “B clade” (sensu Wagner et al. 2007; e.g. *O. flava*, *O. tetraptera*, *O. perennis*, *O. fruticosa*, *O. rosea*, *O. kunthiana*) varied the most in genome size, reflecting polyploidy (exact ploidy level unknown) in one species (*O. fruticosa*), and deep divergence among multiple lineages. Excluding the polyploid *O. fruticosa* (sex), the mean genome size across all species was 0.85 +/- 0.036 s.e. pg, for PTH species 0.85 +/- 0.048 s.e. pg, and for sexually reproducing species 0.86 +/- 0.057 s.e. pg. C-value estimates are available at Dryad (doi:10.5061/dryad.7mf4k).
6.3.2 Phylogenetic signal

We detected significant phylogenetic signal in genome size across the species examined. Pagel’s $\lambda$ was 0.72 ($P = 0.0277$ that $\lambda > 0$) suggesting that phylogenetic non-independence should be taken into account in statistical analyses.

6.3.3 Phylogenetic generalized least squares analysis

We found no significant relationship between sexual reproduction and genome size. The lack of an effect of sex on genome size holds regardless of whether we assume that genome size evolves under a neutral Brownian motion model (df = 29, $P = 0.828$) or moving towards a selective optimum in an Ornstein-Uhlenbeck’s model of stabilizing selection (df = 29, $P = 0.8162$). Comparing the AIC scores suggests that the Brownian motion model (AIC = -103.4203) better describes the data than the Ornstein-Uhlenbeck model (AIC = -42.8108).
Figure 6.1 Molecular phylogeny of *Oenothera* used in the PGLS analysis

Haploid genome size estimates (pg; 1 pg = 978 Mb) are plotted on top. Note, *O. fruticosa* is a polyploid and not included in the statistical analyses. Although divergence times for the whole *Oenothera* clade have yet to be determined, estimates from population transcriptome samples suggest that the divergence times for species in this study range from 5,500 years (*O. biennis* and *O. nutans*), via 12,000 years (*O. laciniata* and *O. heterophylla*), to 500,000 years (*O. rosea* and *O. tetraptera*) (Hollister et al. 2015).
6.3.4 Repetitive content

Repetitive elements were abundant in all three species examined. After filtering for quality, about 80% of all reads formed clusters (Table 6.1). In all three species, TEs made up most of the repetitive content, with the dominant TEs being long terminal repeat gypsy and copia elements (Figure 6.2). In all three species, we estimated that TEs make up ~ 35-40% of the genome.

To investigate whether the genomes differed in other kinds of repeats, we combined the sequences annotated as “simple repeat”, “satellite”, and “low complexity” under the label “short simple repeats”. These repeats made up a larger proportion in the relatively large genomes *O. elata* and *O. biennis* (~ 35%), than in the smaller *O. villaricae* (~ 20%; Figure 6.2) and this difference was statistically significant (Pearson’s chi-square test of independence, $\chi^2 = 59360.79$, df = 2, $P = 2.2 \times 10^{-16}$).
Figure 6.2 Repetitive content of three *Oenothera* species

Repetitive content estimated by RepeatExplorer (Novak et al. 2013), including two species of relatively large genome size (*O. elata* and *O. biennis*) and one of relatively small genome size (*O. villaricae*).
6.4 Discussion

Studies of genome size variation have a long history (Mirsky and Ris 1951; Gregory 2005). References to this variation have featured heavily in the arguments about the role of non-selective processes in the evolution of genome complexity (Lynch and Conery 2003; Lynch 2007; Whitney and Garland 2010; Whitney et al. 2010; Lynch 2011; Whitney et al. 2011) and more recently in the debate associated with the ENCODE Project Consortium’s claim that 80% of the human genome can be assigned a biochemical function (ENCODE Project Consortium 2012; Eddy 2012; Graur et al. 2013; Doolittle 2013; Kellis et al. 2014; Palazzo and Gregory 2014). With estimates from some 7,000 species and a 2,400-fold variation in genome size (Leitch and Leitch 2013), studies of plant genomes have much to contribute to these and other debates.

Here, we presented genome size estimates from thirty species in the evening primrose genus *Oenothera* and performed the first comprehensive study of the role of sex on genome size evolution in plants. We found no evidence that sex explains the almost two-fold variation in genome size. Instead, evolution of genome size was fairly conserved within *Oenothera* and best explained by neutral genetic drift, as opposed to a model of stabilizing selection towards an optimum, or a model that ignores evolutionary history. Moreover, our preliminary analysis based on three genomes of varying size and reproductive system suggests that, contrary to the reasoning outlined in the Introduction, genome size variation in *Oenothera* can be attributed to accumulation of short simple repeats, rather than transposable element abundance. As more whole genome data
becomes available, future work should test the generality of this result across *Oenothera* and other plant systems.

Although TEs have undoubtedly played a key role in driving genome size variation in plants (Tenaillon et al. 2010; Ågren and Wright 2011; Michael 2014), our results are in line with a number of studies highlighting examples of genome size variation due to non-TE repetitive elements. An early example comes from Flax, where a variety with 16% larger genome had 50% more rDNA (Cullis 1976). More broadly, rDNA are particularly common in many large conifer genomes (Ahuja and Neale 2005), including spruce (Bobola et al. 1992) and a general correlation between rDNA abundance and genome size in eukaryotes has been suggested (Birnstiel et al. 1971; Prokopowich et al. 2003). Furthermore, in *Arabidopsis*, the highly selfing *A. thaliana* has fewer TEs and smaller genome than its outcrossing relative *A. lyrata* (Hu et al. 2011), which is likely due to accumulation in the outcrosser rather than a loss of TEs in the selfer (Slotte et al. 2013), consistent with TE transmission advantage in an outcrossing lineage. However, genome size variation within *A. thaliana* does not seem be due to TEs. In particular, Long et al. (2013) detected 10% variation in genome size among 180 Swedish *Arabidopsis thaliana* lines, which was due to differential accumulation of 45S rDNA rather than TEs. In *Eucalyptus*, a member of the same order as *Oenothera* (Myrtales), a difference in abundance of tandem repeats rather than TEs is responsible for the 110 Mb difference in genome size between the closely related *E. grandis* and *E. globulus* (Myburg et al. 2014). Finally, genome size variation within and between maize species can be attributed to various repetitive elements (Rivin et al. 1986; Chia et al. 2012; Diez et al. 2014). For example, 70% of the genome size difference between *Zea mays* and *Z. luxurians*, and
50% of the 1.5-fold difference in size between *Z. mays* and the sister genus *Tripsacum*, is due to TEs (Tenaillon et al. 2011; Chia et al. 2012). However, similar to the *Arabidopsis* example, genome size variation within *Z. mays* is not due to TEs but to chromosomal knob content (Chia et al. 2012). Thus, strong evidence exists for both key roles of TEs and other repetitive elements. An important frontier remains understanding when genome size variation will be due to TEs or other kinds of (simple) repeats.

As outlined in the Introduction, whether a transition to asexuality should be associated with genome expansion or shrinkage will depend on the relative importance of genetic drift and the transmission advantage of TEs in governing genome size evolution (Glémin and Galtier 2012). Although there is evidence from multiple systems that sex may promote the spread of TEs (Zeyl et al. 1996; Arkhipova and Meselson 2000; Schaack et al. 2010), there is also abundant evidence that asexuality is associated with a reduction in the efficacy of selection (reviewed in Glemin and Galtier 2012). Furthermore, if many insertions are recessive and/or ectopic recombination rates are low in translocation heterozygotes, TEs may experience further relaxed selection in diploid asexuals. Indeed, a recent study of 13 sexual and 16 PTH *Oenothera* species found evidence of a reduction in the efficacy of selection in PTH species (Hollister et al. 2015). Thus, any transmission TEs may enjoy in sexual *Oenothera* species will be counteracted with the reduction in the efficacy of selection in PTH species, and this balance may partly explain the lack of effect of sex on genome size in this study.

In addition to the contrasting effects of transmission of and selection against TEs, the short time scale due the recent divergence between species may mean that any effect of reproductive system have yet to materialize. Over longer time scales, asexual lineages
with high TE abundance are expected to go extinct, whereas asexual lineages with low TE abundance are more likely to persist (Arkhipova and Meselson 2005; Dolgin and Charlesworth 2006). However, under shorter timescales predictions are less clear (Docking et al. 2006). Despite the short timescale, the shift to PTH in *Oenothera* has been associated with a shift in a number of traits – ranging from plant defences and diversification rates to molecular evolution (Johnson et al. 2009; 2010; 2011; Hersch-Green et al. 2012; Hollister et al. 2015) highlighting that rapid evolutionary responses to the loss of sex are not implausible. Moreover, even for the oldest diverging PTH lineage (*O. rosea*), we see no signs of genome reduction.

Here, we have performed the first large-scale investigation of the role of sex in genome size evolution in plants. Whereas it has long been clear that sexual reproduction affects various aspects of genome evolution, including in the evening primroses (Ellstrand and Levin 1980; Hersch-Green et al. 2012; Hollister et al. 2015), it appears to have played no role in genome size evolution in *Oenothera*.

### 6.5 Acknowledgements

We thank Niroshini Epitawalage and Victor Mollov for assistance in the glasshouse, Brian Husband and Paul Kron for discussion and advice on flow cytometry, and Elena Ulbricht-Jones and Liliya Yaneva-Roder for providing the DNA samples for Illumina sequencing. Liliya Yaneva-Roder is further acknowledged for excellent technical assistance during the establishment of the phenol-based DNA isolation protocol for
Oenothera. Seeds were generously donated by USDA GRIN, Algonquin Park Visitor Centre staff, S.C.H. Barrett, Ornamental Plant Germplasm Centre (OPGC), J. Meurer, W. Wagner, and R. Raguso. Comments from three anonymous reviewers improved the manuscript. J.A.Â was supported by a Junior Fellowship from Massey College, S.G. by the Max Planck Society and the Deutsche Forschungsgemeinschaft (DFG; GR 4193/1-1), M.T.J.J. and S.I.W. by the Natural Sciences and Engineering Research Council (NSERC). We declare no conflict of interest.
### 6.6 Supplementary Information

#### Supplementary Table 6.1 Seed curation information for *Oenothera* plants used in flow cytometry and genome sequencing

<table>
<thead>
<tr>
<th>Seed curation ID</th>
<th>Species</th>
<th>Genetic System (PTH/SEX)</th>
<th>ploidy</th>
<th>used for</th>
<th>Strain/Locality</th>
<th>Original seed supplier</th>
<th>Collector notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/143 x s (11/205)</td>
<td><em>Oenothera elata</em> ssp <em>elata</em></td>
<td>SEX</td>
<td>diploid</td>
<td>Illumina sequencing</td>
<td>Strain: <em>elata</em> Toluca; originally collected by Munz, Toluca de Lerdo, México, Mexico, 1937</td>
<td>S. Greiner (originally from R. E. Cleland)</td>
<td>Genome: AA-I; Cytology: 7 prs; Evolution (1951) 5: 265-272; Greiner seed (plant) id: 04/143 x s (11/205); inbred</td>
</tr>
<tr>
<td>05/45 x s (12/192)</td>
<td><em>Oenothera biennis</em></td>
<td>PTH</td>
<td>diploid</td>
<td>Illumina sequencing</td>
<td>Strain: <em>biennis</em> Muenchen, syn: <em>biennis</em> Nymphenburg; originally collected by O. Renner, Nymphenburg Garden, Munich, Bavaria, Germany, 1914</td>
<td>S. Greiner (originally from O. Renner)</td>
<td>Genome: AB-II; Cytology: ring of 6, ring of 8; Z Indukt Abstamm Vererbungsl (1917) 18:121-294; Proc R Soc Biol Sci Ser B (1940) 128:509-535; Ber Dtsch Bot</td>
</tr>
<tr>
<td>MTJ-2</td>
<td>Oenothera elata cf. ssp. hookeri</td>
<td>PTH</td>
<td>diploid</td>
<td>Illumina sequencing</td>
<td>Strain: villaricae Antilhue; originally collected by W. Stubbe, along the railroad near Antilhue, Valdivia, Chile, 1961</td>
<td>S. Greiner (originally from W. Dietrich)</td>
<td>Ges (1950) 63:129-138; Greiner seed (plant) id 05/45 x s (12/192), inbred</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------</td>
<td>------</td>
<td>----------</td>
<td>---------------------</td>
<td>-------------------------------------------------</td>
<td>---------------------------------</td>
<td></td>
</tr>
<tr>
<td>MTJ-10</td>
<td>Oenothera perennis</td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Canisbay Twp, Algonquin Park, ON, Canada, (45deg 34'57.34&quot;N, 79deg 27'08.95&quot;W), 2006</td>
<td>M. Johnson (originally from A.P. Staff)</td>
<td>Genome: B^5-I^3; Cytology: ring of 14; Ann Mo Bot Gard (1977) 64: 425–629; Greiner seed (plant) id: 86/224 x s (13/41), inbred, some generations</td>
</tr>
<tr>
<td>MTJ-15</td>
<td>Oenothera villaricae</td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Valdivia Airport, Chile, 2006</td>
<td>M. Johnson (originally from S. Barrett)</td>
<td>From propagated seed grown in greenhouse in 2006</td>
</tr>
<tr>
<td>MTJ-20</td>
<td>Oenothera affinis</td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow</td>
<td>Rocky stream bed, Calle Los</td>
<td>M. Johnson (originally from Bulk collection of)</td>
<td></td>
</tr>
<tr>
<td>MTJ</td>
<td>Species</td>
<td>Method</td>
<td>Ploidy</td>
<td>Genome Size Notes</td>
<td>Collection Site</td>
<td>Collector</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>--------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------------------</td>
<td>-------------------------------------------------------</td>
<td>-----------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>MTJ-22</td>
<td><em>Oenothera biennis</em></td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Wasteland near docks, Toronto, ON, Canada (43deg 38'45.95&quot;, 79deg 21'32.12&quot;), 2005</td>
<td>M. Johnson</td>
<td>12 families</td>
</tr>
<tr>
<td>MTJ-24</td>
<td><em>Oenothera rosea</em></td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Sycamore Canyon, near Nogales, St. Cruz, AZ, USA, 1998</td>
<td>M. Johnson</td>
<td>From propagated seed grown in greenhouse in 2007</td>
</tr>
<tr>
<td>MTJ-39</td>
<td><em>Oenothera flava</em></td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Road cut along Hwy 44, Unitah Co., Unitah Mtns, Ashley N.F., UT, USA, 1978</td>
<td>M. Johnson</td>
<td></td>
</tr>
<tr>
<td>MTJ-40</td>
<td><em>Oenothera albicaulis</em></td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Buena Vista, Colorado, USA, 1995</td>
<td>M. Johnson</td>
<td>(originally collected by R. Raguso, ID: RAR-95-2)</td>
</tr>
<tr>
<td>MTJ-42</td>
<td><em>Oenothera laciniata</em></td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Columbia, SC, USA, 2002</td>
<td>M. Johnson</td>
<td>From propagated seed grown in greenhouse in 2007</td>
</tr>
<tr>
<td>MTJ-46</td>
<td><em>Oenothera longissima</em></td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Riparian habitat, Moab UT, USA, 1996</td>
<td>M. Johnson</td>
<td>(originally collected by R. Raguso, ID: RAR-96-1)</td>
</tr>
<tr>
<td>MTJ-59</td>
<td><em>Oenothera capillifolia</em> (=<em>O. berlandieri</em>)</td>
<td>SEX</td>
<td>Genome size; flow cytometry</td>
<td>Johnson Ranch Rd., 1-40, Donley co., TX, USA, 1999</td>
<td>M. Johnson (originally collected by R. Raguso, ID: RAR 99-66)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------</td>
<td>-----</td>
<td>----------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-103</td>
<td><em>Oenothera rhombipetala</em></td>
<td>SEX</td>
<td>Genome size; flow cytometry</td>
<td>Lake Thunderbird S.P., Cleveland Co, OK (35deg 13'58.05&quot;N, 97deg 14'46.94&quot;W), 2007</td>
<td>M. Johnson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-111</td>
<td><em>Oenothera fruticosa</em> ssp fruticosa</td>
<td>SEX</td>
<td>Genome size; flow cytometry</td>
<td>American Village, at jct of Constitution Ave and Hwy 751, Durham Co, NC, USA (36deg 00'30.81&quot;, 78deg 58'02.97&quot;), 2007</td>
<td>M. Johnson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-128</td>
<td><em>Oenothera kunthiana</em></td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Santa Ana Wildlife Refuge, Hidago Co, TX, USA (26deg 11'09.43&quot;N, 98deg 23'18.75&quot;W), 2007</td>
<td>M. Johnson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-179</td>
<td><em>Oenothera bahia-blancae</em></td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius at dunes ca. 2 km south east of Argerich in the research terrain of the Universidad del Sur, 37 km west of Bahia Blanca, Buenos Aires, Argentina, 9 Jan 1968 (MO-2155721, holotype; CTES,</td>
<td>M. Johnson (donated by OPGC [NSSL 115682], who received it from W. Stubbe, 1981)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Santarius 455; K.A. Santarius (1968/69) Oenothera-Sammelreise nach Südamerika. Jahrbuch der Heinrich-Heine-Universität Düsseldorf; Ann Mo Bot
<table>
<thead>
<tr>
<th>MTJ-186</th>
<th>Oenothera clelandii</th>
<th>PTH</th>
<th>diploid</th>
<th>Genome size; flow cytometry</th>
<th>Kitty Todd Nature Preserve, Lucas Co, OH, USA (41°37′6″N, 83°47′30″W), 2005</th>
<th>M. Johnson (donated by OPGC [OPGC 2487], who received it from R. Gardner et al.)</th>
<th>From propagated seed grown in greenhouse in 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTJ-201</td>
<td>Oenothera glazioviana</td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Varacot, Hungary</td>
<td>M. Johnson (donated by OPGC [NSSL 197319], who received it from P. Lapinskas, Hurst Corp, 1985)</td>
<td></td>
</tr>
<tr>
<td>MTJ-209</td>
<td>Oenothera heterophylla</td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by N. Ellstrand, Stockdale, TX, USA, 1978</td>
<td>M. Johnson (donated by OPGC [NSSL 165776], who received it from W. Stubbe, 1982)</td>
<td></td>
</tr>
<tr>
<td>MTJ-227</td>
<td>Oenothera mendocinensis</td>
<td>SEX</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius at dunes ca. 2 km south east of Argerich, 37 km west of Bahia Blanca, Buenos Aires, Argentina</td>
<td>M. Johnson (donated by OPGC [NSSL 165748], who received it from W. Stubbe, 1982)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-232</td>
<td>Oenothera nana</td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius at a waste places in east part of Potosí, 4,100 m, Tomás Frías, Bolivia</td>
<td>M. Johnson (donated by OPGC [NSSL 115654], who received it from W. Stubbe, 1981)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-236</td>
<td>Oenothera nutans</td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by Bell, Pipers Gap, Carol Co., VA, USA; Oct 1978</td>
<td>M. Johnson (donated by OPGC [NSSL 115798], who received it from W. Stubbe, 1981)</td>
<td>Genome: BB-III; Cytology: ring of 14; Plant Syst Evol (1990) 169: 69-80</td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Accession Type</td>
<td>Species</td>
<td>Ploidy</td>
<td>Genome Size</td>
<td>Location</td>
<td>Donor</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>---------</td>
<td>--------</td>
<td>-------------</td>
<td>----------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>MTJ-244</td>
<td>PTH</td>
<td>Oenothera parodiana ssp parodiana</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius at stony, grazed hills 2 km W of La Pileta in the Sierra de la Ventana, 350-400 m, Buenos Aires, Argentina</td>
<td>M. Johnson (donated by OPGC [NSSL 115699], who received it from W. Stubbe, 1981)</td>
<td>Santarius 469; Santarius (1968/69) Oenothera-Sammelreise nach Südamerika. Jahrbuch der Heinrich-Heine-Universität Düsseldorf; Ann Mo Bot Gard (1977) 64: 425-629</td>
</tr>
<tr>
<td>MTJ-266</td>
<td>PTH</td>
<td>Oenothera ravenii ssp argentinae</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Durazno, Córdoba, Argentina</td>
<td>M. Johnson (donated by OPGC [NSSL 165765], who received it from W. Stubbe, 1982)</td>
<td>Dietrich 107</td>
</tr>
<tr>
<td>MTJ-289</td>
<td><em>Oenothera stricta</em> ssp <em>argentinae</em></td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius east end of Sierra del Volcan, north west of Puerta El Abra, at km 45 of Ruta 226 between Mar de la Plata and Balcarce, at partly grazed terraces between rocks, 150-300 m, 7 Jan 1968 (MO-2155215, holotype; CTES, DUSS, M, isotypes)</td>
<td>Santarius 358; K.A. Santarius (1968/69) <em>Oenothera-Sammelreise nach Südamerika.</em> Jahrbuch der Heinrich-Heine-Universität Düsseldorf; Ann Mo Bot Gard (1977) 64: 425–629</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-298</td>
<td><em>Oenothera tarijensis</em></td>
<td>PTH</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by K.A. Santarius between km 38 and 36 at the road from Villazón to Tarija, 3,200-3,300 m, Tarija, Bolivia, 19 Mar 1968 (MO-2155393, holotype; CTES, DUSS, M, isotypes)</td>
<td>From propagated seed grown in greenhouse in 2007; Santarius 1924; K.A. Santarius (1968/69) <em>Oenothera-Sammelreise nach Südamerika.</em> Jahrbuch der Heinrich-Heine-Universität Düsseldorf; Ann Mo Bot Gard (1977) 64: 425–629</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTJ-300</td>
<td>Oenothera tetrapetra</td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>UK, from Kew B.G.</td>
<td>M. Johnson (donated by OPGC [NSSL 197332], who received it from P. Lapinskas, Hurst Corp. and Kew B.G., 1985)</td>
<td>From propagated seed grown in greenhouse in 2012</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
<td>-----</td>
<td>---------</td>
<td>----------------------------</td>
<td>------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>MTJ-305</td>
<td>Oenothera versicolor</td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>La Paz, Poerto de Transito, Bolivia</td>
<td>M. Johnson (donated by OPGC [NSSL 165729], who received it from W. Stubbe, 1982)</td>
<td>Solomon 5041</td>
</tr>
<tr>
<td>MTJ-356</td>
<td>Oenothera villosa ssp villosa</td>
<td>PTH</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Strain: bauri Standard; originally collected by R. Hölscher near Toruń, Kujawsko-Pomorskie, Poland, before 1944</td>
<td>M. Johnson (donated by J. Meurer/S. Greiner, originally from O. Renner)</td>
<td>Genome: AA-I; Cytology: ring of 14; Flora (1937) 31: 182-226; Flora (1944) 138: 57-92; Greiner seed id: 05/43 x s; inbred</td>
</tr>
<tr>
<td>MTJ-520</td>
<td>Oenothera longituba</td>
<td>SEX</td>
<td>diploid</td>
<td>Genome size; flow cytometry</td>
<td>Originally collected by L. Diers near La Banderita at the road from Andalgala to Concepcion, 1,800 m, Catamarca, Argentina, 11 Feb 1959 (MO-</td>
<td>M. Johnson (donated by S. Greiner, originally from W. Dietrich)</td>
<td>Ann Mo Bot Gard (1977) 64: 425–629; Greiner seed id: 86/2014 x s; inbred, some generations</td>
</tr>
</tbody>
</table>
Supplementary Table  6.2 Genome size estimates in *Oenothera*

<table>
<thead>
<tr>
<th>Genus</th>
<th>species</th>
<th>DNA 1 pg/2C</th>
<th>DNA 2 pg/2C</th>
<th>DNA 3 pg/2C</th>
<th>DNA mean pg/2C</th>
<th>DNA mean pg/1C</th>
<th>Previously published estimates pg/1C (if available)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oenothera</em></td>
<td>affinis</td>
<td>1.44</td>
<td>1.43</td>
<td>1.41</td>
<td>1.43</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Oenothera</em></td>
<td>albicaulis</td>
<td>1.31</td>
<td>1.30</td>
<td>1.30</td>
<td>1.30</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Oenothera</em></td>
<td>bahia-blancaae</td>
<td>1.41</td>
<td>1.43</td>
<td>1.42</td>
<td>1.42</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>biennis</td>
<td>2.21</td>
<td>2.20</td>
<td>2.19</td>
<td>2.20</td>
<td>1.10</td>
<td>1.15; 1.22</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>capillifolia (=O. berlandieri)</td>
<td>1.67</td>
<td>1.67</td>
<td>1.69</td>
<td>1.68</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>clelandii</td>
<td>1.35</td>
<td>1.36</td>
<td>1.34</td>
<td>1.35</td>
<td>0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>elata cf. ssp hookeri</td>
<td>2.21</td>
<td>2.20</td>
<td>2.18</td>
<td>2.20</td>
<td>1.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>flava</td>
<td>1.58</td>
<td>1.57</td>
<td>1.58</td>
<td>1.58</td>
<td>0.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>fruticosa ssp fruticosa</td>
<td>6.09*</td>
<td>6.09*</td>
<td>6.05*</td>
<td>6.08*</td>
<td>3.04</td>
<td>1.45</td>
<td></td>
</tr>
<tr>
<td>Oenothera</td>
<td>glazioviana</td>
<td>2.25</td>
<td>2.20</td>
<td>2.25</td>
<td>2.23</td>
<td>1.12</td>
<td>1.15; 1.19</td>
<td></td>
</tr>
</tbody>
</table>

Kubešová et al. (2010) Preslia 82: 81-96; Zonneveld et al. (2005) Ann Bot 96:229-244 (recorded as O. lamarckiana)
<table>
<thead>
<tr>
<th>Oenothera</th>
<th>heterophylla</th>
<th>1,95</th>
<th>1,92</th>
<th>1,93</th>
<th>1,93</th>
<th>0,97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oenothera</td>
<td>kunthiana</td>
<td>2,30</td>
<td>2,31</td>
<td>2,30</td>
<td>2,30</td>
<td>1,15</td>
</tr>
<tr>
<td>Oenothera</td>
<td>laciniata</td>
<td>1,34</td>
<td>1,35</td>
<td>1,34</td>
<td>1,34</td>
<td>0,67</td>
</tr>
<tr>
<td>Oenothera</td>
<td>longissima</td>
<td>2,21</td>
<td>2,22</td>
<td>2,23</td>
<td>2,22</td>
<td>1,11</td>
</tr>
<tr>
<td>Oenothera</td>
<td>longituba</td>
<td>2,18</td>
<td>2,21</td>
<td>2,18</td>
<td>2,19</td>
<td>1,10</td>
</tr>
<tr>
<td>Oenothera</td>
<td>mendocinensis</td>
<td>1,27</td>
<td>1,27</td>
<td>1,27</td>
<td>1,27</td>
<td>0,64</td>
</tr>
<tr>
<td>Oenothera</td>
<td>nana</td>
<td>1,58</td>
<td>1,57</td>
<td>1,58</td>
<td>1,58</td>
<td>0,79</td>
</tr>
<tr>
<td>Oenothera</td>
<td>nutans</td>
<td>2,32</td>
<td>2,31</td>
<td>2,34</td>
<td>2,32</td>
<td>1,16</td>
</tr>
<tr>
<td>Oenothera</td>
<td>parodiana ssp parodiana</td>
<td>1,41</td>
<td>1,40</td>
<td>1,39</td>
<td>1,40</td>
<td>0,70</td>
</tr>
<tr>
<td>Oenothera</td>
<td>perennis</td>
<td>1,90</td>
<td>1,90</td>
<td>1,91</td>
<td>1,90</td>
<td>0,95</td>
</tr>
<tr>
<td>Oenothera</td>
<td>ravenii ssp argentinae</td>
<td>1,44</td>
<td>1,45</td>
<td>1,46</td>
<td>1,45</td>
<td>0,73</td>
</tr>
<tr>
<td>Oenothera</td>
<td>rhombipetala</td>
<td>2,14</td>
<td>2,13</td>
<td>2,12</td>
<td>2,13</td>
<td>1,07</td>
</tr>
<tr>
<td>Oenothera</td>
<td>rosea</td>
<td>1,58</td>
<td>1,58</td>
<td>1,56</td>
<td>1,57</td>
<td>0,79</td>
</tr>
<tr>
<td>Oenothera</td>
<td>sandiana (=O. rubida)</td>
<td>1,29</td>
<td>1,28</td>
<td>1,29</td>
<td>1,29</td>
<td>0,64</td>
</tr>
<tr>
<td>Oenothera</td>
<td>stricta ssp argentinae</td>
<td>1,44</td>
<td>1,46</td>
<td>1,44</td>
<td>1,45</td>
<td>0,72</td>
</tr>
<tr>
<td>Oenothera</td>
<td>tarijensis</td>
<td>1,34</td>
<td>1,36</td>
<td>1,36</td>
<td>1,35</td>
<td>0,68</td>
</tr>
<tr>
<td>Oenothera</td>
<td>tetraptera</td>
<td>1,21</td>
<td>1,21</td>
<td>1,21</td>
<td>1,21</td>
<td>0,61</td>
</tr>
<tr>
<td>Oenothera</td>
<td>versicolor</td>
<td>1,43</td>
<td>1,46</td>
<td>1,45</td>
<td>1,45</td>
<td>0,72</td>
</tr>
<tr>
<td>Oenothera</td>
<td>villaricae</td>
<td>1,39</td>
<td>1,40</td>
<td>1,39</td>
<td>1,39</td>
<td>0,70</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Oenothera</td>
<td>villosa ssp villosa</td>
<td>2,20</td>
<td>2,19</td>
<td>2,17</td>
<td>2,19</td>
<td>1,09</td>
</tr>
</tbody>
</table>

* The ploidy level for *O. fruticosa* is unknown. Reported values are the DNA content in the nucleus.
CHAPTER 7 CHROMOSOMAL DISTRIBUTION OF CYTO-NUCLEAR GENES IN A DIOECIOUS PLANT WITH SEX CHROMOSOMES

Abstract

The coordination between nuclear and organellar genes is essential to many aspects of eukaryotic life, including basic metabolism, energy production, and ultimately, organismal fitness. Whereas nuclear genes are bi-parentally inherited, mitochondrial and chloroplast genes are almost exclusively maternally inherited, and this asymmetry may lead to a bias in the chromosomal distribution of nuclear genes whose products act in the mitochondria or chloroplasts. In particular, because X-linked genes have a higher probability of co-transmission with organellar genes (2/3) compared to autosomal genes (1/2), selection for co-adaptation has been predicted to lead to an over-representation of nuclear-mitochondrial and nuclear-chloroplast genes on the X chromosome relative to autosomes. In contrast, the occurrence of sexually antagonistic organellar mutations might lead to selection for movement of cyto-nuclear genes from the X chromosome to autosomes to reduce male mutation load. Recent broad-scale comparative studies of N-mt distributions in animals have found evidence for these hypotheses in some species, but not others. Here, we use transcriptome sequences to conduct the first study of the chromosomal distribution of cyto-nuclear interacting genes in a plant species with sex chromosomes (*Rumex hastatulus*; Polygonaceae). We found no evidence of under- or over-representation of either N-mt or N-cp genes on the X chromosome, and thus no support for either the co-adaptation or the sexual-conflict hypothesis. We discuss how our results from a species with recently evolved sex chromosomes fit into an emerging picture of the evolutionary forces governing the chromosomal distribution of nuclear-mitochondrial and nuclear-chloroplast genes.
7.1 Introduction

The intimate relationships between nuclear and organellar genomes in eukaryotes represent some of the most striking examples of co-evolved mutualisms (Gillham 1994; Lane 2005; Aanen et al. 2014). The long co-evolutionary history of nuclear and mitochondrial genomes is perhaps best illustrated by the finding that the vast majority of mitochondrial genes in animals have been transferred to the nuclear genome (Adams and Palmer 2003; Rand et al. 2004; Burt and Trivers 2006). Indeed, animal mitochondria now encode only a few proteins after having lost the majority of their original genes (Berg and Kurland 2000; Ridley 2000; Bar-Yaacov et al. 2012). Moreover, almost one fifth of the Arabidopsis thaliana nuclear genome is of chloroplast origin (Martin 2003), suggesting that organellar-to-nuclear gene movement has played a crucial role in the evolution of plant genetic systems.

The evolution of cyto-nuclear interactions and the chromosomal distribution of the genes involved should be influenced by the contrasting modes of inheritance of organellar genes (maternal inheritance) and autosomal genes (bi-parental inheritance). This difference may, for example, result in conflict between nuclear and organellar genes over sex determination and sex ratio (Cosmides and Tooby 1981; Werren and Beukeboom 1998), and several mitochondrial genes in plants are known to cause male sterility (Burt and Trivers 2006; Touzet and Meyer 2014). In systems with XY sex determination, where males are the heterogametic (XY) and females the homogametic sex (XX), genes on the X chromosome spend 2/3 of their time in females (Rand et al. 2001) and therefore share a female-biased inheritance pattern relative to Y-linked or autosomal genes, which may result in inter-genomic co-adaptation or conflict.
A potential consequence of inter-genomic conflict or co-adaptation between nuclear genes, whose products interact with mitochondrial or chloroplasts (mito-nuclear and cyto-nuclear genes, respectively) and other regions of the genome, is a shift in the chromosomal location of such genes, either becoming more or less abundant on the X chromosome. Several molecular mechanisms have been suggested to be involved in driving gene movement, including gene duplication followed by fixation and subsequent gene loss (Wu and Yujun Xu 2003), and autosomal gene duplications followed by the evolution of sex biased gene expression (Connallon and Clark 2011). The evolutionary mechanisms of this gene movement have also been explored by several recent studies (Drown et al. 2012; Hill and Johnson 2013; Dean et al. 2014; Rogell et al. 2014), and two main processes have been proposed to account for the movement of genes to or from the X chromosome. The co-adaptation hypothesis predicts that the co-transmission of X-linked and organellar genes should result in their co-adaptation, in which selection on beneficial epistatic interactions results in an over-representation of cyto-nuclear genes on the X chromosome relative to autosomes (Rand et al. 2004; Drown et al. 2012). In contrast, the sexual conflict hypothesis predicts the opposite chromosomal distribution, with more cyto-nuclear genes occurring on autosomes to alleviate mutation load in males. To date, empirical evidence for these hypotheses are mixed. Drown et al. (2012) used previously published reference genomes to examine the chromosomal distribution of N-mt genes in 16 vertebrates and found a strong under-representation of such genes on the X chromosomes relative to autosomes in 14 mammal species, but not in two avian species with ZW sex determining systems; note that the co-adaptation hypothesis does not predict that ZW systems should show a bias in the distribution of cyto-nuclear genes. Dean et al. (2014) included seven additional species in their analysis with independently derived sex chromosomes and found that the under-representation of N-mt genes on the X chromosome was restricted to therian mammals and Caenorhabditis elegans.
Here, we use sex-linked and autosomal transcriptome sequences to investigate the chromosomal distributions of cyto-nuclear interactions in the dioecious annual plant *Rumex hastatulus* (Polygonaceae). Examining cyto-nuclear interactions within a plant species is of interest for several reasons (see Sloan 2014). First, plants carry an additional maternally inherited organellar genome that is absent in animals, the chloroplast genome. This provides an opportunity to compare the chromosomal distribution of two independent kinds of cyto-nuclear interacting genes: nuclear-mitochondrial and nuclear-chloroplast. Second, whereas animal sex chromosomes evolved hundreds of millions of years ago (180 MYA in mammals and 140 MYA in birds; Cortez et al. 2014), the origin of plant sex chromosomes is a more recent event (Charlesworth 2013). In *R. hastatulus*, sex chromosomes are thought to have evolved approximately 15-16 MYA (Navajas-Perez et al. 2005) and genes on the Y chromosome show evidence of degeneration, resulting in a considerable proportion of genes that are hemizygous on the X chromosome (Hough et al. 2014). *Rumex hastatulus* therefore provides an opportunity to test whether the early changes involved in sex chromosome evolution are associated with a concomitant shift in the chromosomal location of N-mt or N-cp genes. Moreover, the presence in this system of X-linked genes that have recently become hemizygous provides an opportunity to compare the chromosomal distributions of X-linked genes that are hemizygous versus those that have retained Y-linked alleles (X/Y genes). Hemizygous genes are particularly good candidates for evaluating evidence for co-adaptation and/or sexual conflict because of their relatively older age (Hough et al. 2014), and because beneficial mutations in such genes are exposed to positive selection regardless of dominance and may therefore spread more rapidly.
7.2 Methods

7.2.1 Gene identification and functional annotation

We used sex-linked and autosomal transcriptome sequence data for *R. hastatulus* reported in Hough et al. (2014; GenBank Sequence Read Archive accession no. SRP041588), and obtained three sets of genes with which to test for an over- or under-representation of nuclear-mitochondrial or nuclear-chloroplast genes. In total our analyses included 1167 autosomal genes, 624 X-linked genes, and 107 hemizygous X-linked genes. The X-linked and hemizygous X-linked genes were shared between sex chromosome systems in this species (see Methods and Supplementary Material of Hough et al. 2014 for full details regarding the identification of such genes from transcriptome sequence data). For autosomal genes, we included those previously identified as confidently autosomal in both *R. hastatulus* sex chromosome systems, as well as those uniquely identified in the XYY system. For each gene set, we queried the sequences translated in all reading frames against the *A. thaliana* protein database using the BLASTx homology search implemented in Blast2GO (Conesa et al. 2005), with a significance threshold (BLAST ExpectValue) of $1 \times 10^{-3}$, above which matches were not reported. We limited our searches to the *A. thaliana* protein database because sequence matches to this database returned more detailed functional information than is available for most other species in the NCBI plant database. We obtained BLASTx results for 1073 autosomal genes (90%), 567 X-linked genes (90%), and 95 hemizygous genes (89%). Gene Ontology (GO) terms associated with the hits from BLASTx queries were then retrieved using the ‘Mapping’ function in Blast2GO, which used BLAST accessions to link the queried sequences to functional information stored in the GO database (The Gene Ontology Consortium 2008). Gene names were retrieved using NCBI mapping files ‘gene info’ and ‘gene2accession’, and GO terms were assigned to query sequences.
using the ‘Annotation’ function with an E-Value-Hit-Filter of $1 \times 10^{-6}$ and an annotation cut off of 55 (default parameters). Finally, we ran InterProScan (Quevillon et al. 2005) to retrieve sequence domain/motif information and merged the corresponding annotations with previously identified GO terms. This procedure generated output files containing GO ID’s and functional descriptions for each gene in our data set (files will be uploaded to GitHub). The numbers of genes in our final data set with functional annotations and N-mt and N-cp GO annotations are summarized in Table 7.1.

**Table 7.1 Number of genes in Rumex data set**

Data includes genes for which we obtained functional annotations (see Methods) and those with nuclear-mitochondrial and nuclear-chloroplast GO annotations.

<table>
<thead>
<tr>
<th>Gene sets</th>
<th>Autosomal</th>
<th>X-linked</th>
<th>X-hemizygous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data set</td>
<td>1167</td>
<td>624</td>
<td>107</td>
</tr>
<tr>
<td>With annotation</td>
<td>1073</td>
<td>567</td>
<td>95</td>
</tr>
<tr>
<td>With N-mt GO ID</td>
<td>194</td>
<td>94</td>
<td>13</td>
</tr>
<tr>
<td>With N-cp GO ID</td>
<td>222</td>
<td>102</td>
<td>22</td>
</tr>
</tbody>
</table>

7.2.2 Statistical analyses

We used a similar approach to Drown et al. (2012) and Dean et al. (2014) and estimated the number of N-mt and N-cp genes on the X chromosome and autosomes, and then compared each of these estimates to an expected number. The expected number of N-mt genes was obtained by calculating the product of the proportion of all genes in the data set with mitochondrial annotations (matching GO:0005739) and the number of annotated genes in a given gene set. The expected numbers of N-cp genes were calculated similarly, using GO:0009507. We then
calculated the ratios of the observed-to-expected numbers for both N-mt and N-cp genes in each gene set. The observed-to-expected ratio is expected to equal one when there is no under- or over-representation, and greater than one when there is an over-representation. We note that, unlike for X-linked genes, we did not have information regarding the particular chromosome locations for autosomal genes, and therefore could not obtain the expected numbers of N-mt and N-cp genes per-autosome as in previous studies (Drown et al. 2012; Dean et al. 2014). The expected numbers were thus calculated assuming that the set of autosomal genes represented a random sample of the autosomal chromosomes in this species, which is likely a valid assumption given that the sequences were obtained using whole transcriptome shotgun sequencing (Hough et al. 2014). Calculating the expected-to-observed ratios across X-linked, autosomal, and X-hemizygous genes thus allowed us to determine whether any of these gene sets contained an under- or over-representation of N-mt and N-cp genes compared to the expectation based on the proportion of such genes in the full data set. We tested the significance of over- or under-representation using Fisher’s exact tests, and calculated 95% confidence intervals for the numbers of N-mt or N-cp genes using 10,000 replicate bootstrapped samples. Given our sample sizes of genes with annotations (Table 7.1), Fisher’s Exact Tests allowed us to test for differences in the proportions of cyto-nuclear genes on autosomes versus the X-chromosome that were on the order of 5% with ~80% power, whereas power was reduced for smaller differences (Supplementary Material). Similarly, for hemizygous X-linked genes, we calculate that differences of approximately ~10% could be detected with ~80% power. All data analysis was done in R (R Development Core Team 2013; scripts are download from GitHub).
7.3 Results and Discussion

It has been suggested that cyto-nuclear genes may be either over- or under-represented on the X chromosome compared to autosomes, depending on whether their interactions are driven by co-adaptation or sexual conflict (Rand et al. 2001; Drown et al. 2012; Hill and Johnson 2013; Dean et al. 2014; Rogell et al. 2014). We annotated sex-linked and autosomal transcriptome sequences to test these predictions in the dioecious plant *R. hastatulus*. We found that neither mitochondria- or chloroplast-interacting nuclear genes were under- or over-represented on the X chromosome (Fisher’s exact test, $P = 0.4947$ and $P = 0.3074$, respectively; Figure 7.1). This pattern indicates that neither the co-adaptation nor the sexual conflict hypothesis alone is sufficient to explain the chromosomal distribution of cyto-nuclear genes in *R. hastatulus*.

![Figure 7.1](image)

**Figure 7.1 Representation of the chromosomal location of cyto-nuclear genes in Rumex hastatulus**

Dots represent the observed to expected ratio of mito-nuclear (N-mt) and chloro-nuclear (N-cp) genes on autosomes, the X chromosome, and hemizygous X genes, with the 95% confidence intervals estimated by bootstrapping (10,000 replicates). The vertical dotted line at 1 represents no over- or under-representation
There are several factors that are expected to be important in determining cyto-nuclear gene distributions, and these may explain the lack of bias in *R. hastatulus*. For example, under both the co-adaptation and sexual conflict hypotheses, the age of the sex chromosomes will determine the extent to which selection (either for co-adaptation, or sexual antagonism) has had time to operate, which depends on the rate of gene movement onto and off of the sex chromosomes. Whereas previous studies of cyto-nuclear genes in animals have focused almost exclusively on ancient sex chromosome systems (Drown et al. 2012; Dean et al. 2014; Rogell et al. 2014), our study focused on a dioecious plant species in which sex chromosomes evolved more recently (~15 MYA; Navajas-Perez et al. 2005), and many genes likely stopped recombining much more recently (Hough et al., 2014). The lack of bias in the chromosomal distribution of cyto-nuclear genes may therefore reflect the recent time scale of sex chromosome evolution rather than the absence of biased gene movement. The relatively young age of sex chromosomes may also have played a role in the lack of bias reported in the sex and neo-sex chromosomes in three-spined stickleback, which evolved ~10 MYA (Kondo et al. 2004) and ~2 MYA, respectively (Natri et al. 2013). Comparative studies of sex chromosomes of different age will be central for understanding the rate at which organellar gene movement occurs.

In addition to being evolutionarily older, X-linked hemizygous genes are expected to show a greater effect of over-or under-representation than genes with both X- and Y-alleles because recessive mutations (involved in either co-adaptation or sexual conflict) will be exposed to selection instead of masked by an alternate allele in a heterozygous genotype. We detected a slightly greater under-representation of X-hemizygous N-nt genes compared to autosomes or X-genes with retained Y-alleles, but the effect was not statistically significant (*P* = 0.4947). The opposite pattern was evident for N-cp genes, which were slightly over-represented on hemizygous genes, but again this effect was not significant (*P* = 0.3074). A larger sample of
hemizygous genes would be required to more confidently assess whether such genes are in fact more often involved in cyto-nuclear interactions than other genes on the X chromosome, and to test whether the opposite pattern for N-mt and N-cp hemizygous genes is a result of a different rate of nuclear gene transfer between mitochondrial and chloroplast genomes. In particular, the smaller number of hemizygous X-linked genes in our data set implies that power was reduced for this comparison, such that a ~5% difference could only be detected with ~60% power (see Supplementary Material 7.1).

Another factor that will affect the chromosomal distribution of cyto-nuclear genes is the number of N-mt and N-cp genes that were located on the autosome from which the sex chromosomes evolved. Since the origins of mitochondria and chloroplasts both vastly predate that of sex chromosomes (1.5-2 BYA compared to < 200 MYA; Dyall et al. 2004; Timmis et al. 2004; Cortez et al. 2014), gene transfer from organellar genomes to the nuclear genome began long before the evolution of sex chromosomes. A bias in the chromosomal distribution of cyto-nuclear genes in either direction may therefore arise if the ancestral autosome was particularly rich or poor in cyto-nuclear genes. Indeed, it is striking that autosomes in the animal species previously examined exhibited extensive variation in the relative number of N-mt genes (see Drown et al. 2012 Figure 1 and Dean et al. 2014 Figures 1 and 2). That the ancestral number of N-mt and N-cp genes is likely to be important is highlighted by the fact that the majority of genes involved in mitochondrial DNA and RNA metabolism in A. thaliana are found on chromosome III (Elo et al. 2003). If such a biased autosomal distribution of organellar variation is representative of the ancestral sex chromosomes, the X chromosome could carry significantly more N-mt or N-cp genes because of this ancestral gene number rather than a biased rate of gene movement. This effect is likely exacerbated in early sex chromosome systems, where the majority of genes may not have experienced opportunities for movement. Genetic mapping and
comparative genomic studies of genes that have transferred from organellar genomes after the origin of sex chromosomes may provide a means to control for ancestral differences in gene number and provide a better test of biases in organellar-nuclear gene movement.

To conclude, our study is the first investigation of the extent to which co-adaptation and sexual conflict have shaped the chromosomal distribution of cyto-nuclear genes in a plant species with sex chromosomes. We found no sign of under- or over-representation of either N-nt or N-cp genes on the X chromosome, implying that neither co-adaptation nor sexual conflict alone can explain the chromosomal distributions of these genes. Instead, we suggest that additional factors, including the age of sex chromosomes and the time that has elapsed since X-Y recombination became suppressed, are likely to have been important determinants of the patterns we observed. To determine whether the lack of under-representation of mito-nuclear genes on the X chromosome reflects an absence of gene movement, future studies should focus on quantifying rates of gene movement after sex chromosome origination, and consider the extent to which neutral processes including the number of mito-nuclear genes on ancestral sex chromosomes have played an important role in shaping the current chromosomal distributions of such genes. Cyto-nuclear conflict and co-evolution have undoubtedly played a major role in many aspects of genome evolution in both plant and animal systems, and the previously reported evidence from therian mammals and *C. elegans* (Drown et al. 2012; Dean et al. 2014) suggests that sexual conflict and co-adaptation might represent important mechanisms driving chromosomal gene movement; however, it remains unclear whether these processes have also shaped the chromosomal distribution cyto-nuclear genes in plants.
7.4 Acknowledgements

We thank Rebecca Dean and Devin M Drown for comments on the manuscript. This research was supported by Discovery Grants to SCHB and SIW from the Natural Sciences and Engineering Council of Canada. JH was supported by an Ontario Graduate Fellowship and JAÅ by a Junior Fellowship from Massey College.

7.5 Supplementary Material

7.5.1 Power Analysis

To determine the extent to which a biased distribution of cyto-nuclear genes could be detected given our sample sizes of annotated autosomal, X-linked, and X-hemizygous genes, we calculated the power to detect significant differences based on a Fisher’s Exact Test. Here, power refers to the probability of correctly rejecting the null hypothesis of no difference in the proportion of cyto-nuclear genes among the gene sets, and we used the hypergeometric distribution to calculate the probability of getting the observed data under the null hypothesis that the proportions were the same (with an alpha significance level of 0.05).

To better visualize the difference in power for the two main comparisons of interest (autosomal genes vs. X/Y genes, and autosomal genes vs. X-hemizygous genes), the figure below shows an example in which the true proportion of cyto-nuclear genes on autosomes is assumed to be 0.25 (which is approximately the empirical proportion in our data). Power is then shown as a function of the true proportion of cyto-nuclear genes on the X-chromosome, ranging from 0.1 to 0.5. As discussed in the main text, our sample sizes of annotated X/Y genes (n=567;
see Table 7.1 in main text) were large enough to detect differences between autosomes and X/Y genes with ~80% power given a true difference of ~5%, and as this difference becomes smaller, the power decreases (Supplementary Figure 7.1). Similarly, for hemizygous X-linked genes (with n=95), differences of approximately~10% could be detected with ~80% power, and power was reduced for smaller differences. This analysis was done using G*Power (Faul et al. 2007) and R (R Development Core Team 2013).

Supplementary Figure 7.1 Power analysis of chromosomal distribution
Power to detect a significant difference in the proportion of cytoplasmic genes between autosomes and sex chromosomes as a function of the true proportion on sex chromosomes. Power was calculated assuming the true proportion of cytoplasmic genes on autosomes was 25%.
CHAPTER 8 EVOLUTIONARY TRANSITIONS IN INDIVIDUALITY: INSIGHTS FROM TRANSPOSABLE ELEMENTS

Summary

The history of life has been characterized by evolutionary transitions in individuality, the grouping together of independently replicating units into new larger wholes: genes to chromosomes, chromosomes in genomes, up to three genomes in cells, cells in multicellular organisms that form groups and societies. Central to understanding these transitions is to determine what prevents selfish behaviour at lower levels from disrupting the functionality of higher levels. Here, I review work on transposable elements, a common source of disruption at the genome level, in the light of the evolutionary transitions framework, and argue that the rapid influx of data on transposons from whole genome sequencing has created a rich data source to incorporate into the study of evolutionary transitions in individuality.

Glossary

*Hypercycle* A network for self-replicating units, where the product of a given reaction acts as a catalyst for the next in a cyclical manner. For example, in the hypercycle of members A, B, and C, A catalyses B, B catalyses C, and C catalyses A.

*Genomic conflict* Conflict in the fitness interests between genes in the same genome.

*Linkage disequilibrium* The non-random association of alleles at two or more loci.

*Modifier* A gene that evolved to oppose the effect of a selfish genetic element.
**Selfish genetic element** Genes that have the ability to promote their own transmission at the expense of other (unlinked) genes in the same individual.

**Silencing** The interference with the expression of transposable elements, often resulting in a reduction of the transposition rate to zero. Silencing is often administrated by small interfering RNAs.

**Small interfering RNAs** Short (20-25 base pairs) double stranded RNA molecules that interferes with expression of certain genes with complementary nucleotide sequence. siRNAs play an important role in the silencing of transposable elements.

**Social evolution** The study of the evolution of social behaviour. Traditionally restricted to animal societies, but its expanded version deals with all levels in the hierarchy of life.

**Transposable element** Stretches of DNA that can move to new locations in the genome, either through a “copy-and-paste” mechanism via an RNA intermediate (Class 1 retrotransposons) or via a cut-and-paste mechanism (Class 2 DNA transposons).

### 8.1 Conflict and cooperation across the hierarchy of life

While the study of social evolution has long focused on the baboon troop, the beehive, and more recently that of social microbes, the majority of cooperation occurs within organisms (Queller 1997; Michod 1999; Bourke 2011; Queller and Strassmann 2009). Genes come together in chromosomes, which make up genomes, and several genomes coordinate efforts in the cells that make up multicellular organisms, which in turn often form tightly knitted social groups.
(Maynard Smith and Szathmáry 1995; Michod 1999; Okasha 2006; Bourke 2011). The coming
together of previously independently reproducing units into new larger wholes is what gives life
its hierarchical organization (Buss 1987; Maynard Smith and Szathmáry 1995; Michod 1999;
Calcott and Sterelny 2011; Bouchard and Huneman 2013). Inspired by the research programme
on major transitions in evolution, the points at which one level of the hierarchy becomes
subsumed into another have become known as evolutionary transitions in individuality (Box 8.1;
Table 8.1). The functionality of a given level in the hierarchy is contingent on the suppression of
conflict at lower levels (Leigh 1977; Alexander and Borgia 1978; Buss 1987; Maynard Smith
and Szathmáry 1995; Bourke 2011; Foster 2011). Under this view of life, the problem of
evolution is therefore to understand the mechanisms that make cooperation thrive over conflict at
a given level.

**Box 8.1 Major transitions in evolution and individuality**

Not all major transitions in evolution are evolutionary transitions in individuality. In their
landmark publication, Maynard Smith and Szathmáry (1995) presented eight major transitions
(Table 8.1). In their opinion, what these events had in common was that they changed how the
evolutionary process itself operates. More specifically, most of them share three features. First,
following the transitions, units that previously reproduced independently can now only reproduce
as part of a larger whole. Second, a division of labour characterizes the new units. Third, the
transitions change the language of information and the way information is stored and transmitted.
However, these features do not apply to all transitions. For example, the evolution of the genetic
code is a good example of division of labour and a change in the language, storage, and
transmission of information, but does not involve the coming together of previously
independently reproducing entities. For it to be useful to study transitions as a group, we need conceptual unity (Michod 1999; McShea and Simpson 2011). Without conceptual unity, they may as well be grouped together under the heading “a list of interesting and important events in the history of life”.

Focusing on transitions in individuality offers one route to such unity. Evolutionary transitions in individuality share two main themes (Michod 1999; Bourke 2011). To start, just like the first feature identified by Maynard Smith and Szathmáry (1995), transitions involve the emergence of cooperation among independent units leading to the formation of a new higher-level unit. Second, crucial to the functioning of this new unit, the transitions are associated with the evolution of mechanisms to suppress conflict among lower-level units. This approach typically removes the genetic code, sex and language from the list of transitions, but we may add interspecific mutualisms to the list.
Table 8.1 The major transitions in evolution and individuality

<table>
<thead>
<tr>
<th>Event</th>
<th>Major transition sensu Maynard Smith and Szathmáry (1995)</th>
<th>Evolutionary transition in individuality</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>Replicating molecules(^a)</td>
<td>Populations of molecules in compartments</td>
<td>Yes</td>
</tr>
<tr>
<td>Independent replicators(^a)</td>
<td>Chromosomes</td>
<td>Yes</td>
</tr>
<tr>
<td>RNA as gene and enzyme</td>
<td>DNA + protein (genetic code)</td>
<td>Yes</td>
</tr>
<tr>
<td>Prokaryotes</td>
<td>Eukaryotes</td>
<td>Yes</td>
</tr>
<tr>
<td>Asexual clones</td>
<td>Sexual populations</td>
<td>Yes</td>
</tr>
<tr>
<td>Protists</td>
<td>Animals, plants, fungi (cell differentiation)</td>
<td>Yes</td>
</tr>
<tr>
<td>Solitary individuals</td>
<td>Colonies (non-reproductive castes)</td>
<td>Yes</td>
</tr>
<tr>
<td>Primate societies</td>
<td>Human societies (language)</td>
<td>Yes</td>
</tr>
<tr>
<td>Independent species</td>
<td>Interspecific mutualism</td>
<td>No</td>
</tr>
</tbody>
</table>

\(^a\) These are sometimes group together as one transition.
Each transition in individuality can be divided into three stages (Figure 8.1; Bourke 2011). First, social group formation is the origin and spread of social behaviour in a population. Second, social group maintenance concerns the processes that allow the social group to persist in a stable way. Third, social group transformation is the origin of a new level of individuality. The origin of the genome from independently replicating molecules was not only the first transition, but the genome can also be said to be the most fundamental level of cooperation. Whereas many organisms thrive without the benefits of multicellularity or eusociality, cooperation among genes occurs in all organisms.

The idea that the same principles of social evolution theory can be applied to all transitions in individuality has been around for some time (Leigh 1991; Maynard Smith and Szathmáry 1995; Michod 1999; Queller 2000; Bourke 2011). However, since almost all of these transitions occurred in the distant past, empirical work has often been limited to the later transitions. The origin and maintenance of cooperation among individual organisms has perhaps received the most attention. Pioneering workers in the field concerned themselves primarily with the social lives of animals (Gardner and Foster 2008). Over the last years, many researchers have fruitfully applied the same approaches to the study of interspecific mutualisms (West et al. 2007a) and recent transitions from uni- to multicellularity (Grosberg and Strathmann 2007). Pioneering theoretical work on the transition from genes to genomes was done by Eigen and Schuster through their work on the theory of hypercycles (Eigen and Schuster 1977; 1978a; 1978b; see Glossary) to explain the origin of networks of interacting genes. This approach was later thoroughly extended by others using the conceptual framework of major transitions (e.g. Maynard Smith and Szathmáry 1995; Michod 1999; Szathmáry 2006). Similarly, the empirical study of conflict and cooperation among genes has a long history and stretches back over 100 years (see Burt and Trivers (2006) for a historical overview). Again, while these studies were not
always interpreted within the framework of social evolution, it has long been clear that this can be done (Maynard Smith and Szathmáry 1995; Haig 1997; Ridley 2001; Bourke 2011).

Figure 8.1 The evolutionary transition to the genome

The three stages of an evolutionary transition in individuality (Bourke 2011), as applied to the origin of the genome. First, social group formation is the origin and spread of social behaviour in a population. At the genome level this involved the first cooperative gene-networks and the beginning division of labour among genes. Second, social group maintenance are the processes that allow the social group to persist in a stable way. As explained in the main text, the stability of the genome is maintained through self-regulation and policing. Third, social group transformation is the origin of a new level of individuality. This involved further division of labour, including but not limited to, traits encoded my more than one gene, evolution of genes regulating expression of near (cis) and distant (trans) genes, as well as the evolution of chromosomes with sex-specific functions. Modified with permission from (Bourke 2011).

8.1.1 Empirical studies of evolutionary transitions

Recent years have seen a rapid influx of whole genome data thanks to advances in high throughput sequencing (reference genomes are now available from over 160 eukaryotic and over 2200 bacterial species (http://www.ebi.ac.uk/genomes/) and this has led to the suggestion that the wide-reaching effects that genomic conflict may have on genome evolution remain
underappreciated (Rice 2013). A particularly well-studied form of genomic conflict, and one where our understanding has benefited greatly from whole genome sequencing, is that between transposable elements and the rest of the genome (see below; Glossary; Figure 8.2).

Here, I review recent work on transposable elements in the light of evolutionary transition theory and argue that transposon data offer a rich empirical resource for the study of the second stage in the evolutionary transition in individuality, the maintenance of social groups. To that end, I first discuss the biology of transposable elements and theoretical underpinnings of mechanisms for social group maintenance. I then argue that the study of transposons may improve our understanding of two fundamental questions related to the maintenance of individuality. First, the observed reduction in transposon abundance in highly selfing and asexual lineages provide clues to the circumstances under which the fitness interests of lower and higher levels may align. Second, our improved understanding of evolutionary and ecological processes that affect the epigenetic control of transposon activity may help us understand what prevents selfish behaviour at a lower level from disrupting functionality at higher levels. Finally, I suggest that while the concept of evolutionary transitions provides a useful framework to interpret transposable element data, empirical insights from studies of transposons may in turn help improve the conceptual framework itself.

8.2 The biology of transposable elements

The assumption that genomes function as integrative cooperative units has a long tradition in biology (Mayr 1997; ENCODE Project Consortium 2012). However, conflict among genes is not only possible, but is a dominant feature in eukaryotic genomes (Hurst et al. 1996; Burt and
Trivers 2006). Selfish genetic elements are genes that have the ability to promote their own transmission at the expense of other genes in the genome (the genome here being all the genetic material of an organism; Cosmides and Tooby 1981). This results in a conflict and the spread of selfish genetic elements often leads to selection for modifiers, which are other genes in the same genome that have evolved ways to counter this spread (Werren et al. 1988; Werren 2011).

Genomic conflicts come in a remarkable diversity of flavours (Werren et al. 1988; Burt and Trivers 2006; Werren 2011; Hurst and Werren 2011). Arguably the most successful form of selfish genetic elements is transposable elements, which are mobile genetic elements that can self-replicate and produce copies of themselves that may insert elsewhere in the genome (Figure 8.2). This movement is often associated with mutational effects and therefore creates a conflict between transposons and the rest of the genome (Le Rouzic and Deceliere 2005). Yet, transposons are the most common kind of genetic material (Aziz et al. 2010) and make up more than 80% of the genomes in some species but are virtually absent in others (Kidwell 2002). Examining the evolutionary forces that allow transposons to proliferate so successfully in some species, but not in others, can therefore provide a window of an ongoing genomic conflict and its resolution.
Figure 8.2 Classes of transposable elements

Transposable elements are typically divided into two classes depending on their method of self-replication (Wicker et al. 2007). Class 1 elements, sometimes known as retrotransposons, apply a ‘copy-and-paste’ mechanism and produce an RNA intermediate that is then reverse-transcribed into DNA and integrated into a new location in the genome. However, Class 2 DNA elements replicate via a ‘cut-and-paste’ approach and the transposon is physically excised and integrated elsewhere in the genome. The movement of transposons has been suggested to be deleterious for three reasons (Dolgin et al. 2008; Blumenstiel 2011): (i) new transposon insertions end up in or near genes (Finnegan 1992); (ii) ectopic recombination between nonhomologous elements cause chromosomal rearrangements (Montgomery 1987); and/or (iii) metabolic costs imposed by the process of transposition itself (Brookfield 1991).

8.3 Mechanisms of social group maintenance

Despite the potential for widespread conflict introduced by transposons, the genome does not implode, which suggests some force(s) holding the group together. All levels of individuality are vulnerable to selection at lower levels favouring selfish behaviour (Michod 1999; Maynard Smith and Szathmáry 1995; Bourke 2011) and two routes to social group maintenance have
traditionally been considered (West et al. 2007b). First, if the fitness interests of lower and higher levels are aligned, self-restraint of selfish behaviour is expected to evolve, as the only way to maximize fitness at one level is to maximize it at the other level. Alternatively, cooperation may be promoted by the active suppression of conflict (Frank 2003; El Mouden et al. 2010). One such mechanism is the forcible prevention of selfish behaviour through policing. Here, the opportunity for cheating is removed and the only way to maximize individual fitness is by maximizing group fitness. Policing can enforce cooperation by aligning the fitness interest of interacting entities and so make the group act as one cohesive functional unit. Below, I expand on how these mechanisms may play out at the genome level and discuss how recent studies on transposons suggest the action of both self-restraint and policing in maintaining cooperation in the genome.

8.4 Self-restraint in the genome

Both theoretical and empirical work suggests that the mating system will play an important role in aligning the fitness interest of genes with those of higher levels of the organization. Highly selfing or asexual species are expected to experience a reduction in intragenomic conflict for several reasons (Hickey 1982; Wright and Schoen 1999; Morgan 2001; Nuzhdin and Petrov 2003). Considering the conflict between transposons and the rest of the genome, a particularly important difference between selfing and asexual lineages on the one hand and outcrossing lineages on the other is the extent of linkage disequilibrium in the genome. Linkage disequilibrium is the non-random association of alleles at two or more loci, i.e. a measure of how likely two genes are to be inherited together. This association may favour cooperation among these genes. It is thus conceptually analogous to an iterated prisoners dilemma and we may
therefore expect linked genes to show greater cooperation than unlinked ones (Cosmides and Tooby 1981; Haig 2003; Brown and Levin 2011). The reduction of linkage disequilibrium in outcrossing genomes means that the association between a given transposon and any deleterious mutational effects it may have is more likely to be decoupled by recombination. The transposon can therefore more easily spread in an outcrossing population, compared to a selfing or asexual population, where higher linkage disequilibrium leads to a stronger association between transposons and their deleterious effects. As a consequence, we may expect transposons showing self-restraint to be more common in selfing compared to outcrossing lineages (Charlesworth and Langley 1986).

Empirical support for the role of mating system in social group maintenance in the genome comes from a variety of systems. First, very few transposons are present in the ancient asexual bdelloid rotifers, but common in their sexual relatives (Arkhipova and Meselson 2000). Stronger evidence comes from species that vary in their mode of reproduction. For example, populations of the water flea Daphnia pulex can be either cyclical parthenogens (“sexuals”), which during the growth season reproduce asexually and then switch to sexual reproduction in response to seasonal signals, or, alternatively, completely asexual. Schaack et al. (2010) compared the abundance of transposons in sexual and asexual populations, finding that the sexuals harbour more transposons than do asexuels. Evidence that sex may promote the spread of genomic conflict has also been obtained experimentally. For example, when infected with a novel transposon, rapid spread was observed in sexual but not asexual strains of yeast (Zeyl et al. 1996). Finally, the selfing plant Arabidopsis thaliana has fewer transposons than its outcrossing relative A. lyrata (Hu et al. 2011) and there is evidence suggesting that this is due to accumulation of transposons in the outcrosser, rather than loss in the selfer (Slotte et al. 2013).
8.4.1 Alternative strategies for transposons

The above argument assumes that transposons are deleterious and persist in genomes thanks to their selfish behaviour. An alternative strategy for a transposon would be to adopt the same approach that many other genes do, that is to work with other genes to promote whole organism level fitness. Arguments that transposons may provide adaptive benefits at the organism level have a long history (Le Rouzic and Deceliere 2005; Biemont 2010), and recent studies suggest that examples of transposon-induced adaptations do exist, not only in bacteria, but also in eukaryotes (González and Petrov 2009b; Lisch 2013). An important avenue for future work should thus be to determine how common this is (de Souza et al. 2013). However, the evidence to date is overwhelmingly in support of the hypothesis that transposons are predominately associated with fitness costs, and that their maintenance in populations is due to their ability to transpose despite these costs (Kidwell and Lisch 2001; O’Donnell and Burns 2010). Thus, even if beneficial transposon insertions may occur, most insertions are likely to pose a threat to the maintenance of genome integrity.

8.5 Policing the genome

The enforcement of cooperation through policing has long been recognized as a likely important factor in driving evolutionary transitions (Leigh 1977; Alexander 1987; Frank 2003; Ratnieks and Helanterä 2009). While our understanding of policing mechanisms has made great progress through studies on whole organisms, in particular social insects (Ratnieks et al. 2006), the first
articulation of the argument that conflict suppression could promote cooperation was made in
reference to fair meiosis (Leigh 1971). Furthermore, it has long been recognized that several
features of the genetic system appear to be adaptations to reduce conflict among genes.
Examples of this include randomized segregation during meiosis (Haig and Grafen 1991; Ridley
2001), enclosing genes in cells (Michod 1983; Maynard Smith and Szathmáry 1995), organizing
genes in chromosomes (Maynard Smith and Szathmáry 1995) and the uniparental inheritance of
organelle genomes (Cosmides and Tooby 1981). Still, until recently, technical limitations made
the genome the most inaccessible of all hierarchical levels at which to empirically study
cooperation and our understanding of the molecular mechanisms of policing at the gene level
remained limited.

8.5.1 Empirical examples of transposon policing

Over the last few years, our improved understanding of policing of transposons has been
instrumental in changing this picture. We now know that, in most genomes, the majority of
transposons are not actively transposing at high rates, because their mobility is down regulated
through several mechanisms (Slotkind and Martienssen 2007). Transposons are typically
epigenetically silenced by a variety of small interfering RNAs. These small RNAs are then
incorporated into a larger protein complex, which either destroy transposon transcripts or target
transposon sequences with chromatin modifications or DNA methylation and so preventing the
transposon from transposing (Malone and Hannon 2009).

The relationship between transposons and their silencers can be viewed as an example of
antagonistic co-evolution between selfish genetic elements and their modifiers. This arms race is
often rapid and different populations may therefore fix for different sets of selfish genetic
elements and modifiers. As a consequence, when individuals from different populations or
species interbreed the offspring may suffer from increased levels of intragenomic conflict due to a mismatch of inherited selfish genetic elements and modifiers (Johnson 2010). A good example of this is the experimental cross between the plants Arabidopsis thaliana and A. arenosa, which results in elevated expression of ATHILA transposons in the hybrid (Josefsson et al. 2006). Similarly, crosses between two species of Australian wallaby (Macropus eugenii and Wallabia bicolor) led to a large increase in the size of centromeres due to proliferation of retroelements (O’Neill et al. 1998). Finally, because of proliferation of the Ty3/gypsy retrotransposon, the three Helianthus sunflower species H. anomalus, H. deserticola and H. paradoxus, all products of the same hybridization even between H. annuus and H. petiolaris, have genomes about 50% larger than either parental species (Ungerer 2006).

8.5.2 Efficiency of transposon policing

What may determine how efficient transposon policing is? There is growing evidence that the policing of transposons comes with a fitness cost, as the silencing of transposons can also affect the expression of nearby genes (Hollister et al. 2011). Yet, at any given time and in most genomes, the majority of transposons are actively silenced (Lister et al. 2008). The integrity of the genome is thus appears highly contingent on the efficacy of the silencing system, and the ongoing maintenance of genomic conflict may be driven in part by location-specific costs of silencing. Furthermore, our immune system’s ability to deal with antigens is weakened when our bodies are under stress and the same seem to be true for control of transposons. Examples of stresses that have been linked to increased transposon expression include infections by bacteria and fungi, wounding, high and low temperatures, and water availability (Grandbastien et al. 2005; Hashida et al. 2006). Thus, changes in both the genomic and external environment may
result in the breakdown of transposon silencing. Taken together, these observations of genomic consequences of silencing breakdown highlight the vital importance of policing to the maintenance of genome integrity.

8.6 Social group maintenance in the genome

How much of a threat did transposons present to early genomes? While transposons occur primarily in (sexually reproducing) eukaryotic genomes, they are also found in prokaryotic genomes. Furthermore, both self-regulation (Reznikoff 2008) and policing (Casadesus and Low 2006) appear to play some role in regulating their activity. Durand and Michod (2010) recently applied empirical examples from studies of transposable elements to Eigen and Schuster's theoretical models of hypercycles of replicators to elucidate the early stages of the formation of the genome as a social group. They argued that the division of labour between genes influencing viability and reproduction, respectively, was key to both the formation and transformation of the genome as a level of individuality. However, they also pointed out that to fully understand the origin of the genome, we must also understand what prevents conflict from arising among the genes constituting the genome. That is, what governs the maintenance of social groups at the genome level (Figure 8.1)? The well-developed literature on evolutionary transitions provides a solid conceptual framework to address this question using the large amounts of genomic data on transposable elements currently being generated.

Reciprocally, these new empirical observations may also stimulate the further development of the current conceptual framework of transitions in individuality. Michod (2005) argued that the full decoupling of fitness from lower levels is required before the evolution of a
new level of individuality is complete. Conflict at lower levels must therefore be (almost) completely suppressed; other authors have reached similar conclusions (Gardner and Grafen 2009). While self-fertilizing and asexual lineages are expected to experience less genomic conflict, the majority of multicellular organisms, including plants, are outcrossing (Igic and Kohn 2006). A given level in the hierarchy can thus still function despite extensive conflict at lower levels, consistent with the growing consensus that the argument that complete transitions in individuality require the total suppression of conflict at lower levels is too stringent (Okasha 2006; Gardner 2013). Moreover, the demonstrated importance of silencing of transposons in maintaining genome unity suggests that policing, while playing a minor role in some transitions (El Mouden et al. 2010), may be more important than previously anticipated in others. Thus, while the same principles appear to operate at different levels in the hierarchy, their relative importance may vary, and the only way to test this will be to apply our theoretical models to new data (Queller and Strassmann 2009).

8.7 The future of social evolution is in the genome

There is currently an emerging synthesis marrying social evolution, the theoretical backbone of studies of evolutionary transitions, with genomics (Robinson et al. 2005; Foster 2011; Van Dyken and Wade 2012). This synthesis has two fruitful outcomes. First, it means that the genetic architecture of social behaviour is no longer beyond reach (Robinson et al. 2008). A second, perhaps less appreciated, outcome is that it, for the first time, allows us to fully explore fully the extent to which the theoretical framework of social evolution can be applied to the evolution of
genome architecture. As illustrated in this paper, the potential for this second outcome is promising.

Now is a particularly good time to explore the second outcome. In 1983, Stephen Jay Gould (1983) asked: “How much repetitive DNA is self-centered DNA? If the answer is ‘way less than one percent’ because conventional selection on bodies almost always overwhelms selection among genes, then self-centered DNA is one more good and plausible idea scorned by nature. If the answer is ‘lots of it,’ then we need a fully articulated hierarchical theory of evolution”. One need not embrace Gould’s particular version of a hierarchical theory to recognize that, thirty years on, our improved understanding of genome architecture leads to a resounding confirmation of the ‘lots of it’ alternative (Aziz et al. 2010). In other words, there are almost endless amounts of data that can be approached with the tool kit of social evolution. As has been pointed out elsewhere (Bourke 2011), the application of the ideas of social evolution to entities below that of the individual is still restricted to a few enthusiasts (e.g. Queller and Strassmann 2009; Foster 2011; Haig 1997; Gardner and Welch 2011).

8.8 Concluding remarks

Issues surrounding conflict and cooperation in a hierarchical setting and evolutionary transitions in individuality can be organized around two questions: (i) under what circumstances are the fitness interests of lower and higher levels aligned? and (ii) what prevents selfish behaviour at a lower level from disrupting the functionality of higher levels? Here, I have reviewed empirical evidence from studies of transposable element that can be interpreted in the evolutionary transitions framework. Transposable elements are just one example of genomic conflict; the
broader genomic conflict literature offers a gold mine of data to further examine the social
interactions that occur within the genome. Comparisons between the organization of
multicellular organisms and societies date back a long time (Wheeler 1911), it is now time to
take the final step and integrate the genome, and the empirical resources of genomics, into the
general theory of social evolution and the study of evolutionary transitions in individuality.

8.9 Acknowledgements

This review benefited greatly from the helpful comments received from Carlos Arnillas, Andrew
Bourke, Ellen Clarke, Kevin Foster, Richard Michod, Locke Rowe, Stuart West, Stephen
Wright, Marlene Zuk, the editor and two anonymous reviewers. Thanks also to Susana
Wadgymar, Stephen De Lisle, and Art Weis for organising Fitness: Levels and Landscapes. I am
supported by a Junior Fellowship from Massey College.
CHAPTER 9 CONCLUDING DISCUSSION

This thesis was motivated by two main questions. First, what allows TEs to proliferate so successfully and make the genomes so large in some species, but not so in others? And, second, what prevents the deleterious effects of TE activity to inhibit the production of functioning organisms? To address these questions, I used a combination of population and comparative methods on whole genome and transcriptome data from a variety of plant species. More specifically, I examined the role of mating system shifts in driving variation in TE abundance and genome size across related plant species; tested the importance of co-transmission in explaining the chromosomal distribution of cyto-nuclear genes; and synthesised the empirical TE literature with the conceptual framework of evolutionary transitions in individuality and social evolution. In this final chapter, I first summarize the main insights from the previous chapters and then outline some avenues for future work to further examine the role of mating system in TE evolution in plants, the role of TEs in genome size evolution, and what TEs can contribute to the general study of conflict and cooperation.

9.1 Summary of thesis chapters

9.1.1 Chapter 2 Sizing up Arabidopsis genome evolution

This chapter took as its starting point the publication of the reference genome of Arabidopsis lyrata, the outcrossing relative of the highly selfing A. thaliana. I argued that while the higher TE abundance and larger genome of A. lyrata is consistent with the importance of outcrossing for the accumulation of TEs, comparing the genomes of only two species does not allow us to
rule out the alternative explanation of higher rate of DNA loss in *A. thaliana*. I suggested that the sequencing of the *Capsella rubella* genome would be an important resource in order to anchor the direction of the TE and genome size change and to determine what processes that have been most important in shaping the evolution of genome size and structure in *Arabidopsis*.

9.1.2 Chapter 3 The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution

This chapter described my contributions to the *Capsella rubella* genome project. The goal of my analysis was two-fold. First, I addressed the directionality of the TE and genome size change in *Arabidopsis* outlined above. By using an alignment of the *C. rubella*, *A. thaliana*, and *A. lyrata* genomes, I showed that the regions that have expanded in the *A. lyrata* genome relative to the other two were regions with a higher density of TEs. Second, I used Illumina short read genome and transcriptome data to show that while *C. rubella* has experienced a genome-wide reduction in the efficacy of selection since the recent divergence from its outcrossing progenitor *C. grandiflora*, there has not been a simultaneous shift in TE abundance or expression. This observation presented the first hint that the effect of the evolution of selfing on TE evolution may depend on how long the species has been selfing and provided the motivation to expand the mating system comparison to additional species in *Capsella*.

9.1.3 Chapter 4 Mating system shifts and transposable element evolution in the plant genus *Capsella*

Here, I compared TE abundance, genome distributions, and population frequencies in three *Capsella* species: the recent selfer *C. rubella*, the older selfer *C. orientalis* and their outcrossing progenitor *C. grandiflora*. I demonstrated that the two selfing species have very different TE
profiles. Whereas *C. orientalis* has few and inactive TEs, relative to *C. grandiflora*, there is evidence to suggest that *C. rubella* has the highest overall TE abundance of the three species. Yet, *C. grandiflora* appears to have the highest contemporary TE activity. I used these observations to hypothesize that the effect of the evolution of selfing on TE dynamics may depend on the age of selfing in the species.

9.1.4 Chapter 5 Co-evolution between transposable elements and their host: a major factor in genome size evolution?

In this chapter, I reviewed evidence of a co-evolutionary relationship between TEs and the rest of the genome. Based on this, I argued that by considering the arms race between TEs and the silencing machinery employed to control them, traditional predictions about the rate and direction of genome size evolution may change. I discussed theoretical and empirical work examining factors affecting variation in transposition rate, and how this variation can be linked to differences in genome size. I paid particular attention to the rapidly growing literature on the silencing of TEs by various host mechanisms, such as methylation and the activity of small interfering RNAs. Finally, I provided suggestions of how predictions stemming from this perspective may be empirically tested.

9.1.5 Chapter 6 No evidence that sex and transposable elements drive genome size variation in evening primroses

The goal of this chapter was to test the age-of-selfing hypothesis developed in Chapter 4 in a genus with multiple mating system shifts. While there are important differences between the genomic consequences of selfing and asexuality, the repeated evolution of functional asexuality across the genus made the evening primrose genus *Oenothera* an attractive system. I found that
variation in sexual/asexual reproduction could not explain the almost two-fold variation in genome size. Moreover, the genome sequence analysis suggested that the genome size variation was not associated with the proportion of the genome made up of TEs. Instead, accumulation of simple sequence repeats seems to best explain genome size expansion in this genus.

9.1.6 Chapter 7 Chromosomal distribution of cytonuclear genes in a dioecious plant with sex chromosomes

In this chapter, I used transcriptome data from the dioecious *Rumex hastatulus* to examine whether the elevated cotransmission between the X chromosome and organellar genes compared with autosomal genes, has lead to an over- or underrepresentation of cytonuclear genes on the X chromosome relative to the autosomes, as predicted by the coadaptation and sexual-conflict hypotheses respectively. I found no evidence of either hypothesis, and discuss how additional factors, including the time since the origin of the sex chromosomes as well as the age of the X-Y recombination suppression may underlie the observed patterns.

9.1.7 Chapter 8 Evolutionary transitions in individuality: insights from transposable elements

In this chapter, I reviewed studies of TEs within the context of evolutionary transitions in individuality. I argued that information on TE dynamics shed light on two key questions. First, I pointed out that studies on the role of mating system on TE evolution provide insights to the circumstances under which the fitness interests of lower and higher levels can be expected to align. Second, I suggested that silencing of TEs through siRNAs and methylation is a good example of the kind of mechanisms that prevent selfish behaviour at a lower level from disrupting the functionality of higher levels.
9.2 Avenues for future work

In this thesis, I have taken a variety of approaches to investigate the causes and consequences of TE evolution in plants. Below, I briefly outline three potential extensions of this work.

9.2.1 The rate and direction of TE evolution following a mating system shift

A central theme of this thesis has been to test the effect of mating system shifts on TE and genome size evolution. Taken together, my results do not present strong support for either hypotheses of increase or decrease of TEs following a shift, but instead I have sometimes observed a higher abundance in the outcrosser (Chapter 4), contrasting effects in two selfers (Chapter 4), or no effect (Chapter 6). I have suggested that one way to reconcile these observations is that the effect of mating system depends on age (Figure 9.1). One way to test this would be to apply the approach taken here more broadly in a phylogenetic context to systems with multiple mating system shifts, such as the Solanaceae or Brassicaceae families. Do young selfers in general have a higher, but old selfers lower, TE abundance relative to their outcrossing progenitor?

This divergence time effect could also be investigated theoretically by incorporating repeated shifts in mating system into the theoretical models that underlie the predictions tested empirically here, including the computer simulations of Wright and Schoen (1999) and the analytical framework of Morgan (2001). Extending these theoretical methods would also provide an opportunity to incorporate insights from our improved understanding of silencing of TEs. This
was done successfully for *Drosophila* by Lu and Clark (2010) by extending Dolgin and Charlesworth’s (2008) simulations to include information about TE suppression by piRNAs. Given the evidence that silencing of TEs is more effective in the selfing *A. thaliana* than in outcrossing *A. lyrata* (Hollister et al. 2011), examining the co-evolution between mating system and silencing could be particularly interesting.

![Graph showing TE copy number ratio (selves/outcrosses) over divergence time](image)

**Figure 9.1 Effect of selfing on TE evolution**

The evolution of selfing may initially be associated with an increase in TE copy number relative to the outcrossing progenitor as the reduction in efficacy of selection associated with selfing results in an initial accumulation of TEs inherited from the outcrossing progenitor, but over time the lack of outcrossing prevents the spread of new active TEs.

### 9.2.2 The causes of genome expansion

TE abundance and genome size appear to correlate strongly in plants (Tenaillon et al. 2010; Michael 2014). TE driven genome expansion is also what we observe, at least in part, in *Arabidopsis* (Chapter 3) and the smallest *Capsella* genome has the lowest TE abundance.
(Chapter 4). However, within *Oenothera* the genome variation is best explained by differences in the accumulation of simple repeats. This is in line with observations from within *A. thaliana* (Long et al. 2013), *Eucalyptus* (Myburg et al. 2014), and maize (Rivin et al. 1986; Chia et al. 2012; Diez et al. 2014), where genome size differences on smaller phylogenetic time scales appear to be due to simple repeats, rather than TEs. It would therefore be interesting to build on this thesis by asking if the cause of a genome size difference between a pair of species depend on their divergence time. Are genome size differences within species more likely due to simple repeats and differences between distant relatives due to TE proliferation? This also raises the theoretical challenge of explaining why evolutionary forces allow the accumulation of simple repeats but not TEs, or vice versa.

The influx of new reference genomes of closely related species offers a promising avenue for examining the causes of genome expansion. One example is the *Arabidopsis* and *Capsella* relative *Arabis alpina* (Willing et al. 2015). Here, the genome has expanded relative to other Brassicaceae species such that *A. alpina* has more TEs than the entire *A. thaliana* genome. This massive TE accumulation is associated with a reduced capacity for silencing and removal of LTR (long terminal repeat) retrotransposons in *A. alpina*, consistent with the key role of co-evolution between TEs and the rest of the genome in causing variation in genome size in plants outlined in Chapter 5. A clue to when TEs will be the cause of genome size change could therefore come from determining why the efficacy of silencing varies across species (Kelly and Leitch 2011; Ågren and Wright 2015). A caveat to the expectation that improved silencing efficiency should result in fewer TEs comes from Fedoroff (2012) and Kelly et al. (2015), who have pointed out that if genome size is primarily driven by differences in deletion rates, effective silencing of TEs may in fact result in higher TE accumulation. They argue that silenced TEs will
undergo less ectopic recombination than non-silenced TEs, and therefore experience lower deletion rates.

9.2.3 Transposable elements and the general theory of conflict and cooperation

The premise of Chapter 8 is that we are faced with the same conceptual problems regardless of what level in the hierarchy of individuality (Figure 1.3) that we are interested in and that we therefore have a lot to learn from considering examples from different levels simultaneously. Moving between levels, though, is not without problems. For example, the population genetic models used in studies of TEs have little in common with the game theoretical approaches of researchers of parent-offspring conflict (Charlesworth 2000). Should we try to find mathematical principles that can describe both situations in a way useful for all researchers? Or are we equally well off with separate frameworks? I believe central to addressing this issue will be to apply old modelling tools to new data. One good example of a new empirical resource is the genome. The rapid influx of whole genome data thanks to advances in high throughput sequencing has opened up a gold mine for exploring the extent to which the theoretical framework of (whole organism) social evolution can be applied to the evolution of genome architecture.

A promising avenue of conceptual integration is examples of what Queller (1997) referred to as “egalitarian” cooperation. Whereas ”fraternal” cooperation occurs when the interacting units are genealogical kin, there are also many situations where relatedness is not possible. This includes genes in genomes, the organellar and the nuclear genomes in cells, and between-species mutualisms and symbioses. The key to egalitarian cooperation is an effective
mechanism of conflict suppression. Such mechanisms have been well studied on their own in all examples of egalitarian cooperation, but they are rarely discussed together.

9.3 Conclusion

Transposable elements are the dominant component of plant genomes and any serious attempt to understand the evolution of plant genome structure and size must therefore consider the population biology of TEs. As demonstrated in this thesis, testing classic population genetic models of TE dynamics in whole genome sequences from closely related species of varying mating system offers an excellent window into this problem. Moreover, these studies also provide an opportunity to incorporate TEs into new theoretical frameworks, such as that of social evolution. Overall, the continued effort to uncover the evolutionary causes and consequences of TE proliferation promises to have implications far beyond the field of plant genome biology.
References


Aziz RK, Breitbart M, and Edwards RA.. 2010. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Research* 38: 4207–4217.


Genome Structure and Function. Springer Verlag.


Grover CE and Wendel JF. 2010. Recent insights into mechanisms of genome size change in plants. Journal of Botany Article ID 382732


5654–5666.


Hough J, Hollister JD, Wang W, Barrett SCH, and Wright SI. 2014. Genetic degeneration of old


element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. *Genome Biology and Evolution* 3: 156–167.


Koch MA and Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the


Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, and


R Development Core Team. 2013. R: a language and environment for statistical computing. R


Schaack S, Pritham EJ, Wolf A, and Lynch M. 2010. DNA transposon dynamics in populations of


Shan XH, Liu ZL, and Dong ZY. 2005. Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (Zizania latifolia Griseb.). Molecular Biology


Appendix

Supplementary Note 3.1

Construction of the scaffold assembly

Prior to assembly, organellar reads were removed by screening against mitochondria, chloroplast and rDNA. For Roche 454 linear reads, any read <200bp was discarded. Roche 454 paired reads were split into pairs and any pair with a read shorter than 50bp was discarded. An additional deduplication step was applied to the 454 paired libraries that identifies and retains only one copy of each PCR duplicate. All remaining 454 reads were compared against a Illumina GA2x data totaling 48.22 Gbs and any insertion/deletions in the 454 reads were corrected to match the Illumina alignments. A total of 20,035,171 reads (see Supplementary Table 3.1 for library size summary) were assembled using a modified version of ARACHNE (Jaffe et al. 2003) v.20071016 with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True (see Supplementary Table 3.2 for overall scaffold and contig statistics).

Genetic mapping and pseudomolecule chromosome construction

We used multiplexed shotgun sequencing (MSG; Andolfatto et al. 2011) of 550 F2's from an interspecific cross between C. rubella and C. grandiflora (Slotte et al. 2012) to generate a genetic map to help identify false joins in the assembly and merge scaffolds into pseudomolecules. Our high-density linkage map, which made use of SNP markers genome-wide,
helped enable a precise identification of the positions of false joins, as regions showing discontinuities in linkage. For the final assembly process following the breakage of false joins, the markers were collapsed down into 768 non-redundant markers as previously described (Slotte et al. 2012).

Genetic markers were aligned to the assembly using BLAT (Kent 2002; Parameters: t=dna - q=dna -minScore=200 -extendThroughN). The combination of the genetic map, BAC/Fosmid paired end link support, and A. lyrata synteny were used to identify false joins, and integrate the assembled scaffolds into eight pseudomolecule chromosomes. Scaffolds were broken if they contained a putative false join with both genetic mapping and synteny support coincident with an area of low BAC/fosmid coverage. A total of 16 breaks were identified and broken, resulting in 1,875 scaffolds in the broken assembly. The optimal order and orientation of the broken scaffolds was obtained using all available evidence (768 collapsed markers, synteny, BES). A total of 37 joins were made on 45 scaffolds to form the final assembly containing 8 chromosomes capturing 124.9 Mb (92.7%) of the assembled sequence. Each join is padded with 10,000 Ns. After screening for contaminant, the final assembly contains 853 scaffolds (9,675 contigs) that cover 134.8 Mb of the genome with a contig L50 of 134.1 kb and a scaffold L50 of 15.1 Mb. Significant telomeric sequence was identified using the TTTAGGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly. Plots of the marker placements for the 8 chromosomes are shown in Supplementary Figure 3.1. A dot plot of the C. rubella assembly against the A. lyrata chromosomes is shown in Supplementary Figure 3.3.
Screening and final assembly release

Remaining scaffolds were classified into bins depending on sequence content. Contamination was identified using megablast against the NCBI nucleotide collection (NR/NT) and blastx using a set of known microbial proteins. Additional scaffolds were classified as unanchored rDNA (2 scaffolds), mitochondrion (4 scaffolds), chloroplast (16 scaffolds), repetitive (>95% masked with 24mers that occur more than 4 times in the genome; 123 scaffolds). We also removed 840 scaffolds that were less than 1 kb in sequence length. Resulting final statistics are shown in Supplementary Table 3.2.

Assessment of Assembly Accuracy

A set of 21 BAC clones and 13 Fosmid clones were sequenced in order to assess the accuracy of the assembly. The Fosmid clones were of the same genotype as the reference, and the BAC clones were of a different genotype. Minor variants were detected in the comparison of the fosmid clones and the assembly. In 12 of the 13 fosmid clones, the alignments were of high quality (< 0.50% bp error) with an example being given in Supplementary Figure 3.4a. The overall bp error rate (including marked gap bases) in the Fosmid clones is 0.22% (995 discrepant bp out of 450,806). Supplementary Table 3.4 shows the individual Fosmid clones and their contribution to the overall error rate. Note that one Fosmid clone (4002708) contributes nearly 50% of the discrepant bases. This is due to a small indel in the clone relative to the assembly shown in Supplementary Figure 3.4b (all dot plots were generated using Gepard; Krumsiek et al. 2007). The Fosmid clone that indicated a large discrepancy is shown in Supplementary Figure
3.4c. There are several indels of various sizes in the clone and assembly, typical of a region of degraded transposon.

The BAC clones were generated using the same ecotype as the genomic reads, but different generations, which can introduce discrepancies in the clones relative to the assembly. We found that 11 of the 21 BAC clones exhibited noteworthy discrepancies. The overall error rate in the 10 clones that exhibited high quality alignments was 0.218%, in good agreement with the Fosmid clone error rate. Supplementary Figure 3.4d is an example of a high quality alignment. Supplementary Figure 3.4e is an example of a BAC clone that exhibits a large discrepancy in a region of degraded transposon, while Supplementary Figure 3.4f is an example of a BAC clone that exhibits discrepancies that are potentially caused from the difference in generation between the genomic reads and the BAC clones.

**De novo Illumina assembly in Capsella grandiflora and Neslia paniculata**

**C. grandiflora de novo assembly**

The *C. grandiflora* assembly was performed using the meraculous toolkit (Chapman et al. 2011). Parameter settings for contig generation were: kmer-size 51 (-m 51), minimum depth threshold 3 (-D 3), and minimum contig size 51 (-c 51). The preliminary meraculous mer-frequency analysis indicated a genome size in the range 200-240 Mbp with an estimated heterozygosity of 1.4%. Initial contigs were generated using only the standard fragment library (insert size 289 +/- 25 bp). Two rounds of scaffolding were performed, first using the fragment library with a minimum pairing threshold of 3 (-p 3) and then using the mate-pair library (insert size 4553 +/- 400 bp)
with a minimum pairing threshold of 10 (-p 10). These thresholds were chosen by exhaustive optimization of N50 scaffold length at each round. Isolated contigs with mean kmer depth less than 24 were not included in scaffolding as these are deemed haplotype-specific contigs ("haplotigs"). After scaffolding, gap-closure was performed using the fragment library with the polymorphic-closure setting enabled (-P). The resulting assembly is comprised of 4,997 scaffolds larger than 1kbp, including 105.3Mbp of scaffold sequence (93.8 Mbp in contigs). Half of the assembled sequence is contained within 256 scaffold (2508 contigs) of size at least 98.1 kbp (10.3 kbp). A full table of assembled scaffold/contig sizes is provided in Supplementary Table 3.6.

**N. paniculata de novo assembly**

A partial assembly of the outgroup genome *Neslia paniculata* was undertaken with the aim of generating scaffolds of sufficient length for generally unambiguous alignment against the *C. rubella* reference and coverage of the majority of the non-repetitive sequence. 16.5 Gb of short paired-end sequences were generated (2×108 nt, 280 nt insert size) on the Illumina GAIIX platform to 80x from nuclear enriched DNA. Read data was 3' trimmed for low quality bases (Q<30) and assembled into contigs (N50 ~25 kb) using the Ray (v1.4) assembler (Boisvert et al. 2010) with a Kmer of 31. Contigs were imported into the SOAPdeNovo package (Li et al. 2010) for scaffolding (Kmer 51) with 28 Gb of 2×100 nt mate pair reads (Illumina HiSeq, 5 kb insert size, gently sheared) that were bloom filtered to remove duplicate reads and aligned against the contig reference to remove chimeric reads. This generated a 113 Mb assembly (scaffold L50 = 62 kb) with a 96% called base percentage (see Supplementary Table 3.7 for assembly statistics).
Comparison of *N. paniculata*, *C. grandiflora* and *C. rubella* genome assemblies

Our partial assemblies of the *Neslia paniculata* and *Capsella grandiflora* genomes are available online (http://grandiflora.eeb.utoronto.ca:8086/CG.fasta, http://grandiflora.eeb.utoronto.ca:8086/NP.fasta). While up to 90% of *C. rubella* CDS sequences can be located in these assemblies, a relatively lower 60-70% complete rubella genes can be identified through orthologous-chain liftover at >80% identity (Supplementary Table 3.8). Nonetheless between 80-90% of the *C. rubella* genome can be identified in blocks of sequence in the *C. grandiflora* and *N. paniculata* genomes (Supplementary Table 3.9, Supplementary Figure 3.5).

Assessment of Repetitiveness of Genome Assembly

To assess the degree to which multicopy repeats comprise the remainder of the genome missing from the assembly, we utilized Illumina whole-genome sequence from the reference strain. We took several approaches to assess genome completeness. First, raw fastq reads were mapped onto the reference genome, including chloroplast and mitochondrial scaffolds, using Stampy (Lunter and Goodman 2011) under default parameters. From the results of read mapping, the total number of reads mapping to each nuclear scaffold was assessed. To estimate the total genomic coverage of repetitive nuclear scaffolds, we first estimated the average coverage of the main eight chromosomal scaffolds, and assumed that this represented the single-copy coverage. To be conservative, we assumed that all unmapped reads belonged in the main eight chromosomal scaffolds and failed to map due to base quality. We then estimated the copy number of other genomic scaffolds as the total number of sequence reads mapping to those scaffolds, divided by
the inferred single copy nuclear coverage. As shown in Supplementary Figure 3.6, many of the smaller genomic scaffolds appear to be multicopy, indicating that much of the genome missing from the assembly is comprised of many copies of repetitive elements in our assembly. Summing the inferred sizes of each scaffold gives a total genome size estimate of approximately 187 Mb. Furthermore, full genome-wide estimates of per-base copy number, accounting for the multicopy repeats within the main genome scaffolds lead to an estimate of 198Mb, close to the expected genome size of 224 Mb based on flow cytometry. Although base composition biases and other library preparation biases might affect the quantitative estimates, this indicates that the ‘missing genome’ is mostly accounted for by repetitive scaffolds found within our assembly (Supplementary Table 3.6). Using BLAST (Altschul et al. 1997) searches it was determined that the high-copy scaffolds representing the most abundant multicopy repeats contain centromeric satellite repeats and rDNA units. Other scaffolds showing similarity to centromeric repeats are found to be in 14,000-20,000 copies.

To dissect genome content in more detail, a subset of 16,180,609 Illumina genomic DNA reads were analyzed (Supplementary Figure 3.7). These were serially assigned to each of a number of classes of sequence. After each stage of association, those that could be associated with that class were removed from the read-set and then the next class was aligned.

The stages of classification were:

1. Chloroplast (by alignment to the published *C. bursa-pastoris* chloroplast sequence in Genbank with high similarity)

2. Mitochondrial (by alignment to the *A. thaliana* mitochondrial sequence)

3. Telomeric repeats (by alignment to a series of TTTAGGG sequences)
(4) Centromeric repeats (by alignment to the *C. rubella* centromere satellite sequences in Genbank) and repeat models

(5) rDNA repeats (by alignment to a set of AT rDNA sequences in the SILVA http://www.arb-silva.de/aligner/ database)

(6) Alignment against de novo repeat models by Mathieu Blanchette and RepeatModeller (RepeatModeler Open-1.0 http://www.repeatmasker.org/)

(7) Alignment against the hard repeat masked genome to determine the non-repetitive reads

(8) Alignment against only the repeat masked portion of the genome to determine the repeats missed above

After classification above, very few reads were left at the end of the analysis that would correspond to euchromatic sequences missing from the data. The results generally imply that the repetitive part of the genome should be about twice the size assigned to it in the current build, corresponding to a genome of approximately 206 Mb (Supplementary Figure 3.7), which is close to the estimate of 219 Mb from flow cytometry. Of the residual reads, 19,000 could be aligned to the *A. thaliana* genome. 11,000 of these were mitochondrial but were diverged by more than a few percent from the mitochondrial reference suggesting they might be decaying/mutating nuclear transferred mitochondrial sequences similar to the chloroplast sequences. 7,000 of these reads corresponded to a 200 kb block of the *A. thaliana* genome that seems to be in the *C. rubella* read catalogue but is mostly missing in the assembly of *C. rubella*, at *A. thaliana* coordinates Chromosome 2:3,240,808-3,509,564. Finally, 136 reads came from a small section of *A. thaliana* scaffold 5 that again seems to be in the reads, but not the assembly, at *A. thaliana* coordinates Chromosome 5:4,488,564-4,492,946. This region was identified as being present in
earlier assemblies, but ended up removed from the final assembly as part of the process of breaking false joins and reassembling the genome.

**Genome structure and chromosome rearrangements**

Previous comparative genetic mapping analyses suggested minimal rearrangement events between *Capsella rubella* and *Arabidopsis lyrata*, with both species thought to maintain an ancestral gene order. To investigate the extent of chromosome rearrangements between the species, we conducted dotplot analyses using COGE’s synmap program, using the Quota align algorithm and synonymous substitution rates for inferring syntenic genes. Only three putative major chromosome rearrangements are detected between *C. rubella* and *A. lyrata*. Comparisons with *Schrenkiella parvulum* (Supplementary Figure 3.9) indicate that all three rearrangements are likely to be derived in *A. lyrata* and not *C. rubella*, reflecting derived rearrangements and/or assembly errors in *A. lyrata*. The first and most apparent putative rearrangement event involved a pericentromeric region of Cr 2, which localized to the rightmost end of the Al 1 assembly (Supplementary Figure 3.9). To verify our assembly of *C. rubella* in this region, we identified a BAC clone from *A. thaliana* from this region, T22H22, and conducted DAPI staining of pachytene chromosomes, verifying its location as pericentromeric on Chromosome 2 (Supplementary Figure 3.8b). Since this region has recently been identified as a polymorphic transposition event in *Arabidopsis thaliana* (Long et al. 2013), and the *A. lyrata* assembly in this region had been guided by synteny with *A. thaliana*, Columbia accession, this likely reflects an assembly error in the *A. lyrata* genome. Indeed, DAPI staining of the CVI accession of *A. thaliana* and in *A. lyrata* places T22H22 in the same syntenic position as *C. rubella* (Supplementary Figure 3.8b), confirming an assembly error in *A. lyrata*. 

220
Polymorphism and Expression Analysis

Expression Analysis

After aligning our reads to the *C. rubella* reference genome, the Cufflinks pipeline (Trapnell et al. 2012) was used to assemble transcripts, estimate their abundance and extract differentially expressed genes using the *C. rubella* annotation with known exons. Preliminary analysis indicated overloading of one of the three flow cells during cluster generation. Consequently low quality calls were discarded from this overloaded flow cell and the data was used only for SNP calling. The two remaining flow cells (four samples of each species) with generally high-quality data were used for both expression analysis and SNP calling. The data for gene expression was calculated as fragments per kilobase of transcript per million mapped reads (FKPM). Invariant set normalization across the *C. grandiflora* and *C. rubella* was then performed in Dchip (http://www.dchip.org/). Westfall and Young permutations of sample classifications were used to estimate the FDR for differential expression (DE) assignment under different significance criteria. The aim was to restrict false discovery rates to under 1%. Once a lower signal threshold (19 units, corresponding to the 10th percentile of ranked signal) had been introduced, the combination of fold change and p-value that resulted in the lowest type-II error were fold change > 1.5 and p<0.005. Differentially expressed genes were clustered hierarchically in DChip and samples were clustered by the DE gene list. All samples were appropriately classified, with intra-group variance approximately 10% of the inter-group variance. Ontology assignment of differentially expressed genes was undertaken both in the NIH DAVID (http://david.abcc.ncifcrf.gov/) system in which ontology clustering and pathway enrichment is
Gene expression data were tested for significant differences between species using a t-test, with a significance cutoff of 0.005, a fold change of 1.5 and a lower limit of expression for consideration of 19 normalized FPKMs (covering 90% of expression). Genes showing significant up- and down-regulation were tested for enrichment of functional categories using the Agrigo Bioinformatics database, where a false discovery rate (FDR) for the detection of differentially impacted ontology groups of 0.05 was set.

To further explore the impact of strongly perturbed genes in conjunction with the set of weakly perturbed genes, we identified differentially expressed pathways using a per gene FDR threshold of approximately 5%, corresponding to a fold-change of 1.3 (p<0.05) with the same lower expression limit. This increases the per-gene FDR but at no significant cost to the ontology group FDR. This revealed a number of differentially expressed ontology groups (Supplementary Table 3.11), including a number of groups associated with pollen development and morphology and floral development. We also identified a number of candidate genes found in QTL regions involved in floral evolution between selfing and outcrossing Capsella (Supplementary Table 3.12). Inspection of the intersection of these groups highlighted the Brassinosteroid pathway, involved in pollen development (Wilson and Zhang 2009), as a possible key differentially expressed pathway; of the 11 genes in this pathway with identified orthologues and expression estimates, eight show evidence of down-regulation in C. rubella (Supplementary Table 3.13). Although this could in part reflect simply differences in pollen number between the two flower bud samples, the MS1 transcription factor is found within a pollen number QTL, suggesting a possible direct role in the evolution of pollen number.
SNP Calling

For the polymorphism analysis of the RNA sequencing dataset (for samples, see Supplementary Table 3.14), SNP calls were conducted using the bcftools pipeline (Li et al. 2009). Because of the high heterozygosity of outbred C. grandiflora individuals, we required a minimum depth cutoff of 20x for each individual. Additional filtering included a minimum SNP quality score of 25, a minimum genotype quality score of 20, and to avoid spurious SNP calls around indels and splice junctions we excluded regions within 5 bp either side of an indel and/or a region of total coverage below 220x. For all of the analyses presented here, we included only sites for which all samples passed these thresholds, as well as the presence of informative bases for ancestral state inference from genomic resequencing of the close outgroup Neslia paniculata. We removed 325 loci with internal stop codons, any codons with missing data or gaps in any sample, and loci with 60 bp aligned length prior to analyses. A total of 4,226 loci remained after this filtering procedure, and these loci were analyzed for synonymous and nonsynonymous polymorphism and divergence. Prior to demographic analyses, we also filtered a total of 699 four-fold synonymous SNPs with heterozygous base calls in C. rubella. To estimate error rates, we conducted Sanger resequencing from genomic DNA of three samples of each species at 14 genes, for a total of 104,980 bp of sequence and 425 SNP calls. No false positives were identified in this dataset, suggesting a per base pair false positive error rate less than $10^{-5}$, and a per SNP false positive error rate less than 0.003. 21 false negative SNP calls were identified, mostly at a single gene and due to a highly divergent haplotype that experienced allelic dropout during read mapping, because of high SNP density.
Polymorphism and Divergence

We extracted four-fold synonymous sites and estimated divergence at four-fold synonymous and nonsynonymous sites using Polymorphorama, a perl script written by D. Bachtrog and P. Andolfatto (Bachtrog and Andolfatto 2006). We obtained the joint site frequency spectrum at four-fold synonymous and nonsynonymous sites using the compute and polydNdS programs based on Kevin Thornton’s libsequence (Thornton 2003).

Distribution of Fitness Effects of Nonsynonymous Mutations

We estimated the distribution of fitness effects of new nonsynonymous mutations (DFE) in C. grandiflora using DFE-alpha (http://liberty.cap.ed.ac.uk/~eang33/dfe-alphaserver.html; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). Estimates of the DFE were based only on data for C. grandiflora, and we used 4-fold degenerate synonymous sites as a neutral reference. As in Slotte et al. (2010) we found that most new nonsynonymous mutations are under strong purifying selection in C. grandiflora (88.4% of polymorphisms have \(N_e > 10\); Supplementary Table 13.5). The estimates of the parameters of the gamma distribution of negative fitness effects of new nonsynonymous mutations were –\(N_e\): 602.965338 and beta: 0.3925.

Demographic Model Inference

We estimated the parameters of a model for the split between C. rubella and C. grandiflora using dadi (Gutenkunst et al 2009). Demographic modeling was based on data on the folded joint site frequency spectrum for a total of 20,388 four-fold synonymous SNPs at 438,490 four-fold sites in 4225 loci.
We fit three demographic models that differed with respect to population size change (Supplementary Table 3.16). Parameter estimates for all three models were similar; based on the AIC, the preferred model included an initial bottleneck down to 2.3% of the ancestral population in association with the origin of *C. rubella*, followed by exponential population size expansion in both *C. grandiflora* and *C. rubella*. Population expansion was inferred to be stronger in *C. rubella* than in *C. grandiflora* (alpha 18.6 vs. 7.1) under this model, and parameter estimates for the rate of population size expansion in *C. grandiflora* agree with those in St. Onge et al (2011). Note that our time and demographic estimates differ in part from previous studies because we are using more recent estimates of the per generation substitution rate from direct experimental data (Ossowski et al. 2010), which are also in line with recent fossil dating. Because this substitution rate is substantially lower (2.14 times) than previous estimates, this has the effect of lowering estimates of effective population size and increasing divergence times more than twofold. Accounting for this, the timing of the split is within the confidence intervals of the estimate of Foxe et al. (2009) but the effective sizes are not. This may reflect differences in the sampling strategy or reflect the choice of loci.

**Forward Population Genetic Simulations**

Forward population genetic simulations with purifying selection were conducted in SFS_CODE (Hernandez 2008). We simulated coding sequence data according to the demographic model inferred in dadi, with nonsynonymous mutations having deleterious effects drawn from a gamma distribution with parameters as estimated for *C. grandiflora* in DFE-alpha. The total number of loci in each simulation was 4225, and we assumed that loci were unlinked and equal length, with no recombination within loci and that mutations followed an infinite sites mutation model. The
weighted mean of our per-locus estimate of Watterson’s theta at four-fold synonymous sites in *C. grandiflora*, 0.0168, was used as the per-site scaled mutation rate in these simulations. A total of 500 simulations were done under each model. As a check of model fit, we assessed whether the fit of observed *C. grandiflora* synonymous and nonsynonymous folded site frequency spectra to our simulations. We summarized the results for both *C. rubella* and *C. grandiflora* in terms of numbers of unique and shared polymorphisms and fixed differences at synonymous and nonsynonymous sites. All counts of unique and shared polymorphisms and fixed differences were obtained from the SFS_CODE output using custom perl scripts written by T Slotte. We conducted forward population genetic simulations with selection under the model described above, as well as under a simpler model with an instant size change in both *C. rubella* and *C. grandiflora*. Reassuringly, simulations under these models with a distribution of negative selection coefficients derived from *C. grandiflora* data fit well with the observed site frequency spectra for both synonymous and nonsynonymous sites in *C. grandiflora* (Supplementary Figure 3.11; results shown are for the split+growth model but were similar for the instant change model).

There is also a good qualitative agreement between the relative numbers of shared and unique polymorphisms and fixed differences at synonymous and nonsynonymous sites in our simulations and in our data (Supplementary Figure 3.11). However, some aspects of the model are clearly unrealistic, as our observed data on shared and fixed differences do not fall within the approximate confidence intervals of the simulated data. It is possible that model fit could be improved by a more realistic treatment of linkage among loci, as assuming free recombination between loci can cause confidence intervals to be too narrow. It should also be noted that our models assume additivity of fitness effects, and varying dominance could contribute to departures from model predictions. In any case, the relatively good overall agreement between
our data and simulations suggests that patterns of polymorphism and fixed differences at nonsynonymous sites in *C. rubella* are consistent with a model where there is relaxation of purifying selection due to the population size change alone, but without a change in the shape of the distribution of fitness coefficients.

**Transposon Annotation and Analysis**

**TE annotation**

The TEedenovo pipeline from the REPET v2.0 package (Flutre et al. 2011) was used for de novo identification of repeated sequences in the following genomes: *A. thaliana* (ecotypes Col, Ler, Kro, Bur, and C24; http://www.1001genomes.org/), *Arabidopsis lyrata*, *Arabidopsis lyrata* (courtesy of Dr. George Coupland), *Brassica rapa*, *Capsella rubella*, *Eutrema halophila*, and *Schrenkiella parvulum*. The consensus sequences identified were analyzed using the REPET classifier tool which screens for structural features characteristic of transposable elements, searches for similarity with known TE sequences from Repbase, and probes for virtually all Pfam hidden Markov models. Classification was curated on the basis of the evidence collected for each consensus. Consensus sequences from each species were probed against their respective genome using the TEannot pipeline from the REPET v2.0 package in order to identify those with at least one full-length genomic copy. The selected sequences from all species were combined into the Brassicaceae repeat library to which we also appended the *A. thaliana* repeat library from the Repbase database. TEannot was then run again against the *C. rubella*, *A. thaliana* and *A. lyrata* assemblies using the Brassicaceae repeat library. Tandem repeats were not considered in this
holistic approach and only the unclassified repeats with ten or more copies in the source genome were counted for coverage calculations.

TE expression, insertion frequency, and copy number analysis from Illumina data RNaseq Illumina reads from four C. rubella and four C. grandiflora individuals were aligned to the identified TE database with the Stampy aligner (Lunter and Goodman 2011) using default settings. TE abundance was estimated from the number of reads mapped to TEs relative to the number of reads mapped anywhere in the genome using SAMtools (Li et al. 2009). Genomic abundance of TEs was similarly estimated using Illumina reads from genomic DNA from two accessions each of C. rubella and C. grandiflora.

To identify individual TE insertion sites across the genomes of resequenced samples, PoPoolationTE (Kofler et al. 2012) was run on paired-end Illumina genome sequences from two C. rubella and two C. grandiflora individuals using default settings. This program identifies TE insertion sites by finding paired end reads, where one read maps to a unique position in the repeat-masked genome and the second read maps to a transposable element.
Supplementary Figures

Supplementary Figure 3.1 Marker placements for the genetic map on *Capsella rubella* chromosomes

(a) Marker placements for the genetic map on *Capsella rubella* chromosome 1.
Supplementary Figure 3.1b. Marker placements for the genetic map on *Capsella rubella* chromosome 2.

Supplementary Figure 3.1c. Marker placements for the genetic map on *Capsella rubella* chromosome 3.
Supplementary Figure 3.1d. Marker placements for the genetic map on \textit{Capsella rubella} chromosome 4.

Supplementary Figure 3.1e. Marker placements for the genetic map on \textit{Capsella rubella} chromosome 5.
Supplementary Figure 3.1f. Marker placements for the genetic map on *Capsella rubella* chromosome 6.

Supplementary Figure 3.1g. Marker placements for the genetic map on *Capsella rubella* chromosome 7.
Supplementary Figure 3.1h. Marker placements for the genetic map on *Capsella rubella* chromosome 8.
Supplementary Figure 3.2 Pairwise linkage plot from an initial de novo assembly of the *C. rubella* genome.

(a) Heatmap indicates regions of high linkage (red), intermediate (yellow), low (green) and none (blue). Arrow shows an example of a scaffold with a sharp break in linkage, indicating a false join in the assembly.
Supplementary Figure 3.2b Linkage plot from final genome assembly, indicating that false joins have been removed.

Compare to Supplementary Figure 3.2a.
Supplementary Figure 3.3 Dot plot of the final *C. rubella* assembly (x-axis) against the *A. lyrata* assembly (y-axis).
Supplementary Figure 3.4 Dot plot of Fosmid clones on regions of *C. rubella* scaffolds

(a) Dot plot of Fosmid clone 4002707 on a region of scaffold_7. This alignment is representative of the high quality Fosmid clone alignments in 12 of the 13 available Fosmid clones.
**Supplementary Figure 3.4b. Dot plot of Fosmid clone 4002708 on a region of scaffold_6.**

This clone contains a small insertion relative to the assembly.
Supplementary Figure 3.4c. Dot plot of Fosmid clone 4002704 on a region of scaffold_7 composed primarily of degraded transposons.
Supplementary Figure 3.4d. Dot plot of BAC clone 9285 on a region of scaffold_3. This alignment is representative of the high quality clone alignments in 10 of the 21 available BAC clones.
Supplementary Figure 3.4e. Dot plot of BAC clone 9297 on a region of scaffold_2.

This one is composed primarily of degraded transposons.
Supplementary Figure 3.4f. Dot plot of BAC clone 9295 on a region of scaffold_3, with several indels of various size.

This discrepancy is likely due to the differences in generation between the genomic reads and the clones.
Supplementary Figure 3.5 CoGe dotplot comparison.

CoGe dotplot comparison, using SynMap of *C. rubella* with (A) *N. paniculata*, and (B) *C. grandiflora*. 
Supplementary Figure 3.6 Observed size of scaffolds on the x-axis vs. inferred size based on Illumina coverage on the y-axis.

The 1:1 relation is shown in red. The 8 chromosomal scaffolds are circled.

Supplementary Figure 3.7 Classification of Illumina genomic reads into repetitive and nonrepetitive categories.
Supplementary Figure 3.8 Chromosome painting analysis

(a) Positions of major chromosome landmarks (5S and 45S rDNA loci, interstitial telomere repeats and pericentromeric heterochromatin) within the idiogram of *C. rubella* (left) and on inverted diakinetic bivalents (right). The identity of all 16 chromosome arms was revealed by comparative chromosome painting using *A. thaliana* BAC contigs (see Chapter 3 Methods for details). The 24 ancestral genomic blocks of the Ancestral Crucifer Karyotype (Schranz et al. 2006) are indicated by uppercase letters (A to X).

(b) FISH analysis of the boundary between genomic blocks C and D in *Capsella, Arabidopsis lyrata*, and *A. thaliana* ecotypes Cvi and Col-0. Differentially labeled BAC contig F6D8 – F13D13 (~1.9 Mb in *A. thaliana* Col) hybridized on two chromosomes in *Capsella* and *A. lyrata*, and one chromosome in *A. thaliana*. Note the altered position of BAC T22H22 in Col as compared to the ancestral pattern in Cvi, *A. lyrata* and *Capsella*. Downward-pointing arrows indicate the opposite orientation of block D compared to the position in the Ancestral Crucifer Karyotype. Pachytene chromosomes were counterstained with DAPI. Scale bars, 10 µm.
Supplementary Figure 3.9 Dotplot comparisons of *Arabidopsis lyrata*, *Capsella rubella*, and *Schrenkiella parvula*.

Three major rearrangements are highlighted, which are not detectable between *Capsella rubella* and *Schrenkiella parvula*, consistent with the rearrangements being unique to the *Arabidopsis lyrata* genome assembly.

Supplementary Figure 3.10 Principal components analysis and hierarchical clustering analysis of differentially expressed genes in *C. rubella* and *C. grandiflora*.

Analyses were conducted using the Dchip package.
Supplementary Figure 3.11 Frequency spectra from *C. rubella* and *C. grandiflora*.

(a) Joint site frequency spectra for four-fold synonymous sites (left), and for nonsynonymous sites (right). (b) Fit of site frequency spectra in *C. grandiflora* to simulated data. (c) Fit of shared and unique polymorphisms and fixed differences to observed data. In panels (b) and (c) the left panel shows results for synonymous sites, whereas the right panel shows results for nonsynonymous sites, and for simulated data, arrows indicate approximate 95% confidence intervals.
 Supplementary Tables

 Supplementary Table 3.1 Genomic libraries included in the *Capsella rubella* genome assembly and their respective assembled sequence coverage levels in the final release.

* Indicates that the number reported in the table is the average read length and standard deviation, not insert size.

<table>
<thead>
<tr>
<th>Library</th>
<th>Sequencing Platform</th>
<th>Average Read/Insert Size</th>
<th>Read Number</th>
<th>Assembled Sequence Coverage (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>454 XLR</td>
<td>353 ± 156*</td>
<td>8,998,560</td>
<td>16.88</td>
</tr>
<tr>
<td>GOOT</td>
<td>454 XLR paired</td>
<td>2,390 ± 876</td>
<td>777,790</td>
<td>0.73</td>
</tr>
<tr>
<td>GOOU</td>
<td>454 XLR paired</td>
<td>2,455 ± 880</td>
<td>635,604</td>
<td>0.56</td>
</tr>
<tr>
<td>GOOW</td>
<td>454 XLR paired</td>
<td>2,476 ± 852</td>
<td>735,804</td>
<td>0.67</td>
</tr>
<tr>
<td>HUCP</td>
<td>454 XLR paired</td>
<td>4,094 ± 398</td>
<td>526,360</td>
<td>0.30</td>
</tr>
<tr>
<td>HWZU</td>
<td>454 XLR paired</td>
<td>4,471 ± 368</td>
<td>718,406</td>
<td>0.28</td>
</tr>
<tr>
<td>HUFS</td>
<td>454 XLR paired</td>
<td>8,275 ± 1,380</td>
<td>2,554,138</td>
<td>1.04</td>
</tr>
<tr>
<td>GAYO</td>
<td>454 XLR paired</td>
<td>10,655 ± 1,774</td>
<td>852,924</td>
<td>0.68</td>
</tr>
<tr>
<td>FITO</td>
<td>Sanger</td>
<td>35,317 ± 4,190</td>
<td>308,832</td>
<td>0.82</td>
</tr>
<tr>
<td>FSWU</td>
<td>Sanger</td>
<td>37,158 ± 3,654</td>
<td>24,576</td>
<td>0.11</td>
</tr>
<tr>
<td>BSIA</td>
<td>Sanger</td>
<td>40,132 ± 4,092</td>
<td>10,752</td>
<td>0.04</td>
</tr>
<tr>
<td>BPHG</td>
<td>Sanger</td>
<td>40,299 ± 4,204</td>
<td>81,453</td>
<td>0.14</td>
</tr>
<tr>
<td>CAP</td>
<td>Sanger</td>
<td>138,916 ± 14,821</td>
<td>26,112</td>
<td>0.10</td>
</tr>
<tr>
<td>Total</td>
<td>N/A</td>
<td></td>
<td>20,035,171</td>
<td>22.35</td>
</tr>
</tbody>
</table>
Supplementary Table 3.2 Summary statistics of the output of the whole genome shotgun assembly prior to screening, removal of organelles and contaminating scaffolds and chromosome-scale pseudomolecule construction.

The table shows total contigs and total assembled base pairs for each set of scaffolds greater than the size listed in the left hand column.

<table>
<thead>
<tr>
<th>Minimum Scaffold Length</th>
<th>Number of Scaffolds</th>
<th>Number of Contigs</th>
<th>Scaffold Size</th>
<th>Basepairs</th>
<th>% Non-gap Basepairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Mb</td>
<td>9</td>
<td>1,945</td>
<td>83,800,702</td>
<td>83,207,536</td>
<td>99.29%</td>
</tr>
<tr>
<td>2.5 Mb</td>
<td>15</td>
<td>2,862</td>
<td>109,988,361</td>
<td>109,111,306</td>
<td>99.20%</td>
</tr>
<tr>
<td>1 Mb</td>
<td>22</td>
<td>3,496</td>
<td>120,285,947</td>
<td>119,137,479</td>
<td>99.05%</td>
</tr>
<tr>
<td>500 Kb</td>
<td>29</td>
<td>3,800</td>
<td>125,015,804</td>
<td>123,166,334</td>
<td>98.52%</td>
</tr>
<tr>
<td>250 Kb</td>
<td>36</td>
<td>4,042</td>
<td>127,275,338</td>
<td>124,934,477</td>
<td>98.16%</td>
</tr>
<tr>
<td>100 Kb</td>
<td>52</td>
<td>5,059</td>
<td>129,762,519</td>
<td>126,329,374</td>
<td>97.35%</td>
</tr>
<tr>
<td>50 Kb</td>
<td>71</td>
<td>5,841</td>
<td>131,039,977</td>
<td>127,280,728</td>
<td>97.13%</td>
</tr>
<tr>
<td>25 Kb</td>
<td>108</td>
<td>6,721</td>
<td>132,285,636</td>
<td>128,070,438</td>
<td>96.81%</td>
</tr>
<tr>
<td>10 Kb</td>
<td>211</td>
<td>7,734</td>
<td>133,843,625</td>
<td>128,977,938</td>
<td>96.36%</td>
</tr>
<tr>
<td>5 Kb</td>
<td>407</td>
<td>8,615</td>
<td>135,213,679</td>
<td>129,706,240</td>
<td>95.93%</td>
</tr>
<tr>
<td>2.5 Kb</td>
<td>730</td>
<td>9,487</td>
<td>136,359,310</td>
<td>130,194,008</td>
<td>95.48%</td>
</tr>
<tr>
<td>1 Kb</td>
<td>1,019</td>
<td>10,029</td>
<td>136,845,852</td>
<td>130,460,215</td>
<td>95.33%</td>
</tr>
<tr>
<td>0 bp</td>
<td>1,859</td>
<td>10,916</td>
<td>137,118,633</td>
<td>130,720,267</td>
<td>95.33%</td>
</tr>
</tbody>
</table>
Supplementary Table 3.3 Final summary assembly statistics for chromosome scale assembly.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold total</td>
<td>853</td>
</tr>
<tr>
<td>Contig total</td>
<td>9,675</td>
</tr>
<tr>
<td>Scaffold sequence total</td>
<td>134.8 Mb</td>
</tr>
<tr>
<td>Chromosome Sequence</td>
<td>124.9 Mb</td>
</tr>
<tr>
<td>Contig sequence total</td>
<td>130.1 Mb (3.5% gap)</td>
</tr>
<tr>
<td>Scaffold N/L50</td>
<td>4/15.1 Mb</td>
</tr>
<tr>
<td>Contig N/L50</td>
<td>265/134.1 Kb</td>
</tr>
</tbody>
</table>

Supplementary Table 3.4 Placement of the individual Fosmid clones and their contribution to the overall error rate.

<table>
<thead>
<tr>
<th>Fosmid Clone ID</th>
<th>Length</th>
<th>Scaffold</th>
<th>Start</th>
<th>Stop</th>
<th>Discrepant Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>4002707</td>
<td>38,734</td>
<td>scaffold_7</td>
<td>5,045,895</td>
<td>5,084,629</td>
<td>0</td>
</tr>
<tr>
<td>4002714</td>
<td>41,191</td>
<td>scaffold_1</td>
<td>3,212,963</td>
<td>3,254,155</td>
<td>0</td>
</tr>
<tr>
<td>4002715</td>
<td>36,241</td>
<td>scaffold_2</td>
<td>1,395,421</td>
<td>1,431,662</td>
<td>0</td>
</tr>
<tr>
<td>4002706</td>
<td>32,542</td>
<td>scaffold_6</td>
<td>13,081,692</td>
<td>13,114,232</td>
<td>4</td>
</tr>
<tr>
<td>4002703</td>
<td>38,742</td>
<td>scaffold_5</td>
<td>4,303,844</td>
<td>4,342,579</td>
<td>7</td>
</tr>
<tr>
<td>4002716</td>
<td>40,672</td>
<td>scaffold_1</td>
<td>16,845,541</td>
<td>16,887,448</td>
<td>8</td>
</tr>
<tr>
<td>4002712</td>
<td>40,852</td>
<td>scaffold_6</td>
<td>13,004,766</td>
<td>13,046,277</td>
<td>10</td>
</tr>
<tr>
<td>4002711</td>
<td>44,182</td>
<td>scaffold_8</td>
<td>11,105,345</td>
<td>11,149,515</td>
<td>12</td>
</tr>
<tr>
<td>4002717</td>
<td>29,397</td>
<td>scaffold_5</td>
<td>2,235,314</td>
<td>2,264,946</td>
<td>75</td>
</tr>
<tr>
<td>4002713</td>
<td>37,600</td>
<td>scaffold_5</td>
<td>2,928,104</td>
<td>2,965,878</td>
<td>166</td>
</tr>
<tr>
<td>4002705</td>
<td>41,085</td>
<td>scaffold_7</td>
<td>11,478,240</td>
<td>11,519,089</td>
<td>236</td>
</tr>
<tr>
<td>4002708</td>
<td>29,568</td>
<td>scaffold_6</td>
<td>5,376,721</td>
<td>5,406,090</td>
<td>477</td>
</tr>
<tr>
<td>Total</td>
<td>450,806</td>
<td></td>
<td></td>
<td></td>
<td>995</td>
</tr>
</tbody>
</table>
Supplementary Table 3.5 Placement of the individual BAC clones and their contribution to the overall error rate.

<table>
<thead>
<tr>
<th>BAC Clone ID</th>
<th>Length</th>
<th>Scaffold</th>
<th>Start</th>
<th>Stop</th>
<th>Discrepant Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>9285</td>
<td>134,688</td>
<td>scaffold 3</td>
<td>5,492,673</td>
<td>5,627,354</td>
<td>8</td>
</tr>
<tr>
<td>9279</td>
<td>170,037</td>
<td>scaffold 6</td>
<td>2,833,593</td>
<td>3,003,831</td>
<td>61</td>
</tr>
<tr>
<td>9277</td>
<td>148,953</td>
<td>scaffold 7</td>
<td>4,251,711</td>
<td>4,400,783</td>
<td>123</td>
</tr>
<tr>
<td>9282</td>
<td>125,625</td>
<td>scaffold 5</td>
<td>300,990</td>
<td>426,642</td>
<td>122</td>
</tr>
<tr>
<td>9289</td>
<td>134,212</td>
<td>scaffold 6</td>
<td>298,020</td>
<td>432,319</td>
<td>154</td>
</tr>
<tr>
<td>9296</td>
<td>134,605</td>
<td>scaffold 7</td>
<td>2,691,838</td>
<td>2,829,633</td>
<td>171</td>
</tr>
<tr>
<td>9286</td>
<td>79,708</td>
<td>scaffold 8</td>
<td>11,614,394</td>
<td>11,694,049</td>
<td>134</td>
</tr>
<tr>
<td>9297</td>
<td>155,124</td>
<td>scaffold 2</td>
<td>6,470,939</td>
<td>6,671,470</td>
<td>531</td>
</tr>
<tr>
<td>9287</td>
<td>182,653</td>
<td>scaffold 4</td>
<td>11,881,783</td>
<td>12,064,337</td>
<td>929</td>
</tr>
<tr>
<td>9291</td>
<td>128,808</td>
<td>scaffold 4</td>
<td>13,754,803</td>
<td>13,883,159</td>
<td>806</td>
</tr>
<tr>
<td>Total</td>
<td>1,394,413</td>
<td></td>
<td>3,039</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3.6 Summary statistics of *de novo* assembly of *C. grandiflora*.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Main genome scaffold total</td>
<td>30,490</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main genome contig total</td>
<td>44,658</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main genome scaffold sequence total</td>
<td>112.3 Mb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main genome contig sequence total</td>
<td>100.5 Mb (&gt; 10.5% gap)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main genome scaffold N/L50</td>
<td>56/98.1 kb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main genome contig N/L50</td>
<td>2,508/10.3 kb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of scaffolds &gt; 50 kb</td>
<td>505</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% main genome in scaffolds &gt; 50 kb</td>
<td>65.50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimum Scaffold Length</th>
<th>Number of Scaffolds</th>
<th>Number of Contigs</th>
<th>Total Scaffold Length</th>
<th>Total Contig Length</th>
<th>Scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>30,490</td>
<td>44,658</td>
<td>112,290,236</td>
<td>100,507,575</td>
<td>89.51%</td>
</tr>
<tr>
<td>100</td>
<td>30,482</td>
<td>44,650</td>
<td>112,289,603</td>
<td>100,506,942</td>
<td>89.51%</td>
</tr>
<tr>
<td>250</td>
<td>13,900</td>
<td>28,042</td>
<td>110,057,328</td>
<td>98,274,994</td>
<td>89.29%</td>
</tr>
<tr>
<td>500</td>
<td>9,248</td>
<td>22,548</td>
<td>108,377,980</td>
<td>96,679,626</td>
<td>89.21%</td>
</tr>
<tr>
<td>1 kb</td>
<td>4,997</td>
<td>17,286</td>
<td>105,346,052</td>
<td>93,753,068</td>
<td>89.00%</td>
</tr>
<tr>
<td>2.5 kb</td>
<td>2,291</td>
<td>13,773</td>
<td>101,340,709</td>
<td>98,828,933</td>
<td>88.64%</td>
</tr>
<tr>
<td>5 kb</td>
<td>1,907</td>
<td>13,174</td>
<td>100,067,047</td>
<td>88,571,904</td>
<td>88.51%</td>
</tr>
<tr>
<td>10 kb</td>
<td>1,480</td>
<td>12,175</td>
<td>96,813,647</td>
<td>86,051,737</td>
<td>88.88%</td>
</tr>
<tr>
<td>25 kb</td>
<td>886</td>
<td>9,946</td>
<td>87,093,467</td>
<td>78,265,006</td>
<td>89.86%</td>
</tr>
<tr>
<td>50 kb</td>
<td>505</td>
<td>7,758</td>
<td>73,519,643</td>
<td>66,973,813</td>
<td>91.10%</td>
</tr>
<tr>
<td>100 kb</td>
<td>247</td>
<td>5,226</td>
<td>55,293,634</td>
<td>51,089,345</td>
<td>92.40%</td>
</tr>
<tr>
<td>250 kb</td>
<td>72</td>
<td>2,309</td>
<td>28,004,659</td>
<td>26,281,287</td>
<td>93.85%</td>
</tr>
<tr>
<td>500 kb</td>
<td>8</td>
<td>314</td>
<td>6,178,508</td>
<td>5,972,560</td>
<td>96.67%</td>
</tr>
<tr>
<td>1 Mb</td>
<td>1</td>
<td>59</td>
<td>1,487,256</td>
<td>1,449,874</td>
<td>97.49%</td>
</tr>
</tbody>
</table>
**Supplementary Table 3.7 Assembly statistics for de novo assembly of N. paniculata.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds</td>
<td>6899</td>
<td></td>
</tr>
<tr>
<td>Total size of scaffolds</td>
<td>113370917</td>
<td></td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>455829</td>
<td></td>
</tr>
<tr>
<td>Shortest scaffold</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>Number of scaffolds &gt; 500 nt</td>
<td>5592</td>
<td>81.10%</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 1K nt</td>
<td>4788</td>
<td>69.40%</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 10K nt</td>
<td>2177</td>
<td>31.60%</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 100K nt</td>
<td>225</td>
<td>3.30%</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 1M nt</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Mean scaffold size</td>
<td>16433</td>
<td></td>
</tr>
<tr>
<td>Median scaffold size</td>
<td>2807</td>
<td></td>
</tr>
<tr>
<td>N50 scaffold length</td>
<td>62327</td>
<td></td>
</tr>
<tr>
<td>L50 scaffold count</td>
<td>520</td>
<td></td>
</tr>
<tr>
<td>scaffold %A</td>
<td>30.57</td>
<td></td>
</tr>
<tr>
<td>scaffold %C</td>
<td>17.45</td>
<td></td>
</tr>
<tr>
<td>scaffold %G</td>
<td>17.46</td>
<td></td>
</tr>
<tr>
<td>scaffold %T</td>
<td>30.58</td>
<td></td>
</tr>
<tr>
<td>scaffold %N</td>
<td>3.93</td>
<td></td>
</tr>
<tr>
<td>Percentage of assembly in scaffolded contigs</td>
<td>78.20%</td>
<td></td>
</tr>
<tr>
<td>Percentage of assembly in unscaffolded contigs</td>
<td>21.80%</td>
<td></td>
</tr>
<tr>
<td>Average number of contigs per scaffold</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Average length of break (&gt;25 Ns) between contigs in scaffold</td>
<td>645</td>
<td></td>
</tr>
<tr>
<td>Number of contigs</td>
<td>12044</td>
<td></td>
</tr>
<tr>
<td>Number of contigs in scaffolds</td>
<td>7054</td>
<td></td>
</tr>
<tr>
<td>Number of contigs not in scaffolds</td>
<td>4990</td>
<td></td>
</tr>
<tr>
<td>Total size of contigs</td>
<td>108924905</td>
<td></td>
</tr>
<tr>
<td>Longest contig</td>
<td>165716</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Value</td>
<td>Percentage</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Shortest contig</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>Number of contigs &gt; 500 nt</td>
<td>10266</td>
<td>85.20%</td>
</tr>
<tr>
<td>Number of contigs &gt; 1K nt</td>
<td>8947</td>
<td>74.30%</td>
</tr>
<tr>
<td>Number of contigs &gt; 10K nt</td>
<td>3216</td>
<td>26.70%</td>
</tr>
<tr>
<td>Number of contigs &gt; 100K nt</td>
<td>32</td>
<td>0.30%</td>
</tr>
<tr>
<td>Number of contigs &gt; 1M nt</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Mean contig size</td>
<td>9044</td>
<td></td>
</tr>
<tr>
<td>Median contig size</td>
<td>3189</td>
<td></td>
</tr>
<tr>
<td>N50 contig length</td>
<td>24627</td>
<td></td>
</tr>
<tr>
<td>L50 contig count</td>
<td>1259</td>
<td></td>
</tr>
<tr>
<td>contig %A</td>
<td>31.82</td>
<td></td>
</tr>
<tr>
<td>contig %C</td>
<td>18.16</td>
<td></td>
</tr>
<tr>
<td>contig %G</td>
<td>18.18</td>
<td></td>
</tr>
<tr>
<td>contig %T</td>
<td>31.83</td>
<td></td>
</tr>
<tr>
<td>contig %N</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3.8 Mapping of genic regions in *C. rubella* (CR) onto *C. grandiflora* (CG) and *N. paniculata* (NP) assemblies using liftOver with >80% identity from best orthologous chains.

<table>
<thead>
<tr>
<th></th>
<th>Count in CR</th>
<th>Number mapping to CG</th>
<th>Number Mapping to NP</th>
<th>Median size in CR (nt)</th>
<th>Median size in CG (nt)</th>
<th>Median size in NP (nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>26187</td>
<td>18630</td>
<td>16293</td>
<td>1,979</td>
<td>1,653</td>
<td>1,730</td>
</tr>
<tr>
<td>CDS</td>
<td>136278</td>
<td>123331</td>
<td>109568</td>
<td>135</td>
<td>133</td>
<td>132</td>
</tr>
<tr>
<td>Exons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introns</td>
<td>110091</td>
<td>99362</td>
<td>81440</td>
<td>102</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>Upstream regions (500nt)</td>
<td>26187</td>
<td>18422</td>
<td>11979</td>
<td>500</td>
<td>500</td>
<td>504</td>
</tr>
<tr>
<td>Downstream regions (500nt)</td>
<td>26187</td>
<td>19077</td>
<td>12064</td>
<td>500</td>
<td>500</td>
<td>506</td>
</tr>
</tbody>
</table>

Supplementary Table 3.9 Percentage of the *C. rubella* (CR) genome for which alignments are reliably suggested in the *C. grandiflora* (CG) and *N. paniculata* (NP) assemblies.

<table>
<thead>
<tr>
<th>Gapped Alignment</th>
<th>Size (nt)</th>
<th>Percentage of CR assembly mapped to CG/NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of CR assembly</td>
<td>134834574</td>
<td></td>
</tr>
<tr>
<td>Total size of CR chains (gapped) mapping to NP</td>
<td>123180187</td>
<td>91%</td>
</tr>
<tr>
<td>Total size of CR chains (gapped) mapping to CG</td>
<td>110762997</td>
<td>82%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ungapped alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of CR assembly</td>
</tr>
<tr>
<td>Total size of ungapped alignments to NP</td>
</tr>
<tr>
<td>Total size of ungapped alignments to CG</td>
</tr>
</tbody>
</table>
### Supplementary Table 3.10 Description of the ten non-chromosomal scaffolds with largest inferred genomic content.

<table>
<thead>
<tr>
<th>Scaffold</th>
<th>Observed Scaffold Size (bp)</th>
<th>Inferred Genomic Content (bp)</th>
<th>Inferred Copy Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>76047</td>
<td>3499817</td>
<td>46</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>26</td>
<td>128634</td>
<td>3139803</td>
<td>24</td>
<td>Ribosomal RNA repeat</td>
</tr>
<tr>
<td>31</td>
<td>79818</td>
<td>1965193</td>
<td>24</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>40</td>
<td>52300</td>
<td>1432816</td>
<td>27</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>499</td>
<td>3022</td>
<td>1339713</td>
<td>443</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>36</td>
<td>121157</td>
<td>1279557</td>
<td>11</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>42</td>
<td>37376</td>
<td>1269164</td>
<td>34</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>16</td>
<td>194218</td>
<td>1043964</td>
<td>5</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>11</td>
<td>314570</td>
<td>1034819</td>
<td>3</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
<tr>
<td>17</td>
<td>241894</td>
<td>951260</td>
<td>4</td>
<td>Centromere satellite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sequence</td>
</tr>
</tbody>
</table>
Supplementary Table 3.11 Ontology categories identified as significantly differentiated in expression between *C. rubella* and *C. grandiflora* using the DAVID system.

<table>
<thead>
<tr>
<th>Description</th>
<th>Number in input list</th>
<th>Number of background/reference genes</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-embryonic development</td>
<td>32</td>
<td>705</td>
<td>3.1e-09</td>
<td>3.7e-06</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>13</td>
<td>162</td>
<td>2.6e-07</td>
<td>0.0001</td>
</tr>
<tr>
<td>Floral Whorl development</td>
<td>12</td>
<td>133</td>
<td>2.1e-07</td>
<td>0.0001</td>
</tr>
<tr>
<td>Cell wall modification during multidimensional cell growth</td>
<td>6</td>
<td>29</td>
<td>1.7e-06</td>
<td>0.0005</td>
</tr>
<tr>
<td>Flower development</td>
<td>17</td>
<td>377</td>
<td>1.3e-05</td>
<td>0.0031</td>
</tr>
<tr>
<td>Reproductive development process</td>
<td>29</td>
<td>978</td>
<td>5.9e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Post-embryonic organ development</td>
<td>10</td>
<td>154</td>
<td>3.8e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Anatomical structure formation involved in morphogenesis</td>
<td>9</td>
<td>122</td>
<td>3.4e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Response to endogenous stimulus</td>
<td>31</td>
<td>1068</td>
<td>5e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Developmental process</td>
<td>54</td>
<td>2304</td>
<td>5.9e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Reproductive structure development</td>
<td>29</td>
<td>978</td>
<td>5.9e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Plant-type cell wall modification during multidimensional cell growth</td>
<td>5</td>
<td>27</td>
<td>2.3e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Plant-type cell wall modification</td>
<td>6</td>
<td>50</td>
<td>4.6e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Multidimensional cell growth</td>
<td>6</td>
<td>50</td>
<td>4.6e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Floral organ development</td>
<td>10</td>
<td>152</td>
<td>3.4e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Anatomical structure development</td>
<td>44</td>
<td>1726</td>
<td>4e-05</td>
<td>0.0044</td>
</tr>
<tr>
<td>Plant-type cell wall loosening</td>
<td>5</td>
<td>40</td>
<td>0.0001</td>
<td>0.011</td>
</tr>
</tbody>
</table>
Supplementary Table 3.12 Candidate differentially expressed genes falling in floral QTL regions that have been mapped in interspecific crosses between *C. grandiflora* and *C. rubella*.

<table>
<thead>
<tr>
<th>Candidate genes under QTL</th>
<th>QTL</th>
<th><em>C. rubella</em> ID</th>
<th>Scaffold() and Position</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G70560</td>
<td>Petal size</td>
<td>Carubv10020400m</td>
<td>(2) 10071409-10074104</td>
<td>Essential for the auxin synthesis pathway. Responsible for flower and gynococcus development.</td>
</tr>
<tr>
<td>AT1G20190</td>
<td>Petal size</td>
<td>Carubv10010090m</td>
<td>(1) 7036769-7038449</td>
<td>Responsible for cell morphogenesis. Causes loosening and extension of plant cell walls.</td>
</tr>
<tr>
<td>AT5G09790</td>
<td>Pollen number</td>
<td>Carubv10001176m</td>
<td>(6) 2981770-2983853</td>
<td>Role in cell cycle or DNA replication. Involved in pollen development.</td>
</tr>
<tr>
<td>AT1G63180</td>
<td>Pollen number</td>
<td>Carubv10020560m</td>
<td>(2) 718093-720820</td>
<td>Role in cell envelope biogenesis. Involved in pollen and gametophyte development.</td>
</tr>
</tbody>
</table>

Supplementary Table 3.13 Genes in the brassinosteroid pathway showing changes in gene expression

<table>
<thead>
<tr>
<th><em>Arabidopsis</em> orthologue</th>
<th>Evidence of differential expression from orthology analysis?</th>
<th>Change in expression in <em>C. rubella</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT5G22260 (MS1)</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT4G28395</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT3G42960</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT3G51590</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT1G07340</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT2G18550</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>AT1G61110</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT5G62320</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>AT5G61430</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>AT3G11980</td>
<td>Yes</td>
<td>Down</td>
</tr>
<tr>
<td>AT2G16910</td>
<td>Yes</td>
<td>Down</td>
</tr>
</tbody>
</table>
Supplementary Table 3.14 *Capsella* samples used in RNAseq population genetics analysis.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sample Name</th>
<th>Geographic origin</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. rubella</em></td>
<td>CrTAAL</td>
<td>Taguemont, Algeria</td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>Cr1377</td>
<td>Buenos Aires, Argentina</td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>Cr1GR1</td>
<td>Samos, Greece</td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>Cr81.02</td>
<td>Megalopoli, Greece</td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>Cr75.2</td>
<td>Kalavrita, Greece</td>
</tr>
<tr>
<td><em>C. rubella</em></td>
<td>Cr34.11</td>
<td>San Luca, Italy</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>Cg88.23</td>
<td>Monodendri, Greece</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>Cg103.02</td>
<td>Pili, Greece</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>Cg85.9</td>
<td>Metsovo, Greece</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>CgAxF</td>
<td>F1 of Cg2e_TS1 from Paleokastritsas, Corfu, Greece and Cg918_8_TS1 from Pantokrator, Corfu, Greece</td>
</tr>
<tr>
<td><em>C. grandiflora</em></td>
<td>Cg5a</td>
<td>Troumpeta, Corfu, Greece</td>
</tr>
</tbody>
</table>
Supplementary Table 3.15 Proportions of new nonsynonymous mutations in bins with different negative fitness effects.

For comparison estimates under a model of constant population size.

<table>
<thead>
<tr>
<th></th>
<th>Ne &lt; 1</th>
<th>Ne 1-10</th>
<th>Ne 10-100</th>
<th>Ne &gt;100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step change in Ne</td>
<td>0.063237</td>
<td>0.09263</td>
<td>0.222715</td>
<td>0.621417</td>
</tr>
<tr>
<td>Constant Ne</td>
<td>0.056065</td>
<td>0.120077</td>
<td>0.342698</td>
<td>0.481160</td>
</tr>
</tbody>
</table>

Supplementary Table 3.16 Parameter estimates for three demographic models of the split between *C. grandiflora* and *C. rubella*.

Assuming a mutation rate of $7 \times 10^{-9}$ bp$^{-1}$ generation$^{-1}$. For the split growth model, we estimate that 2.3% of the ancestral population contributed to the formation of *C. rubella*, and estimates for effective population size of *C. grandiflora* and *C. rubella* indicate current sizes.

<table>
<thead>
<tr>
<th>Model</th>
<th>LogL</th>
<th>AIC</th>
<th>NeAnc</th>
<th>NeCg</th>
<th>NeCr</th>
<th>Tsplit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split Growth</td>
<td>-268.22</td>
<td>544.44</td>
<td>111,433</td>
<td>223,588</td>
<td>1,308</td>
<td>96,134</td>
</tr>
<tr>
<td>Instant Change</td>
<td>-287.98</td>
<td>581.97</td>
<td>108,846</td>
<td>163,183</td>
<td>7,613</td>
<td>131,946</td>
</tr>
<tr>
<td>Instant Change in Cr only</td>
<td>-333.44</td>
<td>670.87</td>
<td>115,612</td>
<td>115,612</td>
<td>6,726</td>
<td>115,340</td>
</tr>
</tbody>
</table>
Appendix References


