MOLECULAR POPULATION GENETIC CONSEQUENCES OF EVOLUTIONARY TRANSITIONS FROM OUTCROSSING TO SELFING IN PLANTS

by

Robert W. Ness

A thesis submitted in conformity with the requirements for the degree of

Doctorate of Philosophy

Graduate Department of Ecology and Evolutionary Biology

University of Toronto

© Copyright by Robert W. Ness 2011
The transition from cross-fertilization to predominant self-fertilization is considered the most common evolutionary transition in flowering plants. This change in mating system has profound influences on the amounts and patterns of genetic diversity within and among populations, and on key genetic and demographic processes. The main goal of my thesis is to determine the molecular population genetic consequences of this transition in the annual neotropical aquatic plant *Eichhornia paniculata* (Pontederiaceae) using DNA sequence from individuals sampled from throughout the species’ geographic range. Populations exhibit a wide range of mating patterns associated the evolutionary breakdown of tristyly facilitating specific contrasts between outcrossing and selfing populations.

Analysis of molecular variation supported the hypothesis of multiple origins of selfing, including the evolution of two morphologically distinct selfing variants from Central America and the Caribbean. A survey of 10 nuclear loci from 225 individuals sampled from 25 populations demonstrated the joint influence of mating system, population size and demographic bottlenecks in affecting patterns of nucleotide variation. Small selfing populations exhibited significantly lower genetic diversity compared with larger outcrossing and mixed mating populations. There was also evidence for higher population differentiation and a slower decay of linkage disequilibrium in predominately selfing populations from the
Caribbean region. Coalescent simulations of the sequence data indicated a bottleneck
associated with colonization of the Caribbean from Brazil ∼125,000 years ago.

To investigate the consequences of transitions from outcrossing to selfing across the
genome, I used high-throughput, short-read sequencing to assemble ∼27,000 ESTs
representing ∼24Mbp of sequence. Characterization of floral transcriptomes from this dataset
identified 269 genes associated with floral development, 22 of which were differentially
expressed in three independently derived selfing lineages compared to an outcrossing
genotype. Evidence for relaxed selection in selfing lineages was obtained from an analysis of
a subset of ∼8000 orthologous sequences from each genotype, as predicted by theory. Selfing
genomes showed an increase in the proportion of nonsynonymous to synonymous changes
and relaxation of selection for codon usage bias. My thesis represents the most detailed
investigation to date of the molecular population genetic consequences of intraspecific
variation in the mating systems of plants.
ACKNOWLEDGEMENTS

Throughout my doctoral studies I received invaluable advice and support from many people, without which the research presented in this thesis would not have been possible. First, I would like to thank my supervisor Spencer Barrett for his loyalty and dedication to ensuring the success of my dissertation, and for continued financial support. His guidance has helped greatly in my scientific and intellectual development. I am grateful to John Willis (Duke University) for generously offering to serve as my external examiner. My supervisory committee David Guttman, Nicholas Provert and Stephen Wright have always been available for discussion, and provided critical advice towards many aspects of my project. I would also like to thank Asher Cutter, John Stinchcombe, Jannice Friedman, Rob Colautti, and Aneil Agrawal for valuable insights and conversation over the years. Thanks also to Pauline Wang and John Stavrinides both of whom provided invaluable laboratory advice, and to Justin Giancola for his patient computational assistance. My thesis has benefitted greatly from the fruitful collaborations that I have had with Sean Graham (Univ. of British Columbia), Stephen Wright, Paul Foxe and Mathieu Siol. I also thank Spencer Barrett, Neil White, Suzanne Barrett and Bill Cole for assistance with field collections of Eichhornia paniculata in Brazil and the Caribbean, and Sean Graham for DNA samples of Pontederiaceae and for hosting my visit to Vancouver.

Several people have made my time at the University of Toronto a fun and stimulating place to do graduate training, including Sarah Yakimowski, Bill Cole, Marc Johnson, and Honour McCann and Danielle Marcos. I am also indebted to the whole Barrett lab, past and present, for valuable contributions to my research and for advice on presentations during my graduate work. Finally, I am forever grateful to my parents Janet and Gary Ness for their support and encouragement throughout my graduate and early education. I also owe Rosalind
Murray my deepest gratitude for organizing my life, and for providing partnership, support, fun and distractions, especially over the final months of my thesis.

The Natural Sciences and Engineering Research Council of Canada, Frederick P. Ide Graduate Award in EEB, and the University of Toronto provided financial support for my doctoral studies. The research in this thesis was supported by NSERC Discovery Grants and funding through the Canada Chair’s Program and a Premier’s Discovery Award in Life Sciences and Medicine from the Ontario Government to Spencer C.H. Barrett.
# TABLE OF CONTENTS

Abstract ..................................................................................................................................... ii

Acknowledgements .................................................................................................................. iv

Table of Contents ..................................................................................................................... vi

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

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

CHAPTER ONE: GENERAL INTRODUCTION ................................................................... 1

Mating system variation and evolution ............................................................................. 2

Molecular approaches to investigating the evolution of selfing ....................................... 5

_Eichhornia paniculata_ as a study system .......................................................................... 9

Research objectives ......................................................................................................... 12

CHAPTER TWO: GENOMIC CONSEQUENCES OF OUTCROSSING AND SELFING IN PLANTS ................................................................................................................................. 21

Summary ......................................................................................................................... 21

Introduction ..................................................................................................................... 22

Inbreeding and the efficacy of natural selection ............................................................. 25

Evolution of recombination rate ..................................................................................... 30

Relations between mating system and genomic mutation rate ....................................... 34

Mating systems and genetic conflicts ............................................................................. 37

Transposable elements ................................................................................................. 38

B-Chromosomes ............................................................................................................ 41

Biased gene conversion ............................................................................................... 42

Cytoplasmic male sterility .......................................................................................... 45

Evolution of genome size .............................................................................................. 47
Conclusions and future directions

CHAPTER THREE: EVOLUTIONARY PATHWAYS TO SELF-FERTILIZATION IN A TRISTYLOUS PLANT SPECIES

Summary
Introduction
Evolutionary pathways to selfing
Multiple origins of selfing
Evidence from genetic relations among populations
Evidence from genetic control of mating-system modification
Developmental instability in the formation of selfing flowers
Discussion

CHAPTER FOUR: MATING-SYSTEM VARIATION, DEMOGRAPHIC HISTORY AND PATTERNS OF NUCLEOTIDE DIVERSITY IN THE TRISTYLOUS PLANT

EICHHORNIA PANICULATA

Summary
Introduction
Methods
Population sampling
Marker development
Amplification and sequencing
Polymorphism
Correlates of nucleotide diversity
Population genetic structure
Linkage disequilibrium
Demographic history
Results ............................................................................................................................. 93
Polymorphism .................................................................................................................... 93
Correlates of nucleotide diversity ..................................................................................... 93
Population structure ........................................................................................................ 95
Linkage disequilibrium ..................................................................................................... 96
Coalescent simulations .................................................................................................... 97
Discussion ........................................................................................................................ 98
Population size, morph structure and nucleotide polymorphism ........................................ 99
Timescales and demographic history .................................................................................. 102
Comparing diversity in outcrossing and selfing populations ........................................... 104

CHAPTER FIVE: DE NOVO SEQUENCE ASSEMBLY AND CHARACTERIZATION OF THE FLORAL TRANSCRIPTOME IN CROSS- AND SELF-FERTILIZING PLANTS .. 118
Summary ........................................................................................................................... 118
Introduction ..................................................................................................................... 119
Methods ............................................................................................................................ 122
Sampling and RNA preparation ....................................................................................... 122
Sequencing ....................................................................................................................... 123
De novo assembly ............................................................................................................ 123
Consensus transcriptome generation ................................................................................. 124
Genotype calling and SNP detection ................................................................................ 126
Measurement of gene expression ..................................................................................... 127
Functional annotation ..................................................................................................... 127
Assessment of the accuracy of EST assembly ................................................................. 128
Results ............................................................................................................................. 129
Assembly and consensus transcriptome generation ......................................................... 129
Genotype calling and SNP detection ................................................................................ 130
CHAPTER SIX: GENOMIC CONSEQUENCES OF TRANSITIONS FROM CROSS- TO SELF-FERTILIZATION ON THE EFFICACY OF SELECTION IN THREE INDEPENDENTLY DERIVED SELFING PLANTS ......................................................... 154

Summary ....................................................................................................................... 154

Introduction................................................................................................................... 155

Methods......................................................................................................................... 160

Samples, tissue preparation and sequencing ............................................................. 160

EST analysis, filtering and alignments ...................................................................... 161

Calculation of non-synonymous and synonymous mutations ................................. 161

MK tests of selection ................................................................................................ 162

Assessing selection for codon bias ........................................................................... 163

Phylogenetic inference of selection on codon bias ................................................... 165

Results........................................................................................................................... 166

Selection on non-synonymous mutations ................................................................. 167

MK tests of selection ................................................................................................ 168

Assessing selection for codon bias ........................................................................... 169

Phylogenetic inference of selection on codon bias ................................................... 170
Discussion ..................................................................................................................... 171
Mating system and protein evolution........................................................................ 172
MK tests and evidence for purifying selection ......................................................... 174
Mating system and selection for codon bias............................................................. 175
Factors influencing the efficacy of selection in selfers.............................................. 179

CHAPTER SEVEN: RECONCILING GENE AND GENOME DUPLICATION EVENTS:
USING MULTIPLE NUCLEAR GENE FAMILIES TO INFER THE PHYLOGENY OF
THE AQUATIC PLANT FAMILY PONTEDERIACEAE ................................................. 191

Summary ....................................................................................................................... 191
Introduction ................................................................................................................... 192

Methods ......................................................................................................................... 196
Taxon sampling ............................................................................................................. 196
Primer development and sequencing ......................................................................... 196
Reconstructing genealogical and phylogenetic history .............................................. 198
Rooting the phylogeny of Pontederiaceae ................................................................. 199
Gene duplication and polyploidy ............................................................................... 200

Results ........................................................................................................................... 201
Nuclear-based species tree ......................................................................................... 201
Combined nuclear and plastid species tree ............................................................... 202
Rooting the Pontederiaceae phylogeny ................................................................. 203
Gene duplication and polyploidy ............................................................................... 203

Discussion ..................................................................................................................... 204
Nuclear-based species tree ......................................................................................... 204
Combined nuclear and plastid species tree ............................................................... 207
Rooting the Pontederiaceae species tree ................................................................. 208
Gene duplication and polyploidy ............................................................................... 210
Conclusions.......................................................................................................................... 211

CHAPTER EIGHT: CONCLUDING DISCUSSION ................................................................. 219

Summary of thesis chapters ................................................................................................. 219

Future research...................................................................................................................... 223

Architecture of the selfing syndrome................................................................................ 223

Genetics of adaptation........................................................................................................ 225

Population genomics of mating system .......................................................................... 228

Conclusion .......................................................................................................................... 229

LITERATURE CITED ............................................................................................................. 231
LIST OF TABLES

Table 3.1. Style morph frequencies and the size of *Eichhornia paniculata* populations on the islands of Jamaica and Cuba ................................................................. 75

Table 4.1. Population details of *Eichhornia paniculata* used in Chapter 4 ....................... 106

Table 4.2. Summary statistics of polymorphism for the 10 nuclear loci sequenced in *Eichhornia paniculata* in Chapter 4 ................................................................. 107

Table 4.3. Tajima’s $D$ and $\theta_W$ values for populations of *Eichhornia paniculata* ........... 108

Table 4.4. The statistical relations between population size, the frequency of selfing variants and nucleotide diversity in 25 populations of *Eichhornia paniculata* .......... 109

Table 5.1 Sequencing and assembly statistics from previously published *de novo* transcriptome assemblies ................................................................. 140

Table 5.2. Short-read sequencing statistics for Chapter 5 .................................................... 141

Table 5.3. Summary statistics for reference transcriptome through progressive stages of assembly ......................................................................................... 142

Table 5.4. Results of read mapping using the original trimmed short reads for each of the four *Eichhornia* samples against the reference transcriptome .......................... 143

Table 5.5. Number of single nucleotide polymorphisms between pairs of *Eichhornia* genotypes ......................................................................................... 144

Table 5.6. Pollen and flower development genes that were differentially expressed in both independently derived *Eichhornia paniculata* selfing genotypes and were also identified from selfing *Eichhornia paradoxa* .................................................. 145

Table 6.1. Summary statistics from the dataset on synonymous and non-synonymous sites across all loci of *Eichhornia* ................................................................. 178

Table 6.2. Results from PAML tests of reduced selection efficacy in selfers of *Eichhornia* at the protein level with the genes concatenated ......................................... 179
Table 6.3. Results of the multi-locus MK tests for different models with or without purifying and positive selection on genotypes of *Eichhornia*. ..................................................... 180

Table 6.4. Mean values of expression, gene length, frequency of optimal codon usage and base composition........................................................................................................... 181

Table 6.5. Maximum-likelihood estimates of selection coefficients on codon usage bias in *Eichhornia* genotypes from the program codonbias..................................................... 182

Table 7.1. Species from the family Pontederiaceae chosen for Chapter 7............................ 208

Table 7.2. Summary information about six ESTs used to design primers for amplifying nuclear regions in Pontederiaceae................................................................................. 209
LIST OF FIGURES

Figure 1.1 Images of *Eichhornia paniculata* habitat, inflorescence and a contrast between outcrossing and selfing flowers .............................................................. 18

Figure 1.2 Evolutionary pathway from cross-fertilization to self-fertilization in tristylos *Eichhornia paniculata* via the breakdown of tristyly to semi-homostyly ................. 19

Figure 1.3. Variation in multi-locus outcrossing rate among populations of *Eichhornia paniculata* in N.E. Brazil and Jamaica in relation to population size and style morph diversity........................................................................................................................... 20

Figure 2.1. Expected effects of outcrossing and selfing on patterns of genetic variation and molecular evolution ........................................................................................................ 54

Figure 2.2. Decay of linkage disequilibrium with physical distance in the highly selfing *Arabidopsis thaliana* and the self-incompatible *A. lyrata.* ............................................. 55

Figure 2.3. C-value comparisons for pairs of outcrossing and selfing congeners. ............ 56

Figure 3.1. Evolutionary pathways from cross-fertilization to self-fertilization in tristylos *Eichhornia paniculata* via the origins of semi-homostyly ............................................. 76

Figure 3.2. Images of two selfing variants of *Eichhornia paniculata* illustrating their contrasting positions of sexual organs ............................................................................ 77

Figure 3.3. Neighbour network of 10 nuclear loci displaying relationships among 27 populations of *Eichhornia paniculata* sampled throughout the geographical range.. .... 78

Figure 3.4. Pairwise distribution of single nucleotide polymorphisms from 10 nuclear loci from outcrossing and selfing populations of *Eichhornia paniculata* .................... 79

Figure 3.5. Inheritance of stigma-anther separation in the L- and M-morphs of *Eichhornia paniculata* ....................................................................................................................... 80

Figure 3.6. Developmental instability of stigma-anther separation in *Eichhornia paniculata* from N.E. Brazil.............................................................................................................. 81

Figure 4.1. Map of *Eichhornia paniculata* populations sampled for Chapter 4 ............... 110
Figure 4.2. Estimates of within-population nucleotide diversity ($\theta_\pi$) for 25 populations of *Eichhornia paniculata* ................................................................. 111

Figure 4.3. Joint maximum likelihood surface for nucleotide polymorphism ($\theta_W$) for combined Caribbean and Brazilian populations ........................................ 112

Figure 4.4. Relation between census population size and nucleotide diversity in 25 populations of *E. paniculata* from Brazil and the Caribbean ............................ 113

Figure 4.5. Range-wide genetic structure of *Eichhornia paniculata* based on analyses conducted using InStruct with different numbers of clusters ................................ 114

Figure 4.6. Mean $F_{st}$ for pairwise population comparisons of *Eichhornia paniculata* within geographical regions. ................................................................. 115

Figure 4.7. Decay of linkage disequilibrium with distance in a scattered sample of *Eichhornia paniculata* from Brazil and the Caribbean .................................. 116

Figure 4.8. Marginal posterior probability distributions from coalescent simulations of the demographic history of *Eichhornia paniculata*. ........................................ 117

Figure 5.1. Histogram of the frequency of different contigs sizes in transcriptome assemblies of *Eichhornia* samples ................................................................. 146

Figure 5.2. Proportion of each of the assembled sequences that is aligned to its top BLASTx protein hit versus the proportion of the top hit which is covered by the assembled query sequence ......................................................................................... 147

Figure 5.3. Distribution of genes in the transcriptome assembly assigned to broad GO categories, cellular components and biological processes ...................................... 148

Figure 5.4. Pairwise correlations of gene expression between the four genotypes: *Eichhornia paradoxa* and *Eichhornia paniculata* - Brazil, Jamaica and Nicaragua. .................. 149

Figure 6.1. Counts of synonymous and non-synonymous derived mutations in selfing and outcrossing genotypes of *Eichhornia* ......................................................... 183

Figure 6.2. Heat map of $\Delta$RSCU values of four genotypes of *Eichhornia* in Chapter 6........ 184
Figure 6.3. Comparisons of codon bias, measured as $\Delta \text{RSCU}_+^+$ among each of the four
Eichhornia genotypes and the three selfing genotypes combined............................. 185

Figure 6.4. Schematic representation of estimates of the selection coefficient, $S$, for preferred
codon usage in Eichhornia genotypes, using the maximum likelihood program
codonbias .............................................................................................................................. 186

Figure 7.1. Maximum-likelihood reconstructions of each of five nuclear gene trees used in
Chapter 7 .................................................................................................................................. 210

Figure 7.2. Rooted species tree for the Pontederiaceae inferred in DupTree using five nuclear
genealogies contrasted with unrooted plastid genealogy from Graham et al. (2002)... 211

Figure 7.3. One of four most parsimonious species trees based on simultaneous reconciliation
of five nuclear gene trees and a plastid gene tree .............................................................. 212

Figure 7.4. Species tree of the Pontederiaceae based on combined nuclear and plastid
sequences presented as an unrooted cladogram showing reconciliation costs for root
placements on each branch ................................................................................................. 213

Figure 7.5. Polyploidization and gene duplication events mapped onto the species tree of
Pontederiaceae inferred using five nuclear gene trees and the plastid gene tree ....... 214
"The evolutionary pathway from obligate outcrossing based upon self-incompatibility to predominant self-fertilization has probably been followed by more different lines of evolution in flowering plants than has any other" – G. Ledyard Stebbins (1970)

The reproductive structures of plants are among the most impressive examples of diversity in biological form and function. This is largely because plants are sessile and rely on pollen vectors to transmit gametes and effect cross-fertilization. Evolution to promote efficient pollen dispersal among plants has resulted in diversification in the reproductive biology of angiosperms. Moreover, because hermaphroditism is the most common condition in angiosperms (Yampolsky and Yampolsky 1922), this sexual condition provides opportunities for both cross- and self-fertilization. Selection on the proportion of seed sired by cross- versus self-fertilization, commonly referred to as the mating system, is one of the main driving forces in floral evolution (reviewed in Barrett and Harder 1996). A profusion of floral forms, mating strategies and genetic mechanisms function to promote efficient cross-fertilization and limit the harmful genetic and fitness consequences of self-fertilization (Charlesworth and Charlesworth 1987). However, despite the well-established costs of inbreeding the shift from outcrossing to selfing is considered to be the most common evolutionary transition in the flowering plants (Stebbins 1970). In numerous taxa of herbaceous plants, selfing populations have arisen from outcrossing progenitors. Understanding the genetic and genomic consequences of this transition is one of the central themes of my thesis.
Mating system variation and evolution

Stebbin’s quote underlines the importance of the evolution of selfing in plants and over a century of research has been taken up with asking how and why selfing evolves. Darwin devoted considerable effort to this problem and was particularly interested in the extent to which selfing and outcrossing affected plant fitness. His book “The Effects of Cross and Self Fertilisation in the Vegetable Kingdom” (1876) established for the first time the pervasive effects of inbreeding depression that result from selfing plants that were normally outcrossed. Since then the evolutionary dynamics of mating-system variation have been subject to sustained interest based on both theoretical and empirical studies (reviewed in Barrett and Eckert 1990; Goodwillie et al. 2006). The motivation for this body of work lies not only in explaining the ecological mechanisms that favour the evolution of selfing, but also in understanding the effects that selfing has on genetic transmission (Nagylaki 1976), sex allocation (Charnov 1982) quantitative genetic variation (Charlesworth and Charlesworth 1995a), population genetic structure (Hamrick and Godt 1996; Charlesworth and Pannell 1999) and selection response (Charlesworth 1992). The two main explanations for why selfing evolves are: (1) the genetic transmission advantage experienced by selfers (Fisher 1941), and, (2) the ability of selfing variants to reproduce when access to pollinators or compatible mates is limited (Baker 1955). Much of the literature can be coarsely grouped into these two major traditions, focusing on the population genetics of selfing or the ecological and demographic explanations for the evolution of selfing (reviewed in Cheptou and Schoen 2007). Although of course both ecological and genetic factors are important for understanding this evolutionary transition.
Fisher (1941) first demonstrated the intrinsic genetic benefit of self-fertilization. His simple argument was that if a hermaphrodite individual carried a gene allowing it to self-fertilize, then the gene would incur a 3:2 transmission advantage. This is because on average it would pass on one gene copy via cross-fertilization as a paternal parent, and two additional copies as a maternal parent as a result of self-fertilization. Thus, all things being equal the gene would spread to fixation if not opposed by other forces. The second explanation for the evolution of selfing is known as the ‘reproductive assurance hypothesis’ (Jain 1976) and was first proposed by Darwin (1876). This explanation predicts that selfing commonly evolves as a result of unreliable pollinator service or a scarcity of potential mates. In both cases plants are pollen limited and any variant capable of self-pollination should be favored. There is considerable biogeographical evidence that predominantly selfing populations occupy marginal and/or ephemeral habitats (Brown and Burdon 1987), where the proximity to mates and pollinator services vary considerably owing to demographic conditions (reviewed in Lloyd 1980). For example, evidence for this mechanism comes from the high frequency of self-compatible plants on islands (Barrett 1996). Baker’s Rule (1955) proposes that self-compatible plants will be favored in establishment following long-distance dispersal to islands because they can, in principle, initiate a sexually reproducing population from a single seed. Darwin (1876) was of the opinion that most selfing evolved through reproductive assurance and today this hypothesis is most generally supported (reviewed in Eckert et al. 2006). Indeed, although Fisher’s automatic selection hypothesis is commonly discussed there is remarkably little data demonstrating that it is the primary driver of the evolution of selfing. It seems plausible that both mechanisms may play a role in the transition from outcrossing to selfing.
Despite the advantages to selfing there are several costs that may explain why self-fertilization is not more common. Most important is the sharp reduction in fitness of selfed relative to outcrossed progeny, a phenomenon known as inbreeding depression. When inbreeding depression is greater than 0.5, that is, inbred progeny are less than 50% as fit as outcrossed progeny, the automatic transmission advantage of selfing disappears and outcrossing is maintained. However, there is a dynamic relation between the selfing rate and inbreeding depression mediated by the process of purging in which the deleterious mutations that cause inbreeding depression are selectively removed from the population. Theoretical models demonstrate that to understand the transition from outcrossing to selfing it is important to understand the joint evolution of inbreeding depression and the selfing rate of populations (Lloyd 1979; Lande and Schemske 1985, Uyenoyama and Waller 1991) and these models have stimulated a huge amount of empirical work on the measurement of these parameters (reviewed in Goodwillie et al. 2005). The effect of selfing on relative fitness through inbreeding depression throughout the life cycle is now well established (Husband and Schemske 1996) and it is widely appreciated that inbreeding depression plays a major role in determining the dynamics of mating-system evolution (Charlesworth and Charlesworth 1987). Nevertheless, the underlying genetic architecture of inbreeding depression is less well characterized and the specific ecological mechanisms driving the evolution of selfing are still poorly understood.

Beginning with Stebbins (1957), high levels of selfing have often been considered an evolutionary dead end with selfing species commonly located at the tips of phylogenies and often considered to be short-lived and doomed to extinction (Schoen et al. 1997; Takebayashi and Morrell 2001). The evolution of predominant selfing is almost always associated with a suite of floral traits promoting autonomous self-pollination and high levels of autogamy. The
selfing syndrome is commonly associated with small unshowy flowers, a loss of nectar, reduced pollen production and less well developed floral displays compared with outcrossing progenitors (Lloyd 1965; Morgan and Barrett 1989; Ritland and Ritland 1989; Armbruster et al. 2002). It is unlikely that species which have lost outcrossing adaptations and now possess small selfing flowers could revert back to outcrossing, especially in cases where selfing is largely prevented by self incompatibility (Igic et al. 2008; 2010). However, a few cases are known in which outcrossing has re-established from selfing; however, in these examples predominant selfing was not the endpoint of the initial transition and outcrossing has arisen from populations that are perhaps better characterized as possessing mixed mating (e.g. Barrett and Shore 1987; Olmstead 1990). Several population genetic processes discussed in the next section could reduce the evolutionary potential of selfing taxa by decreasing the amounts of genetic variation and recombination and making populations more vulnerable to mutational decay (reviewed in Charlesworth and Wright 2001). Understanding the interplay between the demographic factors resulting from ecological conditions favoring selfing, and the genetic consequences of increased selfing rates is one of the major motivations for the research in this thesis.

Molecular approaches to investigating the evolution of selfing

Mating systems influence the vertical transmission of heritable variation between generations but also the horizontal transfer among populations through gene flow. The principle genetic consequence of the mating system is that homozygosity at loci in the genome increases as a function of the selfing rate. Increased homozygosity leads to several other changes in the distribution of genetic variation within individuals and among populations and species. Largely homozygous populations tend to carry fewer distinct alleles
reducing effective population size \((N_e)\) (Charlesworth et al. 1993a; Nordborg 2000). In addition, although recombination can occur, its effective rate is lower because crossing over between heterozygous sites is restricted (Nordborg 2000). Therefore, effective population size can be reduced significantly by genetic hitchhiking, including selective sweeps of beneficial mutations and background selection acting against deleterious mutations (reviewed in Charlesworth and Wright 2001). Linkage among weakly selected sites with opposing selective forces also interferes with the ability of selection to act efficiently (McVean and Charlesworth 2000). As a result of these processes, reductions in \(N_e\) are often much greater than the expected two-fold decrease based on selfing alone. In addition, because of the life-history characteristics associated with selfing [e.g. many plants with high selfing rates are annual (Barrett et al. 1996) and many are successful colonizers (Brown and Burdon 1987) populations are often ephemeral, subdivided and prone to frequent genetic bottlenecks and isolation. These demographic processes also act to reduce \(N_e\) and measures of \(N_e\) in populations of selfing species are commonly much reduced in comparison with those of outcrossers (Schoen and Brown 1991).

One of the primary theoretical predictions of the low \(N_e\) values associated with higher selfing rates is a reduction in the amount of genetic variation within populations and strong differentiation among populations. Early studies during the biosystematics era established that mating systems were indeed a strong determinant of the patterns of phenotypic variation within and among plant populations (Baker, 1953). Inbreeding species generally contained relatively low levels of variation within populations but displayed high differentiation among populations. In contrast, populations of outbreeders maintained more variation but were less differentiated from one another. These early generalizations have now been corroborated theoretically (Charlesworth and Charlesworth, 1995) and are supported by surveys of diverse
species using allozyme markers (reviewed in Hamrick and Godt, 1996) Although the total molecular diversity of inbreeders is usually only slightly lower than outbreeders, differences are more dramatic at the population level with inbreeders maintaining, on average, about half of the diversity found in outcrossers. Population structure is more pronounced in inbreeders and this is reflected by significantly higher values of $G_{ST}$ and $F_{ST}$, both measures of population subdivision.

More recent molecular studies have also provided further evidence for reduced nucleotide diversity and greater differentiation among populations of selfing taxa compared to populations of related outcrossing taxa [e.g. *Leavenworthia* (Liu et al. 1998; 1999), *Arabidopsis* (Savolainen et al. 2000; Wright et al. 2002), *Solanum* (Baudry et al. 2001), *Mimulus* (Swiegart and Willis 2003), *Amsinckia* (Perusse and Schoen 2004) and *Capsella* (Foxe et al. 2009; Gou et al. 2009)]. In each case the reduction in diversity was more severe than the two-fold reduction predicted for selfing populations at equilibrium. This indicates that factors in addition to the mating system are reducing diversity, but it has been difficult to uncouple the relative importance of genetic hitchhiking from the ecology and demographic history of selfing populations. These studies of selfing in plants have generally focused on either small samples from a large number of populations or relatively large within-population samples from a small number of populations. Ideally, a deeper sampling both within and among populations combined with independent ecological and historical information is required to improve understanding of the interplay of demography, selection and mating system on shaping diversity. An effort has been made in my studies to achieve this goal by extensive sampling of a plant species in which considerable ecological and demographic information is available.
The reduction of $N_e$ resulting from the transition to selfing is also expected to reduce the efficacy of selection, which is determined by both genetic drift and selection ($N_e s$). With a decrease in the efficacy of natural selection more slightly deleterious mutations and fewer beneficial mutations are predicted to reach fixation. However, recent theoretical work indicates that we may only expect these consequences when there is a large proportion of slightly deleterious, additive mutations (Glemín 2007). If most mutations are strongly selected the reduction in $N_e$ due to selfing may not be sufficient to shift them into the effectively neutral category. In addition, increased homozygosity due to selfing exposes recessive mutations, allowing more efficient selection for advantageous mutations and against deleterious mutations. To date there has been little empirical evidence for an accumulation of deleterious mutations in selfers despite these theoretical predictions (Wright et al. 2002; Glémin et al. 2006; Cutter et al. 2008; Haudry et al. 2008; Wright et al. 2008; Escobar et al. 2010). Several explanations have been proposed for the failure to detect an accumulation of deleterious mutations in selfing lineages. These include: insufficient time since the origin of selfing, insufficient sampling of loci, the confounding effects of interspecific comparisons, and arguments relating to the distribution of fitness effects of mutations and their dominance effects, as mentioned above. At this stage the relative importance of these various explanations is unclear and our overall understanding of the effects of mating system on patterns of selection on the genome remain quite rudimentary despite considerable theoretical work in this area. In an effort to provide further insights into these problems I use the aquatic plant species *Eichhornia paniculata* as a study system for investigating the genetic and genomic consequences of mating-system variation. Below I provide a short summary of the natural history of the species and review its salient life history features. I also provide a short summary of the Pontederiaceae, the family in which *E.*
*Eichhornia paniculata* occurs, because one of the chapters in my thesis focuses on this taxon as a broader macroevolutionary context for the largely population-level investigations that form the basis of my thesis.

**Eichhornia paniculata as a study system**

The aquatic flowering plant family Pontederiaceae (Monocotyledoneae: Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009) is composed of 35-40 species from four major genera: *Heteranthera* (10-12 spp.), *Eichhornia* (8-9 spp.), *Pontederia* (6 spp.) and *Monochoria* (7-8 spp.), plus two to five widely recognized segregates: *Eurystemon* (1 sp.), *Hydrothrix* (1 sp.), *Reussia* (2-3 spp.) *Scholleropsis* (1 sp.) and *Zosterella* (1 sp.). Most species are native to the New World tropics with a few restricted to Africa, Tropical Asia and Australia (Barrett 1978, 2004). Members of the family display a wide range of life history and reproductive strategies, ranging from highly clonal, long-lived taxa of lakes, large river systems and extensive wetlands, to exclusively sexual species that are annual and occur in ephemeral pools, ditches and rice fields. Linking these extremes are species with various combinations of sexual and asexual reproduction and a variety of sexual strategies and mating systems. Evolutionary studies of the family over the past two decades have focused primarily on selected taxa of *Eichhornia* and *Pontederia* which possess tristyloous and homostyloous reproductive systems (reviewed in Barrett and Anderson 1985; Barrett 1988; Barrett et al. 1992; Barrett 1993). Phylogenetic reconstructions using both morphological (Eckenwalder and Barrett 1986) and plastid sequence data (Graham and Barrett 1995; Kohn et al. 1996; Graham et al. 1998; Graham et al. 2002) have been employed to investigate character evolution and the systematic relationships of taxa within the family and its close relatives. These analyses indicated that *Eichhornia* is non-monophyletic with
the 8-9 species separated into four distinct clades, two of which contain polyploid species
whose origin remains unclear. One of these four clades includes three species; the selfing *E.
paradoxa*, a second closely related undescribed species and *E. paniculata*. One of the goals
of my thesis is to investigate further the non-monophyletic status of *Eichhornia*.

*Eichhornia paniculata* is a diploid (*n*=8) facultatively annual aquatic, distributed primarily in
N.E. Brazil and on the Caribbean islands of Cuba and Jamaica with a few isolated collections
from Ecuador and the Matto Grosso of Brazil (Figure 1.1). Although thoroughly self-
compatible the species possess a cryptic trimorphic incompatibility system (Cruzan and
Barrett, 1993) which in association with the floral polymorphism tristyly helps promotes high
outcrossing rates in Brazilian populations (Barrett 1985a; Barrett et al. 1989; Barrett and
Husband 1990). Tristyly is a genetic polymorphism in which populations contain three
morphs (long-styled, mid-styled and short-styled, hereafter L- M-, S-morphs) differing
principally in style length, anther height and pollen size (Figure 1.2). The polymorphism
promotes cross-pollination among the floral morphs resulting in disassortative mating
(Barrett et al. 1987). The inheritance of tristyly in *Eichhornia* is controlled by two tightly
linked, diallelic loci (*S* and *M*), with the *S* locus epistatic to the *M* locus (Barrett et al. 1989
and unpublished data). Thus, the L-morph is homozygous for both recessive alleles (*ssmm*),
the M-morph is homozygous recessive at the *S*-locus but carries at least one dominant allele
at the *M* locus (*ssMm, ssMM*), and the S-morph has at least one copy of the dominant *S* allele
(*Ssmm, SsMm, SsMM, SSmm, SSMM*). Because the three morphs are maintained by
negative frequency-dependent selection at a ratio of 1:1:1 (reviewed in Barrett 1993), the
dominant allele at the *S* locus has the lowest frequency in equilibrium populations. As a
result, the S-morph is lost most frequently from populations by stochastic processes,
including founder events and genetic drift (Barrett et al. 1989; Husband and Barrett 1992b).
In dimorphic populations of *E. paniculata* that have lost the S-morph, the M-morph predominates and has higher relative fitness than the L-morph as a result of reproductive assurance and a transmission advantage which favors the spread of selfing variants of the M-morph (reviewed in Barrett et al. 1993; Kohn and Barrett 1994). These fitness advantages arise because of genetic modifications to the M-morph involving elongation in the position of their “short-level” stamens and the variants are described as “semi-homostylyous”; see Figure 1.2). The genetic and developmental changes to stamen position in these variants is governed by recessive mating-system modifiers and results in autonomous self-pollination (Fenster and Barrett 1994; Vallejo-Marín and Barrett 2009). Several lines of evidence suggest that the breakdown of tristyly to semi-homostyly has been followed repeatedly in different parts of the geographical range (Husband and Barrett 1993; Fenster and Barrett 1994). Populations of *E. paniculata* exhibit the complete range of stages in the breakdown process from highly outcrossing to almost completely selfing populations (Figure 1.3). Demographic and genetic studies of the breakdown process (reviewed in Barrett and Husband 1997; Husband and Barrett 1998), in concert with the wide range of mating patterns, affords excellent opportunities to investigate further the interactions of stochastic and deterministic forces on the transition to selfing and its influence on molecular variation.

The biogeographic relations among populations of *E. paniculata* are important for understanding the historical context in which mating-system transitions have occurred. The majority of populations in Brazil are moderately to highly outcrossing, whereas in some Brazilian, and all Jamaican and Cuban populations, there is a high frequency or fixation of the selfing variant of the M-morph. Previous studies on allozyme variation suggest that Jamaican populations are the result of long-distance colonization events from the South American mainland, probably N.E. Brazil given the high concentration of populations in this
region (Husband and Barrett 1991). The origin of the Cuban populations and their relationship with Jamaican populations was not known at the commencement of this thesis as samples from the island were not available for genetic studies. In addition, two collections of *E. paniculata* from isolated localities from Nicaragua and Mexico were also available for study in this thesis. Samples from both of these populations are a semi-homostylosous form of the L-morph not observed elsewhere in the range. When grown under pollinator-free glasshouse conditions in Toronto, plants from these populations both produce 100% fruit set as a result of autonomous self-pollination. Although mating patterns in these populations have not been measured they are likely to be highly autogamous, given their strong facility for autonomous self-pollination. The genetic relationships between the Nicaraguan and Mexican populations and those from the Caribbean (and Brazil) are not known and another goal of this study was to clarify their relationships in the hope that this information may shed light on their respective origins.

**Research objectives**

The main objective of my thesis is to provide novel insights into the interactions of demography and mating-system variation on the molecular population genetics of *E. paniculata*. The past 30 years of research on *E. paniculata* represents perhaps the most comprehensive microevolutionary investigations of transitions in mating systems in the flowering plants. My goal is to contribute to this growing body of research by extending the approaches that have been used to include the techniques of molecular population genetics and genomics in an effort to better understand the forces shaping the evolution of *Eichhornia* genomes. To achieve this goal I have generated and analyzed DNA sequences from a large sample of genes, individuals and populations. Below, I briefly outline the chapters in my
thesis, each of which contains original research. These have been written as self-contained research papers for submission to journals. Consequently, there is inevitable repetition in the introductions and discussions of some chapters. The citations for chapters that have been published are provided in the summaries below.

**Chapter Two** – *Genomic consequences of outcrossing and selfing in plants.* This chapter is a comprehensive review of our current knowledge and understanding of how transitions from outcrossing to selfing can influence the evolution of the genome and how selfers may avoid declines in fitness leading to extinction. We briefly outline the genomic consequences of self-fertilization on genome evolution including impacts on the effective population size, genetic diversity, linkage disequilibrium and efficacy of selection. We also review theoretical and empirical work on the evolution of recombination, mutation rates, genetic conflicts and genome size. This chapter includes a new analysis of the relative genome size of pairs of closely related outcrossers and selfers and reveals that selfers consistently have smaller genomes than their outcrossing relatives. We propose in this article that large-scale comparative sampling of many taxa and genes is necessary for generalizations about genome evolution in selfers. This chapter was published in a special issue of the *International Journal of Plants Sciences,* on “Major Evolutionary Transitions in Flowering Plant Reproduction” edited by S. C. H. Barrett. (Wright, S. I., R. W. Ness, J. P. Foxe and S. C. H. Barrett. 2008. Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences* 169: 105-118).

**Chapter Three** – *Evolutionary pathways to self-fertilization in a tristyloous plant species.* This chapter is a review of our current understanding of evolutionary pathways to self-fertilization and also presents new molecular and genetic evidence in support of multiple independent origins of selfing in *E. paniculata.* We describe for the first time sampling of
morph ratios in populations from Cuba demonstrating that they closely resemble previous sampling from Jamaica. We also describe the characteristics of self-pollinating variants of the L-morph from Mexico and Nicaragua. Evidence from DNA sequences from 10 nuclear loci in 27 population samples provide support for the hypothesis of multiple independent origins of selfing in *E. paniculata*. We also present analysis from controlled crosses of the L- and M-morphs demonstrating recessive control of the loss of herkogamy (stigma-anther separation). We propose that the demographic conditions in *E. paniculata* combined with the relatively simple genetic control of floral traits influencing self-pollination favour repeated transitions to selfing in *E. paniculata*. This chapter was published in *New Phytologist* (Barrett, S. C. H., R. W. Ness and M. Vallejo-Marín. 2009. Evolutionary pathways to self-fertilization in a tristylos plant species. *New Phytologist* 183:546-556).

**Chapter Four** – *Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylos plant Eichhornia paniculata*. In this chapter we investigate the roles of mating-system variation and demographic history in shaping the distribution of nucleotide variation within and among populations of *E. paniculata*. We analyze 10 nuclear loci in 225 individuals from 25 populations collected from much of the geographic range and use coalescent simulations to investigate demographic and colonization history. We found that highly selfing populations, which were generally small in size, exhibited significantly lower genetic diversity compared with larger trimorphic and dimorphic populations. Bayesian structure analysis revealed strong regional clustering and selfing populations were highly differentiated from one another. Coalescent simulations indicated a bottleneck associated with colonization of the Caribbean from Brazil approximately 125,000 years ago. Our analyses suggest that the recent multiple origins of selfing from diverse outcrossing populations of *E. paniculata* may inflate estimates of genetic diversity compared to

**Chapter Five** – *De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants.* In this chapter we use short-read sequencing to assemble, *de novo*, the floral transcriptomes of an outcrossing and two independently derived selfing genotypes of *E. paniculata*, and a selfing genotype of the sister species - *E. paradoxa*. We sequenced mRNA from various stages of flower development to assemble the floral transcriptome and to quantify gene expression. Using a custom pipeline we assembled 24 Mbp of sequence from ~27,000 contigs that were an average of ~900bp in length. Expression was highly correlated in all three *E. paniculata* genotypes and these were more correlated with one another than with *E. paradoxa*. We also identified 269 genes associated with floral development, 22 of which were differentially expressed in selfing lineages of *E. paniculata* relative to the outcrosser. Many of these genes affected floral traits commonly associated with changes to morphology associated with the selfing syndrome. To our knowledge, our study is the first to demonstrate the use of Illumina short read sequencing for *de novo* transcriptome assembly in a non-model plant species. This chapter involved collaboration with Mathieu Siol and Spencer Barrett and is currently under review.

**Chapter Six** – *Genomic consequences of transitions from cross- to self-fertilization on the efficacy of selection in three independently derived selfing plants.* In this chapter, we exam how transitions from cross- to self-fertilization affect patterns of selection across the floral transcriptomes of four *Eichhornia* genotypes. Because of the predicted reduction in $N_e$, genetic drift is expected to dominate over selection against slightly deleterious mutations, or
in favour of slightly advantageous mutations, rendering them both effectively neutral. We analyze patterns of nucleotide variation for evidence of relaxed selective constraint on protein evolution and codon usage bias. From the floral transcriptomes of four genotypes assembled in chapter five, we selected nearly 8000 putatively orthologous sequences totalling ~3.5 Mb of coding DNA. We detected a small but significantly higher proportion of non-synonymous to synonymous changes in the selfing genotypes of *E. paniculata* compared to the outcrosser. Our findings are consistent with a relaxation of selection in selfers, especially with respect to weakly selected synonymous changes that affect codon usage bias, and to a lesser extent in amino acid substitutions. This chapter involved collaboration with Mathieu Siol and Spencer Barrett and is not yet published.

**Chapter Seven** – *Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae.* In this chapter we revisit the phylogeny of Pontederiaceae. Like most plant phylogenetic inference, previous historical reconstructions for this family have used gene regions from the plastid genome. We present new evidence based on nuclear sequences from five gene families using gene tree parsimony (GTP). GTP infers the rooted species trees most compatible with multiple gene families by fitting gene genealogies to species trees and minimizing the number of duplications needed to reconcile conflicts among them. Our results were highly congruent with trees inferred from plastid data alone. We also provide new evidence placing the root of the family at a branch leading to the rare and locally restricted *Eichhornia meyeri*. In this chapter, we also develop methods to incorporate uncertainty in individual gene trees during reconciliation. Our study contributes to an understanding of the phylogenetic history of Pontederiaceae and also demonstrates the utility of GTP for phylogenetic analysis. This
chapter involved collaboration with Sean W. Graham and Spencer Barrett and is currently under review.
Figure 1.1 *Eichhornia paniculata* infesting a rice field near Camelote, Camagüey, Cuba. Populations in Cuba are predominantly selfing. (B) Inflorescence of *Eichhornia paniculata* M-morph from N.E. Brazil (C) Outcrossing (top) and selfing flowers (bottom) of tristylos *Eichhornia paniculata* from Brazil and Jamaica, respectively. Images by S.C.H. Barrett.
Figure 1.2 Evolutionary pathway from cross-fertilization to self-fertilization in tristyous *Eichhornia paniculata* via the breakdown of tristyly to semi-homostyly. The pathway from trimorphism to dimorphism culminates in the fixation of a selfing semi-homostylous form of the M-morph. Arrows linking anthers and stigmas indicate mating combinations; those linking floral phenotypes indicate evolutionary transitions. The trend to smaller reproductive organs with increased selfing reflects reductions in flower size. An additional pathway to selfing is presented in Chapter 3.
Figure 1.3. Variation in multi-locus outcrossing rate ($t$) among 54 populations of *Eichhornia paniculata* in N.E. Brazil (closed symbols) and in Jamaica (open symbols) in relation to their size ($N$) and style morph diversity ($I$). Trimorphic, dimorphic and monomorphic populations are represented by triangles, squares and circles, respectively. Values of $I$ are based on a normalized index of diversity and range from 1.0, equal frequencies of the floral morphs, to 0, floral monomorphism. After Barrett et al. 1992.
CHAPTER TWO
GENOMIC CONSEQUENCES OF OUTCROSSING AND SELFING IN PLANTS

This chapter resulted from a collaboration with Stephen I. Wright, John Paul Foxe and Spencer C. H. Barrett. All authors contributed to the writing and ideas presented in this chapter. Rob W. Ness conducted analyses on genome size. The manuscript was published in The International Journal of Plant Sciences, 2008, 169:105–118.

Summary

Evolutionary transitions from outcrossing to selfing are expected to cause a reduction in the effective population size, and a corresponding increase in fixation rates of slightly deleterious mutations and decrease in fixation of advantageous mutations. Despite these predictions, evidence from genomic data does not suggest a significant reduction in the efficacy of selection associated with high levels of self-fertilization. Here, we discuss opportunities for selfing populations to avoid an irreversible decline in fitness towards extinction and the implications for genome evolution. Most directly, large population sizes and the purging of deleterious recessive mutations can reduce genetic loads and slow the effects of genetic drift. Theory suggests that recombination rates may also evolve in response to the evolution of mating system, which can offset the harmful effects of inbreeding. Cytological data supporting the evolution of higher recombination rate in selfing species should be supplemented with genetic and molecular methods for estimating this parameter. Mutation rates may also evolve to be higher in selfing plants, due to hitchhiking with advantageous mutations, although this is unlikely to lead to increased fitness. Finally, the abundance and activity of selfish genetic elements may also be reduced in selfing lineages, reducing the accumulation of transposable elements, B chromosomes, biased gene conversion and the spread of cytoplasmic male sterility mutations. This reduction in genomic
conflict can increase mean fitness, reduce deleterious mutation rates, and reduce genome size. We show, using comparative data, that highly selfing plants have significantly smaller genomes in comparison with outcrossing relatives, consistent with reduced activity and spread of repetitive elements in inbred plants. We discuss opportunities for tests of theory as plant genomic data accumulate and argue that a genomic perspective on reproductive transitions in a phylogenetic context should provide important insights into the diversity of reproductive systems in flowering plants.

Introduction

Flowering plants exhibit spectacular diversity in reproductive systems and this can have important effects on the amount and the structuring of genetic variation within and among populations (Hamrick and Godt 1996; Glemin et al. 2006). Reproductive transitions, such as the shift in mating system from outcrossing to selfing, tend to increase the strength and extent of linkage disequilibrium. When considering the probability of fixation of mutations subject to natural selection the strength, efficacy, and the sign of selection acting on mutations can be influenced by the extent of linkage disequilibrium with other sites in the genome. Therefore, evolutionary transitions in reproductive systems should play a central role in genome evolution. However, little is known about the genomic consequences of plant reproductive diversity and how transitions in sexual systems and patterns of mating may influence genome evolution.

At last count there are 41 large-scale plant genome-sequencing projects underway (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi), with more to follow in the next few years. Although many of these projects focus on crop plants, increasing attention is being focused on ecological and evolutionary model systems, including *Mimulus guttatus, Aquilegia*
formosa, Thellungiella halophila, Capsella rubella and Arabidopsis lyrata. (http://www.jgi.doe.gov/sequencing/allinoneseqplans.php). As our ability to compare patterns of genome structure and evolution accelerates, a theoretical and empirical framework for understanding plant genome evolution in wild plant populations becomes increasingly important. Moreover, since many of these non-domesticated species display extensive variation in reproductive traits and patterns of mating there are likely to be rich opportunities for investigating the genomic consequences of reproductive diversity in flowering plants.

Although extensive research has focused on genome evolution in polyploids (Chen 2007) relatively little attention has focused on the potential for mating-system transitions to restructure genomes, perhaps because the effects of polyploidy are immediate and more conducive to experimental manipulation (Husband et al. 2008). In contrast, mating-system transitions are expected to lead to shifts in the selective dynamics of genomic elements over evolutionary time scales. Nevertheless, such changes could be equally important for understanding the evolutionary dynamics of plant genomes.

Mating-system variation has several important effects on the genetic properties of populations (Figure 2.1), which we treat here only briefly since they have been covered in several recent reviews (Charlesworth and Wright, 2001; Glemín et al. 2006). Most directly, homozygosity increases as a function of selfing rate and this reduces the effective size of a population due to a reduction in the number of distinct alleles. The effective size of a completely selfing population is reduced two-fold as a result of homozygosity (Charlesworth et al. 1993b; Nordborg 2000). Further, because of homozygosity, crossing-over rarely occurs between heterozygous sites, increasing linkage disequilibrium among loci (Figure 2.1A and Figure 2.2; and see Nordborg 2000). This results in stronger effects of genetic hitchhiking, in
the form of selective sweeps of positively selected mutations (Figure 2.1B) and background selection acting against deleterious mutations (reviewed in Charlesworth and Wright 2001), which further reduce $N_e$ in the affected regions. Linkage among weakly selected sites with opposing selective forces can also interfere with the ability of selection to act efficiently (McVean and Charlesworth 2000). All of these forces reduce $N_e$ and may be further exaggerated by life-history characteristics associated with selfing that promote population subdivision, isolation, and genetic bottlenecks. Finally, highly selfing species may experience reduced levels of between-species introgression (Sweigart and Willis 2003), leading to further reductions in genetic diversity. Together, these processes should lead to a decrease in the efficacy of natural selection and an increase in the fixation rate of slightly deleterious mutations (Figure 2.1C), with important consequences for evolution at the genome level (Lynch and Conery 2003). Over evolutionary time, increased deleterious mutation accumulation can be important in causing species extinction (Lynch et al. 1995), and this may, in part, explain a lack of persistence of selfing lineages as revealed from comparative and phylogenetic studies (reviewed in Takebayashi and Morrell 2001; Kohn et al. 2008).

There is clear and consistent evidence for a reduction in levels of within-population neutral diversity in highly selfing species (Schoen and Brown 1991; Glemín et al. 2006;). In contrast, as we discuss below, studies of molecular evolution to date have found only limited evidence for elevated fixation of deleterious mutations in selfing species (Wright et al. 2002; Glemín et al. 2006). One possible explanation for this is the inadequacy of sampling to date. There is simply the need to collect data from very large numbers of consistently sampled loci, given the inherent stochasticity of deleterious mutation accumulation. This will soon become feasible with the rapid expansion of whole genome data discussed above. Alternatively, the time scale for the evolution of selfing may be too recent in many lineages
for substantial genomic changes to occur. However, it is also possible that many selfing lineages, particularly those successful model systems that have been the focus of research, can avoid long-term fitness decline through compensatory mechanisms.

In this review, we examine the potential for avoidance of fitness decline associated with the evolution of self-fertilization. First, elevated homozygosity, particularly in large populations, can lead to the purging of recessive, strongly deleterious mutations and enhance the fixation of recessive advantageous mutations. Second, recombination rates are predicted to evolve following the evolution of selfing, which can directly reduce the harmful effects of suppressed recombination. Third, mutation rates may evolve to be higher; although this could increase rates of adaptation, it may also contribute to mutational meltdown and thus is unlikely to lead to increased fitness in selfing lineages. Finally, selfish genetic elements, which represent a major class of genomic deleterious mutation, may be more effectively selected against in highly selfing populations. Theory predicts that the spread and effects of transposable elements, B chromosomes, biased gene conversion, and cytoplasmic male sterility mutations should all be reduced in highly selfing species. Since these elements can spread in outcrossing species despite fitness costs, their elimination in selfing lineages can increase mean fitness and reduce genome size and deleterious mutation rates. We now consider these issues in turn.

**Inbreeding and the efficacy of natural selection**

Reductions in effective population size in inbreeders due to the factors discussed above are expected to elevate fixation rates of deleterious mutations and decrease fixation rates of advantageous mutations. However, the degree to which polymorphism and/or substitution rates are affected by selfing will strongly depend on the distribution of
mutational effects on fitness, which remains very poorly understood (Wang et al. 1999; Weinreich and Rand 2000; Eyre-Walker and Keightley 2007). In particular, high homozygosity due to inbreeding can lead to greater expression of deleterious recessive mutations, leading to their elimination, a phenomenon referred to as purging (Crnokrak and Barrett 2002; see Schoen and Busch 2008). Because of this, selfing leads to a rapid elimination of highly recessive, strongly deleterious mutations from the population. Models show that deleterious mutation accumulation in inbreeding taxa is dominated by slightly deleterious mutations that are closer to additive in their fitness effects (Wang et al. 1999). If a large fraction of deleterious mutations are recessive and strongly selected, the effects of purging may thus dominate over slightly deleterious mutation accumulation, at least in the context of studies of molecular evolution. Furthermore, recessive advantageous mutations can be fixed more effectively in highly selfing populations (Charlesworth 1992), increasing rates of adaptive evolution. In general, the predictions for molecular evolution rely on the assumption of the presence of a large class of mutations that are weakly selected and that are nearly additive in their fitness effects, mitigating any effect of homozygosity in purging deleterious recessive mutations and enhancing the fixation of recessive advantageous mutations.

In addition to the parameters associated with deleterious and advantageous mutations, population sizes can also be important in predicting the rate of fitness decline. Since the extent of linkage disequilibrium is determined by $4N_e r(1-s)$, where $r$ is the rate of recombination and $s$ is the selfing rate (Nordborg 2000), high selfing rates can be partially compensated for by large population sizes, and this will reduce any expected effects on diversity and molecular evolution. Even in highly selfing populations with large population
sizes, linkage disequilibrium may not extend over large genomic regions, and the efficacy of natural selection may not be low.

Reduced effectiveness of selection on amino acid altering mutations can be inferred by increases in the proportion of non-synonymous relative to synonymous substitutions (Wright et al. 2002) or polymorphisms (Bustamante et al. 2002). In addition, a reduced efficacy of selection on biased codon usage can be detected via elevation of un-preferred over preferred synonymous substitutions and polymorphisms (Marais et al. 2004). A comparison of substitution rates and codon bias between the highly selfing *A. thaliana* and self-incompatible *A. lyrata* failed to find evidence for an effect of mating system for both amino acid substitutions and patterns of codon bias (Wright et al. 2002). This suggests that there is not a major decline in effectiveness of selection, at least in these taxa. Similarly, in a comparative analysis of nuclear and chloroplast DNA sequence data from multiple species, Glemín et al. (2006) found no effect of inbreeding on codon usage, with the exception of comparisons in the Poaceae. In this family, differences in base composition were observed for both synonymous sites and noncoding sites, suggesting that this result is unlikely to be caused by differences in selection on codon usage bias (see below).

Although the above comparisons indicate little effect of inbreeding on the effectiveness of selection, these tests focused on effects on substitution rates between species. Such tests would not detect elevation of slightly deleterious mutations if selfing evolved recently, nor could they reveal any reduction in positive selection on amino acids in selfers, since they assume that non-synonymous substitutions are predominantly slightly deleterious in their effects. A high proportion of rare non-synonymous polymorphisms have been reported in the inbreeding wild barley, *Hordeum vulgare* ssp. *spontaneum* (Cummings and Clegg 1998) and *A. thaliana* (Bustamante et al. 2002; Nordborg et al. 2005), although
explicit comparisons of polymorphism and divergence with related outcrossing species have not as yet been undertaken. Using a Bayesian analysis of the ratios of polymorphism to divergence for amino acid replacement and synonymous changes, Bustamante et al. (2002) found evidence for excess amino acid polymorphism and a reduction in positive selection in selfers. However, this study compared *A. thaliana* with *Drosophila melanogaster* and clearly more detailed analyses of polymorphism and divergence for close relatives with contrasting mating systems is needed.

In their analysis of sequence variation in plant species with contrasting mating systems, Glemín et al. (2006) detected a weak but significant elevation in the ratio of amino acid to synonymous polymorphism in species with high selfing rates. This suggests that there may be an elevated frequency of mildly deleterious polymorphisms segregating in selfing species. However, much of the data they reviewed is restricted to a small number of taxa and/or loci and direct comparisons of polymorphism and divergence for the same loci are sparse. Large-scale sampling of polymorphism and divergence among close relatives will be important to further assess whether inbreeding plays a significant role in coding-sequence evolution. Currently, comparisons of patterns of amino acid polymorphism and divergence in *Arabidopsis* do not suggest a significant difference between selfing and outcrossing species (Foxe et al. 2008).

Limited evidence suggests there may be an elevation of non-synonymous polymorphism with increased selfing. However, there is little to indicate that inbreeding leads to a sufficient decline in the effectiveness of selection to influence substitution rates and genome evolution. This suggests that there may be relatively few sites experiencing sufficiently weak, additive selection to be affected by changes in selfing rate. As a result, the
power to detect subtle changes in the patterns of molecular evolution may be too low, given the sample of genes studied to date.

In addition, as discussed above, large population sizes may reduce linkage disequilibrium in the selfing taxa studied. *Arabidopsis thaliana* is a predominantly selfing annual with a near worldwide distribution. This successful colonizing species exhibits high nucleotide diversity and retains a fairly rapid decay of linkage disequilibrium with physical distance in range-wide population samples, despite very high selfing rates (Nordborg et al. 2005). Although within-population diversity is much lower in *A. thaliana* than in populations of self-incompatible *A. lyrata* (Savolainen et al. 2000; Wright et al. 2003b; Ramos-Onsins et al. 2004), this does not necessarily imply a general reduction in the efficacy of selection. If selection against slightly deleterious mutations is effective in a metapopulation-type structure, the effective size across populations in *A. thaliana* could counteract reduced effective recombination rates. Furthermore, outcrossing populations can also experience population bottlenecks, reducing variation and increasing linkage disequilibrium to levels comparable to inbreeders (Wright et al. 2003b). A very broad taxonomic comparison is illustrative; humans have an order of magnitude lower level of variability and much more extensive linkage disequilibrium than worldwide samples of *A. thaliana* (Nordborg et al. 2005), and they show a comparable, or more severe, genome-wide excess of slightly deleterious amino acid polymorphism (Bustamante et al. 2005). Similarly, low linkage disequilibrium has recently been demonstrated in highly selfing wild barley populations (Morrell et al. 2005). In contrast, recent comparisons of outcrossing vs. selfing species of *Caenorhabditis* identified very extensive levels of linkage disequilibrium in the selfing *C. elegans*, in contrast with outcrossing congeners (Cutter et al. 2006b). Thus, variation in the
population history of a species may contribute as much or more to variation in fixation rates at weakly selected sites, swamping out any effect of the mating system alone.

The role of purging and large population sizes in eliminating deleterious mutations in selfing populations may explain the contrast between the results to date for selfing species and those found for non-recombining Y-chromosomes in plants and animals, where recurrent evidence for deleterious mutation accumulation has been found (Charlesworth and Charlesworth 2000). In inbreeders, deleterious mutations are fully expressed as homozygotes, and so fixation is directly dependent on the selection coefficient. In contrast, because of the persistence of functional gene copies on the X-chromosome, Y-chromosome fixation probabilities will be determined by $hs$, the product of the selection coefficient ($s$) and the dominance coefficient ($h$). In the extreme case, mutations that are completely recessive in their fitness effects can fix neutrally on Y-chromosomes, despite strong selection coefficients in homozygotes. Thus, if deleterious mutations are strongly recessive, the permanent heterozygosity of the Y-chromosome brings a much larger proportion of deleterious mutations into a parameter space that allows their fixation. Perhaps even more importantly, the non-recombining portion of Y-chromosomes exhibits near-complete recombination suppression, whereas highly selfing species can more easily ‘escape’ the linkage effects, for example by having larger population sizes. The complete suppression of recombination in Y-chromosomes leads to much stronger hitchhiking than even very high selfing in large populations.

Evolution of recombination rate

The increase in homozygosity as a result of selfing reduces opportunities for recombination to break down associations among alleles by crossing over between
heterozygous loci (Nordborg 2000). This increases linkage disequilibrium and decreases the effective rate of recombination. However few species that engage in high levels of selfing are exclusively selfing (Barrett and Eckert 1990; Igic and Kohn 2006) and therefore some opportunity exists for recombination during rare outcrossing events. It is therefore possible that natural selection may favour modifiers that increase rates of physical recombination or crossing over in selfers to offset the effects of inbreeding. Although theory predicts that mating systems may play an important role in the evolution of recombination rates this area has received relatively little empirical attention.

Recombination is generally thought to be advantageous because it breaks down associations between alleles (linkage disequilibrium) and is favoured under several non-mutually exclusive conditions. For example, with weak negative epistatic associations among mutations (Feldman et al. 1980; Kondrashov 1982, 1988; Barton 1995), when directional selection negatively covaries between habitats (Lenormand and Otto 2000), and when Hill-Robertson interference is strong, such as in populations with small effective population size (Fisher 1930; Muller 1932; Otto and Barton 1997; 2001; Barton and Otto 2005; Keightley and Otto 2006). Under these conditions, recombination increases the variance in fitness, and can introduce higher fitness alleles on the same genetic background.

Models exploring the fate of recombination modifiers have mostly assumed random mating and little attention has been given to how the mating system may alter the outcome of these models. Simulations indicate that hitchhiking between a recombination modifier and a pair of selectively important loci is stronger with selfing and that this difference can under some conditions favour the evolution of higher recombination rates. In contrast, the recombination rate is generally driven downwards with random mating (Charlesworth et al. 1977; Charlesworth et al. 1979; Holsinger and Feldman 1983a). Supporting these results,
Roze and Lenormand (2005) generated an analytical model and found that even small amounts of selfing can greatly increase the range of parameters under which selection favours increased recombination.

In small populations, genetic drift is primarily responsible for generating negative linkage disequilibrium that favours recombination modifiers increasing rates of crossing over (Otto and Barton 2001). The strength of indirect selection is stronger when linkage among loci is tight and population size is small; with tight linkage, a broader parameter space of population size favours recombination modifiers (Barton and Otto 2005). Although not modelled explicitly, these results suggest that with self-fertilization there may be stronger selection on recombination modifiers. Although details of the models vary, there is a general agreement that a fairly broad set of conditions favour an increase in the recombination rate with increased selfing. Support for these predictions is provided by comparisons of chiasma frequency in plant species with contrasting mating systems (Ross-Ibarra 2004; Roze and Lenormand 2005).

An alternative method for estimating the rate of crossing over per base pair is to integrate genetic linkage maps with physical maps. This allows for estimates of the average frequency of crossing over per physical length and a more detailed view of rate heterogeneity in different genomic regions, although the accuracy of cytological versus genetic estimates of recombination rate has been subject to debate (Nilsson et al. 1993). Such integrated maps are available for several model systems and agricultural taxa. However, the only detailed comparison of map-based recombination rates in closely related species with contrasting mating systems is between A. thaliana and A. lyrata (Kuittinen et al. 2004; Hansson et al. 2006; Kawabe et al. 2006). As predicted, overall rates of recombination per unit physical length are higher in A. thaliana. The degree of difference in recombination rate between the
two *Arabidopsis* species varies across different genomic regions.

In the *Arabidopsis* comparison, the contrast in overall rate of recombination is complicated by a shift in genome size. *Arabidopsis thaliana* has a reduced genome size and there is a general trend that rates of recombination per base pair decrease with increasing genome size (Ross-Ibarra 2006). Therefore, it is difficult to rule out an effect of genome size difference, rather than direct selection on recombination modifiers. *Arabidopsis lyrata* has a larger genome, and potentially more heterochromatic, non-recombining DNA than *A. thaliana*. Therefore, the difference in average recombination rates across large regions could result from other aspects of genome evolution in addition to selection on recombination modifiers such as differences in the accumulation of transposable elements (see below).

Nevertheless, using phylogenetically independent contrasts of 142 species with genome size as a covariate, Ross-Ibarra (2006) detected elevated chiasma frequencies in selfers suggesting higher rates of recombination than in related outcrossers. In addition to overall recombination frequencies, outcrossing rates have also been shown to correlate with the ratio of recombination rates in female relative to male function. Lenormand and Dutheil (2005) reported that selfing species tend to have a higher ratio of male to female recombination rates in comparison with outcrossing species. Their interpretation is that because pollen competition is reduced in selfing species this relaxes selection against recombination breaking up favourable epistatic allelic combinations in pollen. Given that most estimates of chiasma frequencies have been determined in pollen, it will be important to test whether sex-averaged recombination rates are generally elevated in selfing species.
Relations between mating system and genomic mutation rate

Beneficial mutations are ultimately necessary for novel adaptations to evolve. However, the relative input of beneficial and deleterious mutations determines the evolution of genome-wide mutation rate, which may be strongly influenced by mating system. In a random mating population a modifier that increases the mutation rate should be selected against because it introduces more deleterious than beneficial mutations. Therefore, the fact that estimates of mutation rate are all above zero presumably reflects physical constraints on further reductions, or a trade-off between fitness effects of new deleterious mutations and the amount of energy required for higher fidelity DNA replication and repair (see Sniegowski et al. 2000). However, in selfing and asexual populations strong linkage between a modifier that increases the mutation rate and a resulting beneficial mutation may result in indirect selection for the modifier. It is therefore possible to achieve an equilibrium rate of mutation that is above zero without invoking any physical constraints. The equilibrium value depends both on the extent of linkage disequilibrium and the relative frequency and effect of deleterious and beneficial mutations. However, while many models have investigated the effects of indirect selection in asexual populations (Kimura 1967; Leigh 1970; Eshel 1973; Painter 1975; Woodcock and Higgs 1996; Orr 2000), only one has explicitly modeled the effect of self-fertilization (Holsinger and Feldman 1983b). Furthermore, there is a paucity of empirical tests of these predictions. Below, we summarize advances made in understanding the effect of mating system, specifically selfing, on the evolution of genomic mutation rate.

The effective rate of recombination in highly selfing populations is low. Therefore, a modifier that increases the mutation rate will have a higher probability of linkage to a beneficial mutation and will be indirectly selected. In a highly selfing population a modifier locus that alters the mutation rate at a second locus is predicted to reach a nonzero
equilibrium mutation rate when overdominance favours the novel heterozygote genotype (Holsinger and Feldman 1983b). This result is robust for a non-trivial amount of outcrossing (<10%), making it more relevant to natural systems which, as discussed earlier, are rarely if ever completely selfing. Johnson (1999) considered a mutation modifier that generates mutations of varying effect across a genetic map, instead of one that acts on a single selectively important locus. With smaller map sizes beneficial mutations will have a stronger influence on the fate of a modifier than previously predicted. Interestingly, the strength of linkage disequilibrium, which is of primary importance, is much higher for selfers than for random mating populations. This suggests that if a high proportion of mutations are beneficial, they may be more important in predominantly or partially selfing populations. However, if beneficial mutations are very rare and weakly selected then deleterious mutations will cause stronger negative selection against increased mutation rate in selfing (or asexual) populations than in outcrossing populations (Kondrashov 1995; McVean and Hurst 1997; Dawson 1998). To distinguish between these possibilities the distribution of mutational effects is required, yet empirical estimates of the relative proportion of beneficial and deleterious mutations are largely unavailable.

Rates of deleterious mutations between predominately outcrossing *Amsinckia douglasiana* and selfing *A. gloriosa* have been compared in a mutation accumulation (MA) experiment (Schoen 2005). Although both lineages showed declines in fitness because of an accumulation of mutations, there was no significant difference associated with the mating system. Only three other MA experiments have been conducted in plants, two with *A. thaliana* (Schultz et al. 1999; Shaw et al. 2000) and another with *Triticum durum* (hard wheat, Bataillon et al. 2000). A problem inherent to MA studies is the difficulty of conducting experiments with organisms with long generation times. Because most mutations
are of small effect many generations are required for the power necessary to detect differences among groups. MA studies are primarily concerned with the rate of deleterious mutations and therefore underestimate the total mutation rate across the genome. Further, the exact measure of $U$ (deleterious mutations per genome per generation) is dependent on the underlying distribution of mutational effects and estimates of $U$ can vary by orders of magnitude depending on the distribution that is assumed (Shaw et al. 2000; 2002).

An alternative method of estimating mutation rates is by comparing nucleotide sequences of divergent lineages. The level of neutral divergence between two lineages is equal to $2\mu t$ where $\mu$ is the mutation rate (per nucleotide per generation) and $t$ is the number of generations since divergence (Kimura 1968). However, this assumes a constant mutation rate and is therefore not useful for testing whether there are rate differences between the lineages. The relative rates test (Sarich and Wilson 1967; Wu and Li 1985; Tajima 1993) can be used to test for heterogeneity of rates among lineages. While there are a number of studies that apply this technique to non-coding and silent sites of nuclear loci in plants (e.g. Gaut et al. 1996, 1999; Wright et al. 2002; Senchina et al. 2003) there are none that do so for enough loci to estimate a mean mutation rate across the genome. In fact, substantial heterogeneity of mutation rates among loci in the above studies demonstrates the necessity of using many loci to estimate a genomic mutation parameter. A parsimony-based study of substitution rates at 23 genes failed to detect a significant difference in neutral mutation rate between *A. thaliana* and *A. lyrata* (Wright et al. 2002). A follow-up comparison did suggest a slight but significant elevation of synonymous substitutions in *A. thaliana* using a sample of 83 genes (Wright 2003). This contrast is complicated by differences in generation time, since *A. thaliana* is annual and *A. lyrata* is biennial or perennial. However, Whittle and Johnston (2003) found no association of generation time and mutation rate in a comparison of 24
phylogenetically independent pairs of annual and perennial plants. Complete uncoupling of generation time from mating system will require broader-scale comparisons in a phylogenetic context.

Could higher mutation rates enhance the ability of selfers to adapt, counteracting the harmful effects of inbreeding? Both theory and data suggest that this is unlikely. In particular, because most mutations are deleterious, elevated mutation rates are likely to be transient outcomes of hitchhiking with beneficial mutations, rather than long-term shifts which increase adaptation (Sniegowski et al. 2000). Overall, higher mutation rates should enhance the mutational meltdown predicted in inbreeding populations, rather than reduce its deleterious consequences.

**Mating systems and genetic conflicts**

‘Selfish’ genetic elements that enhance their own transmission, despite null or negative fitness consequences for the genome, represent a dominant component driving evolutionary divergence in eukaryotic genomes (Burt and Trivers 2006). The persistence and spread of selfish genetic elements can be generally understood as a balance between their transmission advantage and any deleterious effects related to their activity. If the rate of spread of selfish genetic elements predominates over reduced fitness, such elements can persist and even become fixed in natural populations. Their activity can also represent a significant contribution to spontaneous mutation rates and genetic load (Lai et al. 1994, Houle and Nuzhdin 2004).

As we discuss below, transitions in mating system including increased rates of selfing are generally expected to have an important effect on the outcome of genomic conflicts. In general, highly inbred or asexual taxa are expected to experience reduced intragenomic
conflict, since the lack of outcrossing prevents the spread of selfish genetic elements into other genetic backgrounds thereby increasing the variance in fitness and leading to stronger purifying selection (Cavalier-Smith 1980; Hickey 1982). Furthermore, because selfish genetic elements will remain in greater linkage disequilibrium with replicate elements causing harmful mutations, there may be selection on the elements themselves for reduced activity, as models have shown for the evolution of transposition rates (Charlesworth and Langley 1996). Finally, because many selfish genetic elements increase transmission rates by gaining over-representation in gametes when competing with alternative alleles, the presence of high homozygosity in selfing species acts as a further deterrent to alleles that make use of meiotic drive to increase in frequency. With homozygosity, the competition among alleles which leads to drive is absent, since individuals have either two copies of the ‘driving allele’ or the wild type allele, but not both. Therefore, selfing ensures a second level of homogeneity not experienced by asexuals, and should act further to inhibit the spread of selfish elements. Given the potential importance of selfish genetic element activity for mean population fitness and deleterious mutation rates, selective elimination of selfish elements by inbreeders could offset the predicted decline in fitness associated with selfing.

Transposable elements

Equilibrium models of transposable element evolution indicate that stable copy numbers can be maintained in populations by a balance between transposition increasing copy number and the action of negative selection removing insertions from the population (Charlesworth and Langley 1989). In inbreeders, computer simulations (Wright and Schoen 1999) and analytical models (Morgan 2001) show that the spread and accumulation of TEs can be inhibited by the lack of outcrossing, due to reduced spread of elements between
individuals, as well as the purging of insertions with deleterious recessive effects on fitness.

As mentioned above, self-regulated transposition is also more likely to evolve in inbreeders (Charlesworth and Langley 1996), bringing down the copy number compared with outcrossing relatives.

By contrast, if natural selection acts predominantly against TE insertions through ectopic (i.e. between-element) recombination events causing chromosomal rearrangements in heterozygotes, there may be a strong relaxation of natural selection against TEs in selfers (Charlesworth and Charlesworth 1995b). This could lead to rapid accumulation and population fixation (Wright and Schoen 1999; Morgan 2001). The net outcome will depend on the underlying nature of selection on TEs and the history of selfing. For example, recent transitions to inbreeding may relax selection against ectopic exchange and reduce effective population size, leading to increased frequencies and fixation of insertions. However, over the long-term the inability of new insertions with deleterious effects to spread through selfing populations may limit a rampant accumulation process. Simulations show that, even under the ectopic exchange model, stochastic loss of elements from selfing populations is more frequent, potentially resulting in a net loss of TEs from selfing genomes (Wright and Schoen 1999). With a recent transition to selfing, this could lead to an effect in one direction on polymorphism patterns (i.e. increased TE frequencies at individual sites inherited from the ancestor), without a large copy number increase, or perhaps a decrease in copy number in selfers. Note that in either case, transposable element activity reduces fitness less in inbreeding than outbreeding populations.

Early results to date suggest a decline in TE abundance in selfing species, consistent with the basic predictions of deleterious insertion models. Morgan (2001) reviewed evidence from a number of species pairs, showing a general reduction in copy number in inbreeding
species. Furthermore, given the strong correlation between TE abundance and genome size, preliminary evidence for reduced genome size in selfers (see below) is at least consistent with inbreeding playing a role in the elimination of selfish genetic elements. In a population genetic study, Ac-like TE insertions were at higher frequencies in populations of the selfing A. thaliana compared with A. lyrata (Wright et al. 2001). This suggests a relaxation of natural selection, but this was not accompanied by an accumulation of new element insertions, consistent with inhibition of new element activity. Similar results were obtained in a study of retrotransposable elements in tomato (Tam et al. 2007), although a lack of evidence for purifying selection for most transposons even in outcrossing taxa suggests that the patterns may primarily reflect neutral insertion polymorphism. Analysis of transposable element distributions in A. thaliana also suggested that recombination rate heterogeneity does not appear to influence transposable element accumulation in this species, as has been observed in outcrossing genomes (Wright et al. 2003a), and this pattern is also apparent in the selfing nematode C. elegans (Duret et al. 2000). Selection against ectopic recombination may be a weak force in most selfing genomes.

As with inbreeding, asexual species are expected to show reduced TE activity and abundance, particularly due to the inability of active elements to spread among individuals (Hickey 1982). In contrast to this prediction, analysis of sequence evolution of three retrotransposon families in four asexual plant species identified many copies and evidence for selective constraint on TEs in these taxa, suggesting that elements remain active (Docking et al. 2006). However, given the likely recent evolution of asexuality in these plant species (Docking et al. 2006), simulations indicated that the evidence for residual selective constraint in TEs was not unexpected, even if long-term reductions in activity is occurring. More direct comparisons of abundance and polymorphism patterns between related sexual and obligate
asexual taxa would be useful to provide a direct test of the effects of transitions to asexuality on TE evolution.

**B-Chromosomes**

B-chromosomes are non-essential chromosomes found in addition to the basic set of chromosomes. They have now been identified in over 2000 species (Burt and Trivers 2006), including over a thousand plant species (Jones 1995), and are often morphologically distinct, usually smaller than essential chromosomes, and show numerical variation within and between individuals. B-chromosomes may be neutral, positive, or more often, harmful in their effects and are maintained in the genome via a meiotic drive mechanism, ensuring a greater representation in gametes than expected by chance (Camacho 2006).

Models have shown a strong effect of outcrossing rate on the equilibrium frequencies of B-chromosomes due to reduced transmission and greater variance in fitness. For example, Burt and Trivers (1998) found that outcrossing rates below 50% lead to complete elimination of B-chromosomes from populations, reducing significant fractions of ‘junk DNA’ and any fitness costs associated with B-chromosome maintenance. A comparative survey of the distribution of B-chromosomes in 353 plant species from the UK (12.5% of which contained B-chromosomes), using phylogenetically independent contrasts, confirmed a strong and consistent effect of mating system on the presence of B-chromosomes (Burt and Trivers 1998). Three independent analyses performed to test for an association between the presence of B-chromosomes and mating system revealed a positive correlation between outcrossing and the presence of B-chromosomes in 16 out of 19 taxonomic contrasts. Three predominantly inbreeding species were found to contain B-chromosomes, namely *Desmazaria rigidum*, *Poa annua* and *Luzula campestris*. In these cases, it is possible that B-
chromosomes are beneficial. Alternatively, the time scale of mating-system evolution may have been too recent for selection to have effectively purged these genetic elements. Theoretical work incorporating the transition to selfing will be important to better understand the time scales required for the selective elimination of B chromosomes following shift in mating system.

Biased gene conversion

Gene conversion is the non-reciprocal copying of one stretch of DNA into another during recombination (Marais 2003). It has been argued that the genetic systems involved in this repair are biased and can lead to transmission distortion in favour of GC bases. For example, if an individual is heterozygous at a site for a G/T polymorphism, biased gene conversion will lead to an over-representation of G- over T gametes, thus leading to biased transmission. The net effect is a selective advantage of GC over AT bases, with the selection coefficient being determined by the rate of biased gene conversion (Marais 2003). Several studies to date, using patterns of genome structure, population genetic data, and evidence from DNA mismatch repair processes, have found evidence for biased gene conversion (BGC) in yeast (Birdsell 2002), Drosophila (Galtier et al. 2006) and in mammals (Duret et al. 2006). However, there is little evidence to date for the presence of biased gene conversion in plant genomes.

Because of the transmission advantage inherent in the process, GC bases under this model are effectively selfish genetic elements, where there is a fixation bias towards GC bases due to biased transmission. Because the process leads to an increased fixation probability of GC bases it can have important consequences for base composition evolution.
GC biased gene conversion could lead to reduced fitness at sites under weak selection favouring A-T nucleotides, such as plant introns (Ko et al. 1998).

In an outcrossing species, the efficacy of biased gene conversion is given by the product of $N_e \gamma L c$ where $\gamma$ is the probability per generation that a given site is affected by a gene conversion tract $L$ and $c$ is the bias in favour of the GC allele (Nagylaki 1983a, 1983b; Galtier et al. 2001). However, because biased gene conversion will only occur in heterozygotes, the effective rate of this process will be reduced dramatically in highly selfing species. Analytical results have shown that the strength of biased gene conversion, and thus GC content, will be reduced as a direct product of the selfing rate; since BGC will only occur in heterozygotes, an organism with an outcrossing rate of 1% will experience only 1% the level of biased gene conversion of an equivalent, highly outcrossing species (Marais et al. 2004). As a result the mating system could play a role in eliminating any effect of this process on base composition.

If biased gene conversion plays a major role in structuring genomic base composition, inbreeding species should generally evolve to become more AT-rich, particularly at sites less constrained by other forms of selection, e.g. synonymous sites. In their analysis of genomic diversity in angiosperms, Glemín et al. (2006) examined the GC content of 10 species with contrasting mating systems. Although no effects were identified in most comparisons, outcrossing species in the Poaceae were found to have significantly higher GC content compared to selfers, as measured by total GC, GC at third codon positions, and GC in introns. Given the consistent elevation of GC across non-coding and coding sites, this pattern is unlikely to be explained by a difference between species in the efficacy of natural selection on codon usage bias. Instead, it is more likely to be the result of contrasting rates of biased gene conversion.
Similarly, an analysis of base composition in the Brassicaceae revealed a consistent reduction in GC content at third codon positions in *A. thaliana* compared with outcrossing *Brassica oleracea* and *A. lyrata* (Wright et al. 2007). Analysis of base composition evolution controlling for gene expression and codon preferences indicated that the contrasting patterns were unlikely to result from differences in the effectiveness of selection on codon usage. Instead, the pattern appeared to be the direct result of a consistent decline in GC-richness in *A. thaliana*, potentially due to a reduction in biased gene conversion following the recent evolution of selfing in this lineage (Nasrallah et al. 2004). Although alternative explanations, including shifts in the patterns of mutation bias, are possible, the difference is consistent with expectations under biased gene conversion. Comparisons of polymorphism and divergence for GC → AT versus AT → GC changes should help untangle the role of biased gene conversion vs. mutation bias. In particular, if bias gene conversion is strong, AT → GC changes are expected to segregate at higher frequencies than GC → AT changes. This should result in higher frequencies and fixation rates of AT → GC changes in outcrossing species compared with selfing species, under the model of biased gene conversion. By contrast, no such effect is expected if the difference is driven by changes in mutation bias.

In addition to an overall reduction in GC content, the reduced opportunity for biased gene conversion could act to reduce heterogeneity in base composition across the genome in selfers. In particular, variation in GC content may be driven, in part, by variation in rates of recombination, assuming a tight correlation between estimated rates of recombination and rates of biased gene conversion. However, in a highly selfing species the overall reduction in biased gene conversion should result in a weaker relation. Consistent with this, recombination rate and GC content are not positively correlated in *A. thaliana*, as they are in
other genomes (Marais et al. 2004). In the future it will be important to test for such a
correlation using the complete genomes of related outcrossing taxa.

*Cytoplasmic male sterility*

Because of the predominant maternal inheritance of mitochondria and chloroplasts,
cytoplasmic mutations reducing male fertility will be selectively favoured in these genomes
if they cause even a slight increase in female fitness, even if total reproductive output is
reduced. However, increased female frequencies will often favour the spread of nuclear
restorers suppressing male sterility. *Cytoplasmic male sterility* (CMS) is a maternally
inherited condition leading to male infertility as a result of an inability to produce viable
pollen. CMS has been documented in over 150 plant species and can arise spontaneously in
natural populations or following inter-specific hybridization (Schnable et al. 1998). Although
theoretical work on the population dynamics of plant cytonuclear conflict has focused on
gynodioecious species (reviewed by Saur Jacobs and Wade 2003), any outcrossing
hermaphrodite is potentially susceptible to mitochondrial mutations that increase
female fertility at a cost to pollen fertility. Indeed, the discovery of CMS in many inter-
specific crosses is suggestive of a hidden history of cytonuclear coevolution. With high
selfing, the selective advantage for pollen sterility mutations in the cytoplasm disappears,
since CMS mutants reduce female fertility as well. Because of this, we expect much less
selection on cytonuclear interactions in highly selfing species, and thus the spread of fitness-
reducing pollen sterility mutations should be reduced or eliminated.

Theoretical models of gynodioecy (reviewed in Saur Jacobs and Wade 2003) suggest
that fixation of both the CMS and restorer alleles is a common evolutionary outcome, with
fixation of the restorer allele bringing the population back to hermaphroditism. Evidence for
the exposure of ‘cryptic’ CMS through inter-specific crosses of hermaphroditic plants is consistent with this hypothesis. If distinct restorer and CMS types have been fixed in different species or populations, wide crosses re-expose this evolutionary history. There are several conditions under which CMS and non-CMS (or restored and unrestored) individuals may be maintained in a population leading to stable gynodioecy, including pollen limitation experienced by females as they increase in frequency, and a fitness cost of the restorer allele (Budar et al. 2003). However, historical ‘epidemics’ of CMS and restoration may be the common outcome, and outcrossing hermaphrodites may have a hidden history of CMS-restorer evolution.

In inter-specific crosses, cryptic CMS should be exposed when the maternal parent is outcrossing, but not when it is highly selfing. Recently Fishman and Willis (2006) report cryptic CMS revealed in *Mimulus guttatus* (an outcrosser) X *M. nasutus* (highly selfing) hybrids. They found that hybrid sterility was differentially expressed in the *M. guttatus* cytoplasmic background and that pollen-less anther phenotypes were recovered in F2 hybrids with *M. guttatus* cytoplasm but not in the reciprocal hybrids. The lack of CMS phenotypes in the reciprocal F2 hybrids is consistent with predictions from theoretical models, as we would not expect to find any evidence for a CMS mutant in a highly inbreeding species. Does cryptic CMS exposed in inter-specific crosses generally derive from highly or partially outcrossing species as opposed to highly selfing taxa? This question is difficult to address, due to the limited numbers of species for which quantitative estimates of outcrossing rate are available. It would be interesting to systematically examine pairs of closely related taxa with contrasting selfing rates to test for the degree to which CMS is exposed in inter-specific crosses.
Since CMS most often results from novel chimeric mitochondrial proteins (Delph et al. 2007), we might expect the rate of mitochondrial genome structure evolution to be elevated in outcrossers compared to selfers. In addition, CMS/restorer dynamics may lead to a proliferation of members of the PPR gene family acting as restorers in the nuclear genome. Evidence from genetic mapping of fertility restorers indicates that novel restorers often represent new or mutated forms of a multi-gene family of PPR genes (Brown et al. 2003; Wang et al. 2006).

In addition to CMS dynamics, antagonistic coevolution between the sex functions should generally be reduced in highly selfing populations, due to the absence of conflict when the maternal and paternal parents are the same. Brandvain and Haig (2005) hypothesize that this effect will tend to cause asymmetric hybridization success, since sexual conflict will tend to lead to growth suppression of offspring by the maternal parent and promote offspring growth by the paternal parent. The net effect in inter-specific hybrids is that crosses will be more successful when the selfing species is the maternal parent than in the reciprocal case, since sexual conflict is reduced in the inbreeder. A review of the literature generally supports this model and suggested that genomic imprinting driven by sexual conflict has declined in selfing species (Brandvain and Haig 2005).

**Evolution of genome size**

There is more than 1200 fold variation in nuclear DNA content (C-value) in the angiosperms alone. A variety of phenotypic traits have been shown to be associated with C-value including cellular characteristics such as nucleus and cell size (Mirsky and Ris 1951; Price et al. 1973), duration of cell division (Van’t Hof and Sparrow 1963; Van’t Hof 1965; Bennett 1972), seed size (Beaulieu et al. 2007) and annual or biennial life form (Bennett
1972; Vinogradov 2001). It is possible that genome size has direct effects on these traits and is selected in association with these phenotypes. On the other hand, an alternative class of explanation invokes genetic drift as the primary determinant of genome size. Species with small effective population sizes will experience reduced efficacy of natural selection, leading to the accumulation of slightly deleterious insertions resulting in increased genome size (Lynch and Conery 2003). Finally, genome size evolution may reflect the selective dynamics associated with transposable element evolution and other genomic conflicts, and/or differences among species in the amount and pattern of DNA repair.

In highly selfing species, genome loss may predominate. Proposed evolutionary forces predicted to reduce genome size in selfers include mechanistic explanations such as a lower accumulation of transposable elements and other selfish genetic elements (see above), or increased fixation of large under-dominant deletions as a result of higher homozygosity in selfers (Charlesworth 1992). In addition, faster cell replication and generation time, both of which are negatively associated with C-value, may be selected for in selfing species, leading to an association between selfing and smaller genomes. Because annuals tend to have reduced genome size (Bennett 1972), this would confound effects of selfing and life history, since annuals often display higher rates of selfing than perennials (Barrett et al. 1996). Alternatively, if increased genome size is driven by reduced effective population size (Lynch and Conery 2003), or relaxation of selection against ectopic recombination between transposable elements drives accumulation in selfers, we would expect increased genome size in selfing species.

There is preliminary evidence that mating system may have an effect on genome size evolution. In Veronica, selfers show a significant reduction in genome size with mating system showing a stronger effect than between annual and perennial sister taxa (Albach and
Greilhuber 2004). In their analysis of B-chromosomes and genome size, Trivers and colleagues (2004) also found a strong positive correlation between outcrossing and genome size among UK plants (correcting for B-chromosome presence/absence), although this effect was lost when they used phylogenetically independent contrasts.

A limitation of many comparative studies in relation to mating system is the lack of quantitative estimates of selfing rate. Many comparisons rely on descriptions from floras, usually based on floral morphology. Because most models predict a strong effect of inbreeding only with very high selfing rates, analyses based on morphological inferences could mask effects of mating-system transitions. Here, we re-explore the effect of inbreeding on genome size by presenting a comparison of C-value and genome size in pairs of outcrossing and highly selfing congeners. We refer to DNA amount in a post-replication nucleus; this is equivalent to the 4C DNA amount (C-value). However, it is important to distinguish this measure from genome size, defined as the amount of DNA in a monoploid chromosome set. We therefore report genome size as the 4C DNA amount divided by the ploidy of the sample, consistent with the rationale provided by (Bennett et al. 1998). A species was considered ‘highly’ selfing if the outcrossing rate was less than 10% ($t_m < 0.10$).

We combined two databases of outcrossing rate from Barrett and Eckert (1990) and modified by Igic and Kohn (2006) with the KEW C-value database (http://www.rbgkew.org.uk/cval) and selected those species for which we had both a C-value and an outcrossing estimate or evidence of self-incompatibility. We then identified pairs of congeners with contrasting mating system. From this there were 14 pairs from nine genera, representing eight families including both eudicots and monocots. A sign test was used to test the prediction that predominantly selfing species have smaller genomes.

While there was no significant pattern of C-value reduction in selfers uncorrected for
ploidy, there was a significant reduction in genome size in selfing species (Figure 2.3). This pattern is due to four pairs in which the selfer is polyploid and therefore has a much higher DNA amount. In all four of these pairs the actual genome size is reduced in the polyploid selfer, and there is a significant reduction in genome size across all pairs after controlling for ploidy (sign test, $P = .006$). This is consistent with a reduction in genome size in response to either polyploidization or selfing, or both. Previous work has shown a general reduction in genome size in polyploids (Leitch and Bennett 2004), likely reflecting DNA loss following whole-genome duplication (Ku et al. 2000). If we exclude pairs with polyploid selfers from the analysis both C-value and genome size are significantly reduced in selfers suggesting a reduction in predominantly selfing species independent of ploidal level (sign test, $P = .016$). Further, after excluding those pairs including annual and perennial congeners the pattern of reduced genome size in selfers remains significant (sign test $P = .016$). A weakness of our analysis is that in the contrasts involving monocots two outcrossing species were used repeatedly in comparisons with different selfing species. If selfing evolved independently in each lineage, these should represent independent contrasts. If the selfing species have a shared evolutionary history the analysis will introduce non-independence. However, if we take the average contrast for these groups as single data points, there is still a significant reduction in genome size in the selfing species in our data (sign test $P< 0.01$).

These preliminary results are consistent with a reduction in genome size in self-fertilizing species but they should be interpreted with caution because of the small sample size. Nevertheless, this is the first confirmation of a reduced genome size in selfers using taxonomically-paired contrasts. It suggests that changes in selection associated with genome conflict and/or life-history evolution may be more significant for driving genome evolution in selfers than reduced effective population size. The fact that there are four polyploid selfers
out of the 14 contrasts may indicate that selection against increased DNA amount is not driving the pattern. Instead, changes to the monoploid genome size, such as a reduction in transposable elements, are possibly driving this trend. Also, the general association between polyploidy and selfing (and see Husband et al. 2008) may have obscured the correlation between selfing rate and genome size in previous work, particularly if some plants in the C-value database were incorrectly assumed to be diploid. The role of inbreeding on genome size evolution deserves more thorough investigation as more data become available, but the results to date are consistent with the prediction of increased fitness and reduced deleterious mutation rates via genomic conflict in inbreeding taxa.

Conclusions and future directions

The preliminary evidence suggests a role for mating-system differences in genome evolution. However, the relation between mating system transitions to selfing and reduced fitness associated with the harmful effects of linkage is not clear. Highly inbred taxa may in some cases avoid deleterious mutation accumulation through purging of deleterious mutations, large population sizes, the evolution of higher recombination rates, and the elimination of selfish genetic elements. Although there have been few rigorous comparative or experimental tests of the predictions and patterns outlined above, future investigations should become increasingly tractable. Completion of the genome of A. lyrata, along with the outgroup Capsella rubella will present the first opportunity for large-scale comparison of substitution patterns and genome evolution between closely related species (A. thaliana vs. A. lyrata) with contrasting mating systems. By examining molecular evolutionary rates and genome structure in these species, on a whole-genome scale, there will be extraordinary power to detect differences among species.
Although whole-genome analysis in *Arabidopsis* will be informative, taxonomic replication and a broader sampling of species, particularly using ecological model systems, will be crucial for robust generalizations to be made. This is because studies involving a single comparison will be confounded with the history of the lineages involved. Despite independent evolutionary and coalescent history of genes across the *Arabidopsis* genome, providing a form of evolutionary replication, the entire genome has been influenced by the evolutionary and demographic history of species sampled. Recent evidence suggests historical bottlenecks in some populations of *A. lyrata* (Wright et al. 2003b; Ramos-Onsins et al. 2004), large species-wide effective population size in *A. thaliana* (Nordborg et al. 2005), and the recent and potentially independent evolution of selfing from standing variation in *A. thaliana* (Nasrallah et al. 2004). Collectively, these all point to a complicated demographic history, which may reduce the equivalence of the taxa. More widespread taxonomic replication would allow for a detailed picture of the role of mating-system evolution and its interaction with ecology and demography.

If most mating system transitions have been recent, we need a better theoretical understanding of the timescales and dynamics of genomic changes associated with the transition to selfing. Nearly all of the theory discussed above assumes long-term equilibrium selfing populations, and have not considered the transition to selfing. Such models will be important to understand the factors determining the success and/or extinction of selfing lineages. Species in which there is evidence for multiple independent shifts from outcrossing to selfing should offer opportunities to examine the genomic consequences of mating-system transitions and enable investigation of the interaction of mating patterns and demographic factors (e.g. *Amsinckia spectabilis* - Schoen et al. 1997; *Eichhornia paniculata* - Husband and Barrett 1993). Polymorphism and divergence data should also be integrated with genetic
information on population structure, gene flow, effective populations size and phylogeographic history. It is critical to consider these effects because, while mating system alone can influence effective population size through the processes outlined above, diverse ecological factors can also play a role in shaping population genetic structure. A more comprehensive understanding of genome evolution in plants will require information on the interactions between demography, life history and mating system and how these govern genetic parameters.
Figure 2.1. Expected effects of outcrossing and selfing on patterns of genetic variation and molecular evolution. (A) Effects on linkage disequilibrium, the blue and red circles represent neutral polymorphic mutations, and each pair of lines represents a diploid individual. In outcrossing species, polymorphic sites will often be found in heterozygotes, and crossing over between polymorphic sites can generate novel haplotypes, breaking up linkage disequilibrium. With selfing, although physical crossing over occurs, the low level of heterozygosity leads to an effective reduction in recombination rate, maintaining linkage disequilibrium. (B) Effects on diversity. The yellow circle is an advantageous mutation arising in one chromosome. In outcrossers, the effectively high rate of recombination can uncouple the fate of the advantageous mutation from linked variation, maintaining neutral variation as the advantageous mutation gets fixed. With selfing, the fixation of the advantageous mutation is accompanied by fixation of linked neutral variants, due to hitchhiking. (C) Effects on the fixation of deleterious mutations. Deleterious mutations are shown in black. With outcrossing, the deleterious mutations can be eliminated by selection independently of the fixation of an advantageous mutation. With selfing, fixation of the advantageous mutation can be accompanied by fixation of a linked deleterious mutation, provided that the net selection coefficient is highest on the linkage group with both the advantageous and deleterious mutations.
Figure 2.2. Decay of linkage disequilibrium with physical distance in the highly selfing *Arabidopsis thaliana*, and the self-incompatible *A. lyrata*. The Y axis depicts the average squared correlation coefficient between pairs of polymorphic sites and the X axis represents their physical distance. Data from species-wide samples and are reported in Wright et al. (2006) and Nordborg et al. (2005) for *A. lyrata* and *A. thaliana*, respectively.
Figure 2.3. C-value comparisons for pairs of outcrossing (y-axis) and selfing (x-axis) congeners. Each point represents the genome sizes for a single pair of congeners estimated as the 4C-value divided by the ploidy. The line is the 1:1 line of equality. Note that *Bromus arvensis* and *Hordeum bulbosum* are used as the outcrossing species in comparisons with all selfing congeners in their respective genera. Species pairs are listed with the outcrosser followed by the corresponding selfer and their respective genome size in pg: *Amaranthus cruentus* (1.06), *A. hypochondriacus* (0.96); *Arabidopsis lyrata* (0.94), *A. thaliana* (0.32); *Bromus arvensis* (12.5), *B. japonicus* (11); *B. rubens* (4.85); *B. squarrosus* (11.5); *Cerastium arvense* (1.36), *C. fontanum* (0.73); *Glycine argyrea* (2.56), *G. soja* (2.3); *Hordeum bulbosum* (11), *H. jubatum* (10.8); *H. spontaneum* (11); *H. vulgare* (11.1); *Phaseolus coccineus* (1.36), *P. vulgaris* (1.2); *Plantago lanceolata* (2.4), *P. major* (1.7); *Senecio squalidus* (1.8), *S. vulgaris* (1.58). Genome size data from the Kew C-value database (Bennett and Leitch 2004. Angiosperm DNA C-values database (release 5.0, Dec. 2004) http://www.rbgkew.org.uk/cval/homepage.html). Information on mating systems of species from Barrett and Eckert (1990) and modified by Igic and Kohn (2006).
CHAPTER THREE

EVOLUTIONARY PATHWAYS TO SELF-FERTILIZATION IN A TRISTYLOUS PLANT SPECIES

This chapter resulted from a collaboration with Spencer C.H. Barrett and Mario Vallejo-Marín. Spencer C.H. Barrett contributed ideas and to the writing of the manuscript. Mario Vallejo-Marín contributed data collection and analysis of crosses conducted in the glasshouse and to the writing of the manuscript which was published in New Phytologist, 2009 183:546-556.

Summary

Evolutionary transitions from outcrossing to selfing occur commonly in heterostylosous genera. The morphological polymorphisms that characterize heterostyly provide opportunities for different pathways for selfing to evolve. Here, we investigate origins and pathways by which selfing has evolved in tristylos Eichhornia paniculata by providing new evidence based on morphology, DNA sequences and genetic analysis. The primary pathway from outcrossing to selfing involves stochastic loss of the S-morph from trimorphic populations followed by the spread of selfing variants of the M-morph. However, the discovery of selfing variants of the L-morph in Central America indicates a secondary pathway and distinct origin for selfing. Comparisons of multi-locus nucleotide sequences from 27 populations sampled from throughout the geographical range suggest multiple transitions to selfing. Genetic analysis of selfing variants of the L- and M-morphs demonstrate recessive control of the loss of herkogamy, although the number of factors appears to differ between the forms. Early stages in the establishment of selfing involve developmental instability in the formation of flowers capable of autonomous self-pollination. The relatively simple genetic control of herkogamy reduction and frequent colonizing
episodes may often create demographic conditions favouring transitions to selfing in *E. paniculata*.

**Introduction**

In many herbaceous plants morphological and physiological traits that function to reduce the incidence of inbreeding have been lost, leading to high rates of self-fertilization and evolution of the “selfing syndrome” (Lloyd 1965; Ornduff 1969; Ritland and Ritland 1989; Armbruster et al. 2002). The origin of high levels of self-fertilization (autogamy) from obligate cross-fertilization has been considered the most frequent evolutionary transition in flowering plants (Stebbins 1974). There are no estimates of the number of origins of selfing but it is likely that this transition has occurred many hundreds of times. Most transitions probably go undetected because selfing lineages are often short-lived and usually occur at the tips of phylogenetic trees (Schoen et al. 1997; Takebayashi and Morrell 2001; Ige et al. 2008).

The paradox of why selfing should evolve, given its limited evolutionary future and the generally poorer performance of selfed offspring relative to outcrossed offspring, has been an enduring source of curiosity for evolutionary biologists. Charles Darwin (1876) drew attention to the reproductive advantage that selfing individuals have when pollinators are scarce owing to “reproductive assurance”. Later, Fisher (1941) pointed out that genes that increase the rate of selfing have a transmission bias when they arise in an outcrossing population (Jain 1976). Since then this transition has been the focus of considerable theoretical and empirical research aimed at understanding how and why selfing evolves from outcrossing (reviewed in Lloyd 1980; Uyenoyama et al. 1993; Holsinger 1996).
Selfers are represented in many floras, especially those in which there is a marked dry season and an abundance of ephemeral habitats. Selfing species are commonly annual and many have prolific colonizing ability and extensive geographical ranges (Jain 1976; Lloyd 1980; Barrett et al. 1996). This is associated with the ability of individuals to establish at low density or following long-distance dispersal (Darwin 1876; Baker 1955; Pannell and Barrett 1998). The evolution of selfing from outcrossing can therefore have important ecological, demographic and biogeographic consequences.

Heterostylous groups provide a rich source of reproductive diversity for investigating the evolution of selfing. In many heterostylous taxa, obligate outcrossing, enforced by heteromorphic self-incompatibility, has been replaced by predominant selfing as a result of the origin and spread of self-compatible homostyles (distyly) or semi-homostyles (tristyly). These floral forms have the capacity for autonomous self-pollination because they either have all anthers (homostyles), or one set of anthers (semi-homostyles), closely adjacent to stigmas. The transition to predominant selfing in heterostylous groups is evident from phylogenetic analysis (e.g. Pontederiaceae - Kohn et al. 1996; Amsinckia - Schoen et al. 1997; Primula - Mast et al. 2006), and at the intra-specific level through population-level studies (Crosby 1949; Ornduff 1972; Ganders 1975). Homostyles often occur at range margins, or on islands, a finding consistent with their ability to produce seed in the absence of pollinators or mates (Baker 1966; Barrett 1985b; Barrett and Shore 1987). This pattern implicates reproductive assurance as playing an important role in the selection of homostyly.

Heterostylous populations are reproductively sub-divided into two (distyly) or three (tristyly) morphologically distinct mating groups. The pathways from outcrossing to selfing therefore have the potential to be more diverse than in non-heterostylous species because of sexual polymorphism. For example, the floral morphs could participate equally in the breakdown
process, or, as occurs in some distylovous taxa (e.g. Primula - Charlesworth and Charlesworth 1979; Turnera - Barrett and Shore 1987), a particular floral morph may be favoured because of distinctive features of its morphology, genetics, or mating ability. The morphological and physiological polymorphisms that characterize the heterostylous syndrome therefore provide opportunities for different mechanisms and pathways for selfing to evolve.

Here, we review what is known about the evolution of selfing from outcrossing in Eichhornia paniculata (Pontederiaceae), an annual Neotropic tristylovous species. Populations of E. paniculata occupy ephemeral aquatic habitats, including ponds, drainage ditches, wet pastures and rice fields. The main concentration of populations occurs in N.E. Brazil, with smaller foci in Jamaica and Cuba and isolated disjunct populations in Nicaragua and Mexico. In tristylovous species the three floral forms are known as the long-, mid- and short-styled morphs (hereafter L-, M-, S-morphs). Eichhornia paniculata has been the subject of extensive research on the ecological and demographic factors responsible for mating-system variation (reviewed in Barrett et al. 1992). To provide a context for our current investigations we briefly summarize earlier studies on the evolutionary processes responsible for the breakdown of tristyly. We then present new biogeographical and molecular genetic evidence on the evolutionary pathways by which selfing has evolved based on a wide geographical sampling of populations. Also, using controlled crosses we examine the inheritance of mating-system modification and describe the expression of selfing modifiers when they first appear in populations.

Evolutionary pathways to selfing

Large-scale surveys of the frequency of style morphs in populations of E. paniculata in N.E. Brazil and Jamaica have provided evidence for a primary evolutionary pathway from
outcrossing to selfing that has been followed repeatedly (Barrett et al. 1989; Husband and Barrett 1993; Figure 3.1). In N.E. Brazil stylar trimorphism is the commonest condition, followed by dimorphism and monomorphism. Most dimorphic populations are missing the S-morph and contain the L-morph and selfing variants of the M-morph (Figure 3.2A). Monomorphic populations are dominated by selfing variants of the M-morph. Surveys indicate that ~30% of populations are missing at least one style morph and genetic drift and founder events are largely responsible for the loss of morphs from populations (Husband and Barrett 1992a; 1992b). Populations missing morphs are significantly smaller than those that are tristylos and the observed pattern of S-morph loss is that predicted by stochastic models of the influence of finite population size on the two-diallelic locus (S, M) genetic system controlling tristyly (reviewed in Barrett 1993). Style morph frequencies on the island of Jamaica are distinct from those in N.E. Brazil (Barrett 1985b; Barrett, et al. 1989; Husband and Barrett 1991). No tristylos are known in Jamaica and most populations are composed exclusively of selfing variants of the M-morph. A few populations contain low frequencies of the L-morph. Recent surveys of style morph frequencies in Jamaica (2007) and Cuba (2008) confirm these overall patterns; most populations were fixed for selfing variants of the M-morph and a low frequency of the L-morph occurred in a single population on both islands (Table 1). The absence of the S-morph from the Caribbean is likely the result of a founder event associated with long-distance dispersal from Brazil.

Of particular significance is the observation that in dimorphic populations in both Jamaica and Cuba the L-morph exhibits conspicuous herkogamy. This limits opportunities for autonomous self-pollination in contrast to selfing variants of the M-morph. The morph-specific difference in herkogamy has a profound influence on mating patterns and fertility of the two morphs. Plants of the L-morph are largely cross-pollinated by the M-morph but have
low seed set, whereas those of the M-morph produce large quantities of self-fertilized seed (see Figure 1 in Barrett et al. 1989). Theoretical models of selection in dimorphic populations indicate that this asymmetrical mating system is difficult to maintain and that fixation of selfing variants occurs under most conditions (Barrett et al. 1989; Pannell and Barrett 2001).

Collectively these style morph surveys, in combination with field studies of fitness components support a pathway involving two main stages: 1) stochastic loss of the S-morph from trimorphic populations through drift and founder events; 2) selective loss of the L-morph by spread and fixation of self-pollinating variants of the M-morph as a result of automatic selection and reproductive assurance. The pathway to selfing represents an example of how a transition from outcrossing to selfing can be triggered by genetic drift. It has been identified as one of the few cases that meet several of the key conditions in Sewall Wright’s shifting balance theory of evolution (Coyne et al. 1997).

Additional exploration in Central America has revealed a second pathway for the evolution of selfing in *E. paniculata* (Figure 3.1). Single isolated populations in Nicaragua and Mexico were both composed of a single floral phenotype not observed elsewhere in the range. At the present time only a single population is known from Mexico and three from Nicaragua (A. Novelo R, personal communication). Plants in both populations that were sampled are semi-homostylous forms of the L-morph, rather than the M-morph (Figure 3.2B). They possess purple pigmentation of styles, a feature typical of the L-morph, and when crossed to homozygous plants of the M-morph (*ssMM*) the progeny are uniformly mid-styled. This demonstrates that they are derived from the L-morph rather than the other two style morphs (S.C.H. Barrett, unpublished data). The Nicaraguan and Mexican populations exhibit the smallest flowers observed in *E. paniculata* and “mid-level anthers” in these populations are in contact with stigmas of the long style. Plants from these populations
grown under pollinator-free glasshouse conditions at Toronto produce 100% fruit set as a result of autonomous self-pollination. Although we have not measured mating patterns in these populations they are likely to be highly autogamous, given their strong facility for autonomous self-pollination. The origin of these semi-homostylous L-morph populations is unclear. They could be descended from a single long-distance dispersal event from N.E. Brazil or, more likely, from a dimorphic population from the nearby Caribbean islands of Jamaica or Cuba. Molecular evidence presented below is equivocal on this question.

**Multiple origins of selfing**

*Evidence from genetic relations among populations*

Patterns of allozyme variation at 24 loci from 44 populations of *E. paniculata* from N.E. Brazil provided evidence for multiple origins of selfing in the M-morph (Husband and Barrett 1993). Selfing populations were dispersed across a dendrogram of genetic relationships and, for the most part, were more genetically similar to nearby trimorphic populations than to more geographically distant selfing populations. Here, we present additional evidence for multiple transitions to selfing using DNA sequence data from 10 EST-derived nuclear loci sampled from 4-12 individuals from each of 27 populations, for a total of 229 individuals (for detailed localities of populations sampled see Table S1). The populations we investigate here represent a broader geographic sample and involve different populations from those we previously investigated. We have not estimated outcrossing rates in these populations but assume, based on our previous work, that the morph structure of populations (trimorphic, dimorphic, monomorphic), especially the occurrence of self-pollinating variants in all monomorphic populations that we sampled, has a major influence on their mating patterns (see Barrett and Husband 1990; Barrett et al. 1992).
Using the program SplitsTree (Huson and Bryant 2006) we estimated the network of genetic relationships among populations by randomly sampling one individual per population and concatenating all 10 loci (Figure 3.3). In a neighbour-network single nucleotide polymorphisms (SNPs) that are not consistent with a simple bifurcation process, either due to incomplete lineage sorting or recombination, are represented by reticulation in the network. This method is more appropriate for depicting the genetic relationships among sequences for a species with recombination than methods of historical reconstruction that assume a bifurcating process of divergence. To ensure that our results were robust we also generated the neighbour network for all 229 individuals in our sample. The resulting network was qualitatively similar to that displayed in Figure 3.3. We also ran the SplitsTree program for each of the 10 loci individually and in seven of the 10 cases the same overall topology was evident. The three instances where this was not the case involved sequences with fewer informative segregating sites. In addition, we tested other phylogenetic methods [e.g. Bayesian analysis (Mr. Bayes), Neighbour joining, and parsimony (PAUP*)] by concatenating all sequences for each individual. The trees obtained from each of these methods were concordant with the overall patterns obtained in Figure 3.3 with four deep clusters separating the major geographical regions occupied by *E. paniculata*.

Within N.E. Brazil, populations are geographically subdivided into northern and southern ranges separated by an extensive arid zone (see Figure 2 in Barrett et al. 1989). This subdivision was evident in our data with a split in the network (78.1% bootstrap support) that likely reflects a barrier to gene flow imposed by arid conditions inimical for aquatic plant growth. Within the southern cluster, selfing monomorphic populations (B182, B183) were separated on the tree and each clustered with nearby dimorphic and trimorphic populations.
However, bootstrap support for relationships among the Brazilian populations in the southern range was weak, perhaps reflecting limited divergence and/or gene flow.

Populations from Jamaica and Cuba formed a distinct cluster (100% bootstrap support) indicating that they are likely to be descended from a single dispersal event to the Caribbean from mainland South America. Lastly, population samples from Mexico and Nicaragua formed a fourth cluster separated from the Brazilian and Caribbean populations by a long branch indicating a long period of isolation. The dispersed distribution of monomorphic populations across the network is consistent with the occurrence of multiple transitions from outcrossing to selfing.

The distribution of SNPs among populations of *E. paniculata* provides additional support that selfing has likely arisen on multiple occasions. If populations of selfers are derived from a particular source region they are likely to be less differentiated from that source than from other selfing populations that are independently derived from a different region. To test this proposition, we used alignments of all 10 loci for all individuals excluding the Central American plants, for which we did not have adequate population samples. Using the software package SITES (Hey and Wakeley 1997) we estimated $F_{st}$ for each locus among all pairs of populations (Figure 3.4A). Average pairwise differentiation ($F_{st}$) among selfing populations from the Caribbean (CRB) ($F_{st\, \text{CRB}} = 0.402$, SE = 0.053) is lower than differentiation of these populations from selfing populations from Brazil (BRA) ($F_{st\, \text{CRB vs. BRA}} = 0.493$, SE = 0.095), but higher than differentiation of Caribbean populations from outcrossing populations from Brazil (OUT) ($F_{st\, \text{CRB vs. OUT}} = 0.338$, SE = 0.087).

Furthermore, the two monomorphic selfing populations from Brazil (B182 and B183) are more differentiated from one another than each is to trimorphic outcrossing populations ($F_{st\, \text{B182 vs. B183}} = 0.562$, SE = 0.088; $F_{st\, \text{B182 vs. OUT}} = 0.284$, SE = 0.060 and $F_{st\, \text{B183 vs. OUT}} = 0.227$, SE = 0.060).
Lastly, selfing populations from the Caribbean are more differentiated from Brazilian selfing populations (BRA) than to outcrossing populations from Brazil \((F_{ST}^{BRA \text{ vs. CRB}} = 0.493, SE = 0.095)\). These patterns would be predicted if different portions of the allelic variation segregating in outcrossing Brazilian populations were sampled during the origin of selfing populations in the Caribbean and N.E. Brazil.

The proportion of fixed, shared and unique SNPs in these populations provide further insight on the patterns of differentiation (Figure 3.4B). 93.4% of nucleotide polymorphisms in selfing populations from Brazil were shared with outcrossing populations and there were no fixed differences between the groups. This pattern is consistent with the hypothesis that these populations are of relatively recent origin and represent early stages in the breakdown of tristyly. Morphological studies of these populations support this hypothesis (Vallejo-Marín and Barrett 2009 and see below). Comparisons of nucleotide variation between selfing populations from the Caribbean and outcrossing populations from Brazil indicated that these populations share only 15% of nucleotide variation and there were three fixed differences between the two regions (Figure 3.4B). Further, comparisons of selfing population from the Caribbean versus Brazil indicate that populations from the two regions have more fixed differences (11) and share few polymorphisms (9.1%). These patterns indicate that most of the variation in selfing populations represents different sub-samples of variation maintained among outcrossing populations, a pattern consistent with multiple origins.

**Evidence from genetic control of mating-system modification**

Fenster and Barrett (1994) used controlled crosses of selfing and outcrossing phenotypes of the M-morph from Brazil and Jamaica to investigate the inheritance of variation in herkogamy, the principle floral trait governing selfing rates in *E. paniculata*. 
Their results indicated that the loss of herkogamy causing selfing was fully recessive to the ancestral outcrossing condition. They also provided evidence that different recessive modifiers were responsible for the loss of herkogamy, based on crosses among selfing phenotypes from different regions, a finding consistent with multiple origins of selfing.

To investigate further the genetic basis for the loss of herkogamy we conducted controlled crosses of outcrossing and selfing phenotypes of the M- and L-morphs. Here, we present selected crosses to illustrate general patterns. A more detailed treatment will be presented elsewhere (M. Vallejo-Marín and S.C.H. Barrett, unpublished results). For the M-morph we crossed an individual with a single stamen adjacent to the stigma (Figure 3.2A) from a monomorphic selfing population to an unmodified M-morph from a trimorphic outcrossing population. Both plants were from N.E. Brazil. In the case of the L-morph, we used a semi-homostylosous individual of the L-morph from Mexico (Figure 3.2B) and crossed it to an unmodified (outcrossing) L-morph from a trimorphic population in Brazil. Crosses were performed in both directions, using plants in each cross as maternal and paternal parents. To generate an F2 population, we self-fertilized a single F1 plant per cross type, including separate F1 parents for crosses done in both directions. Because we found no evidence that the direction of crosses had a significant influence on the distribution of herkogamy variation we have pooled data presented in Figure 3.5.

Our results support previous findings that the loss of herkogamy in the M-morph is controlled by recessive factors. F1 offspring strongly resembled the outcrossing parent with respect to their herkogamy values (Figure 3.5). Under a simple model of monogenic or oligogenic inheritance of herkogamy modification, we would expect that a fraction of the F2 would recover the parental selfing phenotype. The distribution of herkogamy values in the F2 indicates a discontinuous distribution showing two main classes of plants, one with
unmodified flowers with a mode near the outcrossing parent (>2mm), and a second class of individuals with herkogamy values (< 2mm) closer to the selfing parent (Figure 3.5A). The segregation ratio for these two classes of plant was 142:17, respectively, which is statistically different from Mendelian expectations for either one ($\chi^2=17.36$, 1 d.f., $P < 0.001$) or two ($\chi^2=5.35$, 1 d.f., $P = 0.02$) recessive gene(s) controlling the loss of herkogamy.

Crosses between selfing and outcrossing phenotypes of the L-morph also indicate that the loss of herkogamy involves the action of recessive factors. Herkogamy values in the F1 resembled the outcrossing parent (Figure 3.5B). The distribution of herkogamy values in the F2 generation was continuous but included two modes, one roughly intermediate between the two parental values, and a second close to values for the selfing parent. This distribution suggests quantitative genetic control but involving the segregation of both minor and major genes.

**Developmental instability in the formation of selfing flowers**

An unusual feature of the early stages of the evolution of selfing in *E. paniculata* is the occurrence of plants exhibiting developmental instability in the production of flowers capable of autonomous self-pollination (Barrett 1985a; Richards and Barrett 1992; Seburn et al. 1990; Barrett and Harder 1992; Vallejo-Marín and Barrett 2009). This conspicuous within-plant variation is a feature of the M-morph only. In some flowers usually one, sometimes two, short-level stamens elongate to a position close to mid-level stigmas resulting in self-pollination. In the remaining flowers there is no modification to the positions of short-level stamens. This phenomenon results in a bimodal distribution of herkogamy values within a plant and has the potential to promote mixed mating. To our knowledge discontinuous variation in stigma-anther separation is unique among angiosperm species and
raises the question of the proximate mechanisms responsible and their ecological and evolutionary consequences.

The genetic basis of developmental instability in *E. paniculata* is not well understood but it seems probable that it is associated with inbreeding. To initiate investigations on the inheritance of developmental instability we self-pollinated an individual of the M-morph that displayed a bimodal distribution of herkogamy values for two generations. In the S₂ generation (*N* = 75 plants) we measured stigma-anther separation on an average of 9.7, SE = 0.25 flowers per plant and classified them as either modified (herkogamy < 2mm) or unmodified (herkogamy ≥ 2mm). The frequency distribution of herkogamy values among all flowers produced by S₂ plants was strongly bimodal (Figure 3.6A) resembling the distribution observed in the parental plant (see B189-7-1 Figure 5 in Vallejo-Marín and Barrett 2009). However, despite all S₂ plants sharing the same level of inbreeding there was considerable variation among individuals in the proportion of modified flowers. The majority of plants (67%) produced only one class of flower, either modified (31%) or unmodified (36%) while the remaining plants (33%) displayed instability in herkogamy (Figure 3.6B).

The expression of herkogamy in *E. paniculata* is also influenced by environmental factors. Using cloned genotypes grown under different resource levels, Barrett and Harder (1992) and Vallejo-Marín and Barrett (2009) were able to modify the frequency of self-pollinating flowers. Under more stressful conditions plants produced more flowers with near zero herkogamy values. Vallejo-Marín and Barrett (2009) further demonstrated genetic variation in the level of plasticity in stigma-anther separation using genotypes sampled from trimorphic, dimorphic and monomorphic populations from N.E. Brazil. Reductions in water availability, nutrients, and pot size resulted in significantly smaller herkogamy values in some but not all genotypes. Plants from trimorphic and monomorphic populations exhibited
much less developmental instability in herkogamy. In contrast, plants from dimorphic populations produced a mixture of flowers with high versus low herkogamy values. These results obtained under uniform glasshouse conditions confirm field observations indicating that developmental instability is most common in dimorphic populations where selfing modifiers first spread.

**Discussion**

“Some species exist under two forms, the one bearing conspicuous flowers adapted for cross-fertilisation, the other bearing inconspicuous flowers adapted for self-fertilisation”

Charles Darwin 1876 p. 445

As recognized by Darwin (1876), some plants exhibit intra-specific variation in mating system including predominant outcrossing and high levels of self-fertilization. In species maintaining outcrossing and selfing forms it is of interest to determine the pathway(s) by which selfing has evolved and how often it has originated. Although questions related to the origins of selfing have often been addressed using species-level phylogenies (e.g. Kohn et al. 1996; Barrett et al. 1996; Schoen et al. 1997; Goodwillie 1999), there have been fewer efforts to consider how often the shift to selfing may occur within species (Allen et al. 1991; Wyatt et al. 1992), presumably because species displaying wide intra-specific variation in mating patterns are relatively uncommon.

Biogeographical surveys have revealed two distinct selfing forms of *E. paniculata* independently derived from the L- and M-morphs of heterostylosous populations (Figures 3.1 and 3.2). Over most of the geographical range the principal form involves modified variants of the M-morph with one (Figure 3.2A) to three short-level stamens close to mid-level stigmas. Semi-homostylosus M-morphs are reported in several other tristylos species and
appear to be the commonest type of breakdown product (Stout 1925; Ornduff 1972; Barrett 1988). Nevertheless, the two populations of *E. paniculata* from Central America were composed exclusively of the semi-homostyloous L-morph and to our knowledge these forms occur nowhere else. Significantly, both of the semi-homostyloous forms of *E. paniculata* have also been described from the related *E. heterosperma* and *E. diversifolia* (Barrett 1988) where they co-exist within populations rather than occurring allopatrically, as in *E. paniculata*. Although polymorphism involving two selfing forms could potentially occur in dimorphic populations of *E. paniculata*, the L-morph in these populations exhibits well-developed herkogamy (Figure 3.1). Why are selfing variants of the L-morph absent from Jamaica and Cuba, the closest concentrations of populations to Central America? This question is particularly perplexing since the ecological conditions on these islands clearly favour selfing, as indicated by the predominance of semi-homostyloous M-plants.

The genetic architecture of tristyloous and semi-homostyloous morphs may help to explain these puzzling geographical patterns. The most probable historical scenario is that Cuba or Jamaica was originally colonized from Brazil by a self-pollinating heterozygous variant of the M-morph (*ssMm*) with the L-morph (*ssmm*) then arising through segregation. The very similar patterns of nucleotide variation among populations from these islands suggest that they are descended from a single long-distance dispersal event (Figure 3.3). In dimorphic populations, the semi-homostyloous M-morph is predominantly selfing whereas the L-morph is largely outcrossed by pollen from long-level anthers of the M-morph (Barrett et al. 1989; Figure 3.1). This asymmetrical mating pattern combined with M-morph abundance will cause the recessive alleles governing the L-morph to spend much of their time sheltered from selection in heterozygous M-genotypes. Such an effect should reduce the intensity of selection for selfing on the L-morph as well as guarantee a long persistence time of this
morph on these islands despite its lower fertility in comparison to the M-morph (e.g. Haldane 1924).

Studies of inheritance indicate that selfing modifiers causing the loss of herkogamy in the M-morph are morph-limited in expression (Fenster and Barrett 1994; Vallejo-Marín and Barrett 2009). Although the alleles are transmitted to the L-morph through segregation in dimorphic populations they do not alter the stigma-anther separation or mating patterns of this morph. Selfing in the semi-homostylous L-morph involves different genetic modifiers affecting the position of mid-level rather than short-level stamens. These modifiers were only evident in Central American populations. This indicates that selfing has arisen by two distinct genetic pathways in *E. paniculata*, although in both cases the modifications are under recessive gene control (Figure 3.5). Favourable recessives have a low chance of establishment in large random mating populations because of the rarity of homozygotes for low frequency alleles (“Haldane’s Sieve” – Haldane 1924). However, in populations with moderate to high rates of self-fertilization there is both theoretical and empirical evidence that favourable recessive mutations can play a role in the evolution of adaptations (Charlesworth 1992). Genetic drift can also result in chance fixation of recessive alleles in small populations because of reductions in effective population size (Pollak 1987). As discussed earlier, both selfing and drift are characteristic features of *E. paniculata* populations and the finding that selfing modifiers in both the L- and M-morph are recessive is therefore not entirely unexpected.

Long-distance dispersal of the L-morph to Central America would result in monomorphic populations because self-fertilization of this morph can only result in L-morph progeny. In Central America strong selection for reproductive assurance may have led to the evolution of semi-homostyloous forms of the L-morph. It is possible that in contrast to
Jamaica and Cuba the absence of the M-morph from this region may have fostered the evolution of semi-homostyly in the L-morph. Selfing modifiers may be more rapidly fixed in the M-morph than the L-morph because fewer alleles of larger effect appear to be involved (Figure 3.5). According to this hypothesis the dynamics of selfing evolution in *E. paniculata* may be conditional on which morphs are represented in populations because of the different genetic pathways involved in the origins of selfing. However, the coexistence of both selfing morphs within populations of related *Eichhornia* species (Barrett 1988) argues against this interpretation unless their co-occurrence has arisen secondarily after polytypic origins of semi-homostyly.

The patterns of molecular genetic variation among our sample of 27 populations of *E. paniculata* are consistent with multiple independent transitions to self-fertilization. However, our analyses cannot provide a concrete estimate of the number of independent origins. Gene flow among populations could allow recombination between nuclear markers and the loci responsible for the selfing phenotypes resulting in a decoupling of the evolutionary history of loci within the genome. However, our data on the distribution of SNPs among populations, and the proportion that are fixed, shared and unique, indicate that populations containing selfing variants have likely been derived from outcrossing populations in different parts of the geographical range, a pattern repeated across 10 presumably putatively unlinked nuclear markers. This result, in conjunction with the occurrence of two distinct semi-homostylous phenotypes derived independently from the L- and M-morphs, and genetic evidence indicating different selfing modifiers in geographically distinct populations, argues strongly against a single origin for selfing in *E. paniculata*. However, it is possible to explain the patterns of molecular genetic differentiation that we observed with alternative scenarios in which all of the selfing populations derived from the M-morph resulted from a single origin.
This is because differentiation among selfing populations is likely to have been affected after the evolution of selfing by gene flow, as well as drift and isolation. These factors will complicate inferences on the origins of selfing from patterns of differentiation alone. Although we cannot confidently estimate the number of independent transitions to selfing with our data, especially in N.E. Brazil where gene flow is most likely, it seems unlikely that all selfing populations are descendants of a single ancient transition to selfing in *E. paniculata* based on the different lines of genetic evidence presented here.

Developmental instability is determined by a variety of factors including inbreeding and homozygosity at regulatory loci, mutation, breakdown of adapted gene complexes, and various environmental stressors (Polack 2003). Early stages in the establishment of selfing involve developmental instability in the production of flowers capable of autonomous self-pollination. Plants of the M-morph in dimorphic populations carry selfing modifiers producing both unmodified and modified flowers. Our studies have demonstrated both genetic and environmental components to this variation. The widespread occurrence of developmental instability in dimorphic populations of *E. paniculata* raises the issue of whether it is maladaptive, as is often assumed, or whether this form of within-plant variation represents a mixed strategy enabling plants to adjust their mating to match heterogeneous environmental conditions. Although we do not know the answer to this question in *E. paniculata* the relatively simple genetic basis for selfing modifiers, frequent colonizing episodes, and the species capacity for long-distance dispersal seem likely to often create conditions favouring transitions to selfing.
Table 3.1. Style morph frequencies and the size of *Eichhornia paniculata* populations on the islands of Jamaica and Cuba surveyed in 2007 and 2008, respectively. Surveys were conducted following methods described in Barrett et al. (1989).

<table>
<thead>
<tr>
<th>Population</th>
<th>Locality</th>
<th>Habitat</th>
<th>Pop. Size</th>
<th>L-morph</th>
<th>M-morph*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jamaica</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J28</td>
<td>Treasure Beach, St. Elizabeth</td>
<td>wet pasture</td>
<td>200</td>
<td>0.03</td>
<td>0.97</td>
</tr>
<tr>
<td>J29</td>
<td>Fullerswood, St. Elizabeth</td>
<td>marsh</td>
<td>28</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>J30</td>
<td>Catabo, St. Elizabeth</td>
<td>seasonal pond</td>
<td>20</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>J31</td>
<td>Slipe, St. Elizabeth</td>
<td>wet pasture</td>
<td>4</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>J32</td>
<td>Little London, Westmoreland</td>
<td>sugar cane field</td>
<td>25</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>Cuba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>Yara, Granma</td>
<td>ricefield</td>
<td>500</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>C2</td>
<td>Manzanillo, Granma</td>
<td>ricefield</td>
<td>800</td>
<td>0.15</td>
<td>0.85</td>
</tr>
<tr>
<td>C3</td>
<td>Chorerra, Granma</td>
<td>ricefield</td>
<td>120</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>C4</td>
<td>Baracoa, Guantánamo</td>
<td>drainage ditch</td>
<td>10</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>C5</td>
<td>Camelote, Camagüey</td>
<td>ricefield</td>
<td>600</td>
<td>0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

* All plants of the M-morph are semi-homostylos
Figure 3.1. Evolutionary pathways from cross-fertilization to self-fertilization in tristyloous *Eichhornia paniculata* via the origins of semi-homostyly. The primary pathway from trimorphism to dimorphism culminates in the fixation of selfing semi-homostylovous forms of the M-morph. A less common pathway leads to populations monomorphic for semi-homostylovous forms of the L-morph. Arrows linking anthers and stigmas indicate mating combinations, those linking floral phenotypes indicate evolutionary transitions. The trend to smaller reproductive organs with increased selfing reflects reductions in flower size.
Figure 3.2. Two selfing variants of *Eichhornia paniculata* illustrating their contrasting positions of sexual organs. (A) In the M-morph the first stage in the transition to selfing involves elongation of a single stamen from short-level stamens so that it takes up a position next to mid-level stigmas resulting in autonomous self-pollination. Plants with this phenotype are very common in dimorphic populations from N.E. Brazil where the individual illustrated originated. (B) Semi-homostyloous flower of the L-morph from Mexico. All three stamens are adjacent to the stigma resulting in autonomous self-pollination. Plants from Nicaragua possess sexual organs in similar positions.
Figure 3.3. Neighbour network of 10 EST derived nuclear loci displaying relationships among 27 populations of *Eichhornia paniculata* sampled throughout the geographical range. Population codes are indicated at branch tips. Brazilian, Cuban, Jamaican, Mexican and Nicaraguan populations are represented by codes that begin with B, C, J, M and N, respectively. The morph structure of populations is indicated with blue squares, green triangles or red circles representing trimorphic, dimorphic or monomorphic populations, respectively. We constructed the network with the program SplitsTree (Huson and Bryant 2006) by randomly sampling one individual per population and concatenating the sequences from all 10 loci. Bootstrap values from 10,000 replicates are displayed for nodes with greater than 70% support.
Figure 3.4. Pairwise distribution of single nucleotide polymorphisms (SNPs) from 10 EST-derived nuclear loci from outcrossing (OUT, N = 111 individuals) and selfing populations of *Eichhornia paniculata*. Selfing populations from Jamaica and Cuba are combined (CRB, N = 64 individuals) and Brazilian populations (B182 and B183, N = 10 and 12 individuals, respectively) are combined in pairs as BRA. The first population group listed in the axis labels is Group 1 and the second Group 2. A) Mean pairwise differentiation ($F_{ST}$) between population groups based on alignments of 10 sequences. Each point represents the mean across all loci and populations pairs with error bars indicating ±1 SE. (B) Proportion of unique and shared SNPs and fixed differences between population groups. We calculated all statistics using the program SITES (Hey and Wakeley 1997). Nucleotide sequences were collected from 4-12 individuals per population for a total of 6678bp per individual.
Figure 3.5. (A) Inheritance of stigma-anther separation (herkogamy) in the M-morph of *Eichhornia paniculata* from N.E. Brazil. Two individuals of the M-morph (parental generation) representing contrasting selfing and outcrossing phenotypes were used to generate two filial generations (F1 and F2). The parental plant on the left represented a “modified” (selfing) phenotype (M’) collected from a monomorphic population with values of herkogamy near zero. The parental plant on the right represented an “unmodified” phenotype (M) with higher herkogamy values (~3mm) from a tristylos population. (B) Inheritance of stigma-anther separation (herkogamy) in the L-morph of *E. paniculata* from Mexico and N.E Brazil. The Mexican plant was semi-homostylos (L’) with no stigma-anther separation (Figure 3.2B), the Brazilian plant was an unmodified L-morph (L) from a tristylos population. The histograms in (A) and (B) illustrate the frequency distribution of herkogamy values for offspring from each self or cross type. The F1 was produced by crossing the two parents and the F2 by selfing a single F1 individual from each cross-type. P1 and P2 illustrate the segregation of herkogamy variation following self-fertilization in the modified and unmodified parents, respectively. Samples sizes for flowers measured per plant were 2.76±0.08, mean±SE, for the F1, P1 and P2; and 9.59±0.13 for the F2.
Figure 3.6. Developmental instability of stigma-anther separation (herkogamy) in *Eichhornia paniculata* from N.E. Brazil. (A) Bimodal distribution of herkogamy values in 724 flowers from 75 individuals in the S$_2$ generation. This line is the product of two generations of selfing by single seed descent of an individual (B189-7-1) that displayed a bimodal distribution of herkogamy values (see Figure 5 in Vallejo-Marín and Barrett 2009). (B) The average proportion of modified flowers relative to total flowers ($N=9.66$, SE = 0.25) in each plant. Developmental instability occurs in individuals producing both “modified” and “unmodified” flowers. Among the 75 plants, 33.3% displayed developmental instability.
CHAPTER FOUR

MATING-SYSTEM VARIATION, DEMOGRAPHIC HISTORY AND PATTERNS OF NUCLEOTIDE DIVERSITY IN THE TRISTYLOUS PLANT
EICHORNIA PANICULATA

This chapter resulted from a collaboration with Stephen I. Wright and Spencer C. H. Barrett. Stephen I. Wright contributed ideas and to the writing of the manuscript. Spencer C. H. Barrett contributed ideas and to the writing of the manuscript which was published in Genetics, 2010 184:381-392.

Summary

Inbreeding in highly selfing populations reduces effective size and, combined with demographic conditions associated with selfing, this can erode genetic diversity and increase population differentiation. Here we investigate the role that variation in mating patterns and demographic history play in shaping the distribution of nucleotide variation within and among populations of the annual neotropical colonizing plant Eichhornia paniculata, a species with wide variation in selfing rates. We sequenced 10 EST-derived nuclear loci in 225 individuals from 25 populations sampled from much of the geographic range and used coalescent simulations to investigate demographic history. Highly selfing populations exhibited moderate reductions in diversity but there was no significant difference in variation between outcrossing and mixed mating populations. Population size interacted strongly with mating system and explained more of the variation in diversity within populations. Bayesian structure analysis revealed strong regional clustering and selfing populations were highly differentiated based on an analysis of $F_{st}$. There was no evidence for a significant loss of within-locus linkage disequilibrium within populations, but regional samples revealed greater breakdown in Brazil than in selfing populations from the Caribbean. Coalescent simulations indicate a moderate bottleneck associated with colonization of the Caribbean from Brazil
~125,000 years before present. Our results suggest that the recent multiple origins of selfing in *E. paniculata* from diverse outcrossing populations results in higher diversity than expected under long-term equilibrium.

**Introduction**

The rate of self-fertilization in hermaphrodite organisms is expected to affect a number of important features of population genetic structure and diversity. Most directly, homozygosity increases as a function of the selfing rate and thus reduces the effective population size ($N_e$), up to twofold with complete selfing (Pollak 1987; Charlesworth et al. 1993a; Nordborg 2000). Further, because of increased homozygosity crossing-over rarely occurs between heterozygous sites thus increasing linkage disequilibrium (LD). Higher LD causes stronger hitchhiking effects such as selective sweeps, background selection and Hill-Robertson interference, all of which are expected to further reduce the amount of neutral genetic variation within populations (reviewed in Charlesworth and Wright 2001).

Population genetic processes resulting from inbreeding may be further augmented by demographic and life-history characteristics associated with the selfing habit. In particular, selfing populations can be founded by single individuals resulting in striking reductions in diversity as a result of genetic bottlenecks and reproductive isolation. The capacity for uniparental reproduction gives many selfers prolific colonizing ability and the capacity to establish after long-distance dispersal, especially in comparison with obligate outcrossers (Baker 1955; Pannell and Barrett 1998). The colonization-extinction dynamics typical of many selfing species and limited pollen-mediated gene flow also increases differentiation among populations resulting in considerable population subdivision (Hamrick and Godt 1990; 1996; Schoen and Brown 1991). Although the total amounts of among-population
variation may be less affected by these processes (Pannell and Charlesworth 1999; Ingvarsson 2002) the demographic and life history characteristics of many selfing species are likely to result in non-equilibrium conditions occurring in selfing populations.

In many taxa where selfing has evolved it may be of relatively recent origin (Schoen et al. 1997; Takebayashi and Morrell 2000; Foxe et al. 2009; Guo et al. 2009). Where selfing has recently established demographic forces associated with colonization may be as important as the mating system per se in structuring patterns of diversity. For example, if selfing originates through the establishment of a small number of founders we would expect a sharp reduction in diversity relative to the outcrossing progenitor and a strong signature of a genetic bottleneck. In contrast, if selfing has evolved recently through the spread of genetic modifiers of small effect, newly established populations may retain significant amounts of ancestral polymorphism from their outcrossing progenitors. In this latter case populations may retain considerably more variation than expected under long-term equilibrium predictions.

Molecular evidence for reduced nucleotide diversity and greater differentiation among populations of selfing taxa compared to populations of related outcrossing taxa has been reported from *Leavenworthia* (Liu et al. 1998; 1999), *Arabidopsis* (Savolainen et al. 2000; Wright et al. 2002); *Solanum* (Baudry et al. 2001), *Mimulus* (Sweigart and Willis 2003), *Amsinckia* (Perusse and Schoen 2004) and *Caenorhabditis* (Graustein 2002; Cutter et al. 2006b; Cutter 2008). In each case the reduction in diversity was more severe than the two-fold reduction predicted for selfing populations at equilibrium. This indicates that factors in addition to the mating system are reducing diversity, but it has been difficult to uncouple the relative importance of genetic hitchhiking from the ecology and demographic history of selfing populations. This challenge parallels similar difficulties in efforts to distinguish
selective from demographic explanations in population genetic studies of *Drosophila* (Haddrill et al. 2005; Ometto et al. 2005; Thornton and Andolfatto 2006; Jensen et al. 2008). However, in many plant populations, especially those with annual life histories and small structured populations, demographic processes may play a more prominent role in causing reduced diversity than increased hitchhiking associated with selfing.

Molecular population genetic studies of selfing in plants have generally focused on either small samples from a large number of populations (e.g. Sweigart and Willis 2003; Nordborg et al. 2005), or relatively large within-population samples from a small number of populations (e.g. Baudry et al. 2001). Ideally, a deeper sampling both within and among populations combined with independent ecological and historical information is required to improve understanding of the interplay of demographic and selective factors. Here we address these issues by examining patterns of nucleotide diversity within a large sample of populations of *Eichhornia paniculata* (Pontederiaceae), an annual species for which there is considerable ecological and demographic information (reviewed in Barrett and Husband 1997).

*Eichhornia paniculata* occurs primarily in N.E. Brazil and the Caribbean islands of Cuba and Jamaica. Various lines of evidence suggest that Brazil is the original source region for Caribbean populations (reviewed in Chapter 3). Populations of *E. paniculata* exhibit striking mating-system diversity, ranging from predominantly outcrossing to those that are highly selfing (outcrossing rate, $t = 0.002 - 0.96$; $n = 54$ populations; Barrett and Husband 1990; Barrett et al. 1992). Variation in mating system is associated with the evolutionary breakdown of the species’ tristyloous genetic polymorphism and the spread and fixation of selfing variants capable of autonomous self-pollination (Barrett et al. 1989). Populations of *E. paniculata* are characterized by three morph structures – trimorphic with long-, mid- and
short-styled morphs (hereafter L-, M-, S-morphs); dimorphic – with two floral morphs, most commonly the L- and M-morphs; and monomorphic – primarily composed of selfing variants of the M-morph. The morph structure and presence of selfing variants within populations explains ~60% of the variation in outcrossing rates among populations (Barrett and Husband 1990). Trimorphic populations are largely outcrossing, dimorphic populations display mixed mating and monomorphic populations are highly selfing. Patterns of allozyme variation indicate a reduction in diversity with increased selfing rates and greater among-population differentiation (Glover and Barrett 1987; Barrett and Husband 1990; Husband and Barrett 1993; 1995). Finally, studies of the inheritance of mating-system modifiers (Fenster and Barrett 1994; Vallejo-Marín and Barrett 2009) in combination with allozyme (Husband and Barrett 1993) and molecular evidence (Chapter 3) indicate that the transition from outcrossing to selfing in *E. paniculata* has occurred on multiple occasions.

The goal of our study was to investigate the relation between mating-system variation and neutral molecular diversity for a large sample of *E. paniculata* populations encompassing most of the geographical range. This was accomplished by collecting multi-locus nucleotide sequence data from 225 individuals sampled from 25 populations including trimorphic, dimorphic and monomorphic populations. Because it has been previously demonstrated that this sequence of morph structures is strongly associated with increasing rates of self-fertilization (see Barrett and Husband 1990), we predicted a decrease in neutral diversity and increases in $F_{st}$ and linkage disequilibrium from floral trimorphism to monomorphism. This extensive population-level sampling across a wide range of selfing rates allowed us to investigate the relative importance of mating system, geography and current population size in structuring genetic variation. We also applied the approaches of Bayesian clustering (Pritchard et al. 2000; Falush et al. 2003; Gao et al. 2007) and divergence population genetics
(Hey and Nielson 2004; Wakeley and Hey 1997; Becquet and Przeworski 2007) to investigate the demographic history of *E. paniculata* and to provide a framework for understanding island colonization and the transition from outcrossing to selfing.

### Methods

**Population sampling**

We sampled open-pollinated maternal seed families from *E. paniculata* populations occurring in N.E. Brazil (May 2005), Jamaica (January 2006) and Cuba (March 2008). Because of a significant break in the distribution of *E. paniculata* in N.E. Brazil, corresponding to an arid zone inimical for aquatic plant growth, we distinguish a northern (Ceará) and southern (Pernambuco, Alagoas) portion of the Brazilian range. This is evident in Figure 4.1, which maps the location of all populations used in this study. Information on morph structure, morph diversity, the frequency of selfing variants within populations, and current population size are provided for each population in Table 4.1. Details of the methods we used to sample population size and morph diversity can be found in Barrett et al. (1989). We germinated seeds and grew plants for DNA extraction in a glasshouse at the University of Toronto. We used a single individual from each maternal family to minimize the number of closely related individuals and to provide the best estimate of population-level diversity. Sample sizes for the number of individuals sequenced in each population are provided in Table 4.1 (mean number of individuals per population = 9, range 4-12).

**Marker development**

We developed nuclear DNA primers based on EST sequences collected from a cDNA library from leaf tissue of a single plant. We purified poly-adenylated RNA from total RNA.
using the Ambion Micro Poly(A) Purist kit and reverse transcribed the mRNA using the InVitrogen Superscript cDNA synthesis kit. We cloned cDNA into *E. coli* and sequenced approximately 480 clones.

From these clones, we selected sequences that aligned to well-annotated nuclear sequences from other plants and designed primers to amplify both coding regions and intron sequence where possible. We initially tested the primers by sequencing the loci in four inbred lines derived from selfing populations. Our rationale being that there should be few heterozygous sites and any loci with these could be excluded as likely paralogs. From this screening we chose 10 EST-derived nuclear markers (Table 4.2) that were then amplified and sequenced in all individuals. Open reading frames in coding loci were identified and annotated using a combination of BLASTx, the original EST sequence and the ORF prediction software GeneMark-E* (Lomsadze et al. 2005).

*Amplification and sequencing*

We extracted DNA from all 225 individuals and this was used to PCR amplify the 10 loci in each individual. We sequenced both forward and reverse strands with an ABI 3730XL fluorescent-based capillary sequencer at the Centre for Applied Genomics facility at Sick Kids Hospital, Toronto, Ontario Canada. We assembled and aligned sequences using Sequencher 4.7 and edited chromatographs and alignments manually to ensure that all base calls and polymorphisms including heterozygotes were reliably scored.

*Polymorphism*

Of the analyses described below, both InStruct and analyses of linkage disequilibrium (LD) include all sites while the remaining analyses used only silent sites (synonymous and
noncoding). We calculated both silent and nonsynonymous polymorphism statistics including $\theta_w$ (Watterson 1975) and $\theta_n$ (Tajima 1983) for all loci and populations using the program SITES (Hey and Wakeley 1997). We calculated Tajima’s $D$ (Tajima 1989) at silent sites with the program SITES and compared each locus in each population to simulation results to detect significant deviations from neutrality using the program HKA (Hey and Wakeley 1997). In addition, we calculated multi-locus estimates of nucleotide polymorphism ($\theta_w$) using a maximum-likelihood method based on the number of segregating sites as implemented by Wright et al. (2003b) to test for significant differences in polymorphism between Caribbean and Brazilian regions. Significant differences in $\theta_w$ were assessed using twice the relative ln likelihood following the $\chi^2$ approximation. We also calculated the same maximum-likelihood estimate of $\theta_w$ for each population. However, for analyses of polymorphism at the population level we discuss only silent $\theta_n$ (Table 4.3 for $\theta_w$) because the patterns and significance attained are equivalent using both measures of genetic variation.

**Correlates of nucleotide diversity**

To understand the factors influencing genetic variation we investigated the relation between nucleotide diversity ($\theta_n$) and five population-level variables: population size, morph diversity, frequency of selfing variant, morph structure (tri-, di-, or monomorphic) and region (Caribbean or Brazil). Values for the variables are presented in Table 4.1. Morph diversity and the frequency of the selfing variant were highly correlated (Kendall’s $\tau = 0.900$). We therefore report results for analyses with the frequency of the selfing variant only, as results were similar using morph diversity. We tested for associations among population size, frequency of the selfing variant, morph diversity and $\theta_n$ using Kendall’s rank correlations. We also calculated Kendall’s rank partial correlations (Kendall 1942) for each pairing of
these variables to examine their associations in the absence of correlated effects from the third variable. Because region and morph structure were nominal variables we compared mean nucleotide diversity within populations to these variables using a Kruskal-Wallis test and compared means between pairs of populations with the Tukey-Kramer HSD test.

Population genetic structure

To investigate the genetic structure of populations we used the Bayesian clustering program InStruct (Gao et al. 2007). The program is similar to the algorithm STRUCTURE developed by Pritchard et al. (2000), but allows for varying levels of inbreeding within subpopulations. We removed singletons from all alignments and sequences were converted to haplotypes and the phase of heterozygous sites was inferred using the program PHASE v2.1 (Stephens et al. 2001; Stephens and Scheet 2005). Simulations were conducted using InStruct version two to estimate structure with both admixture and selfing. We assumed all 10 loci are unlinked which is supported by analyses of linkage disequilibrium among loci. We simulated $10^6$ iterations with a burnin length of $10^5$ steps. This was repeated over cluster number ($K$) from one to the number of sampled populations ($K = 1$ to $K = 25$) and each simulation was run three times with the optimal number of clusters determined by the software using the deviance information criteria.

We calculated mean $F_{ST}$ (Wright 1931) for silent sites within regions using the program SITES (Hey and Wakeley 1997) and averaged pairwise $F_{ST}$ estimates for each population across all populations within its own region (Table 4.1). We also conducted an analysis of molecular variance (AMOVA) to investigate the partitioning of variation among and within populations and regions using the program GenAlex 6 (Peakall and Smouse 2006).
Linkage disequilibrium

We calculated all within-population and within-region, pairwise associations ($R^2$) between single nucleotide polymorphisms (SNPs). For regional samples we present results for a ‘scattered sample’ that includes one individual per population. This sampling strategy has been shown to most accurately reflect the deeper coalescent history of structured populations (Wakeley and Lessard 2003; Städler et al. 2009), which is likely in *E. paniculata* (see Husband and Barrett 1998), with minimal effects of population subdivision in generating excess LD. We estimated the expected decay in LD with distance within loci using the method of Cutter et al. (2006a) and equation 3 from Weir and Hill (1986):

$$E(r^2) = \frac{10 + \Gamma}{22 + 13\Gamma + \Gamma^2} \times \left[ 1 + \frac{(3 + \Gamma)(12 + 12\Gamma + \Gamma^2)}{n(22 + 13\Gamma + \Gamma^2)^2} \right]$$

Where $\Gamma$ is the product of the recombination rate ($\rho = 4N_e\mu$) and distance in base pairs. We fitted this equation to our data using values of associations between pairs of SNPs ($R^2$) estimated with Weir’s (1996) algorithm for unphased data as implemented by MacDonald et al. (2005). Values for $R^2$ were calculated only for pairs of SNPs within loci; however, we pooled values across all loci to estimate the decay in LD with distance. We complimented this method by using Hudson’s (2001) estimator of the population recombination parameter $\rho$, which employs a composite likelihood approach based on pairwise disequilibrium between pairs of SNPs within loci. We excluded loci with fewer than five SNPs in this analysis because this method does not accurately estimate $\rho$ when there are few variable sites.
Demographic history

We conducted coalescent simulations of the divergence between Caribbean and Brazilian populations of *E. paniculata* using the isolation-migration software MIMAR (Becquet and Przeworski 2007). We chose MIMAR over other available methods because it allows for intralocus recombination, which we observed in the Brazilian and Caribbean regions. In MIMAR we simulated a single ancestral population that instantaneously split into two derived populations (Brazil and Caribbean) at some time \( T \) in the past. All populations were assumed to be at a constant size and to be linked by gene flow. Because we did not have a close outgroup sequence to reliably reconstruct ancestral states for SNPs, we used a modified version of MIMAR as implemented by Foxe et al. (2009), which does not assume knowledge of the ancestral state of polymorphic sites. We included silent-site variation from all 10 loci in our input, excluding positions with 3 or more variants. We allowed for locus-specific mutation rates by dividing \( \theta_{SIL} \) from each locus by the mean \( \theta_{SIL} \).

We initially used wide prior limits for simulations to determine better priors for later runs. All \( \theta \) priors had a uniform distribution and individual populations had prior probabilities of \( \theta_{CRB}: 0.0005 - 0.002, \theta_{BRA}: 0.0018 - 0.006 \) and \( \theta_{ANC}: 0.0075 - 0.018 \) for the Caribbean, Brazilian and ancestral populations, respectively. We also conducted simulations where \( \theta_{BRA} = \theta_{ANC} \); however this model did not influence the general results of the analysis and is not included. Simulation runs that included migration indicated a mode at zero and little effect on other parameters, and we therefore report only runs assuming no gene flow. The time of the split between contemporary Brazilian populations and the Caribbean was bounded between 10,000 and 175,000 years ago with a uniform distribution. The mutation rate for *E. paniculata*, a monocot, is unknown; therefore, to estimate time in generations and the effective population size, we used the mean synonymous-site mutation rate from the grass
ADH sequence which is $\mu = 6.5 \times 10^{-9}$ substitutions per site per generation (Gaut et al. 1996). We ran simulations for 20,160 minutes (2 weeks). This allowed for approximately $1.61 \times 10^6$ iterations with a burn-in of 100,000 steps. Marginal posterior probabilities for all parameters were generated in R v2.8.1 (R Development Core Team 2008). To estimate the portion of the posterior distribution that encompasses 90% of the simulated values, and describe confidence limits around point estimates of each parameter, we calculated the 90% highest probability density (90% HPD) for each parameter using the boa package for R (Smith 2007).

Results

Polymorphism

Species-wide estimates of silent nucleotide variation in *E. paniculata* were $\theta_w = 0.0101$ and $\theta_\pi = 0.0064$. The ratio of replacement to silent segregating sites in all of the coding regions is consistent with purifying selection constraining changes in coding sites (see Table 4.2). Neutral genetic diversity was distributed unevenly within and among populations and the mean within-population diversity ($\theta_w = 0.0034$, $\theta_\pi = 0.0020$) was substantially lower than species-wide estimates. An AMOVA of all populations revealed that 64.3% of the molecular variation was partitioned within populations with the remaining 35.7% distributed among populations.

Correlates of nucleotide diversity

There was striking variation in silent nucleotide diversity among regions and populations within each floral morph structure (Figures 4.2, 4.3). However, as predicted, monomorphic populations maintained significantly less variation than trimorphic or dimorphic populations ($\theta_\pi$, monomorphic = 0.00083, $\theta_\pi$, trimorphic = 0.001989, $\theta_\pi$, dimorphic =
0.001719, Kruskal-Wallis test, $\chi^2 = 11.45, P < 0.01$). To examine whether this difference in $\theta_\pi$ was significantly distinguishable from the two-fold reduction in diversity predicted in selfing populations, we doubled the mean of $\theta_\pi$ in monomorphic populations and repeated the test. There was no longer a significant effect of morph structure on diversity, a result consistent with theoretical predictions of a 50% reduction due to selfing (Kruskal-Wallis test, $\chi^2 = 0.837, P = 0.658$). There was no significant difference in the amount of nucleotide diversity within trimorphic versus dimorphic populations (Tukey-Kramer, $P = 0.78$). Based on measures of $\theta_\pi$, a Kruskal-Wallis test revealed that region of origin had a significant effect on within-population diversity ($\theta_\pi_{\text{Caribbean}} = 0.0009, \theta_\pi_{\text{Brazil}} = 0.00172, \chi^2 > 4.795, P < 0.05$). Additionally, using the joint maximum likelihood estimate of $\theta_W$, the Caribbean sample treated as a single population was significantly less diverse than the total Brazilian sample (Figure 4.3, $\theta_W_{\text{Brazil}} = 0.0068, \theta_W_{\text{Caribbean}} = 0.0023, \chi^2 \cong 20.55, P < 0.0001$).

Nucleotide diversity was strongly associated with estimates of population census size and the frequency of the selfing variant within populations (Table 4.4). Population size was highly correlated with $\theta_\pi$ (Figure 4.4; Kendall’s $\tau = 0.512, P < 0.001$) but retained a lower partial correlation (0.105) after controlling for variation in the frequency of the selfing variant. However, while the frequency of the selfing variant correlated strongly with $\theta_\pi$ (Kendall’s $\tau = -0.449, P < 0.005$), it retained very little partial correlation (0.032) after we controlled for the influence of population size. However, neither of these partial correlations was significant because population size and the frequency of the selfing variant were so highly correlated (Kendall’s $\tau = 0.629, P < 0.0001$).

Multi-locus estimates of Tajima’s $D$ revealed that 40% of the populations sampled exhibited significantly negative average multi-locus values (Table 4.3). This pattern varied among regions. For example, within the southern portion of the Brazilian range, six of nine
populations had significantly negative values (mean within population $D = -0.57$, SE 0.075), whereas one of five populations from the northern portion of the Brazilian range showed a significant negative Tajima’s $D$ and another population showed a significant positive value. When considered regionally, southern Brazil showed a highly significant negative Tajima’s $D$ ($D = -1.159$), while populations sampled from northern Brazil did not differ from neutrality ($D = -0.134$). Only two of five Jamaican populations and no Cuban populations had significantly negative values of Tajima’s $D$. However, when Cuban and Jamaican populations were pooled, they showed a significant excess of rare variants ($D = -0.59$). It should be noted that the coalescent simulations as implemented in the HKA program assume no recombination, and thus this test should be conservative with respect to inferring departures from neutral equilibrium.

**Population structure**

Using InStruct, we explored genetic structure of *E. paniculata* across its geographical range. At $K = 2$, the clusters reflected the large geographic divide between populations sampled from Brazil and the Caribbean (Figure 4.5). At $K = 3$, the clusters corresponded to geographical sub-division within Brazil, plus individuals from the Caribbean populations. As the number of clusters increased, each larger region became subdivided into groups that did not correspond to the geographic locations of populations. The optimal number of clusters, as determined by the deviance information criterion, was $K = 9$. At this $K$-value there was still little change to the clusters from the three major regions as defined by $K=3$. Individuals from Caribbean and Brazilian populations from the northern portion of the range were largely composed of single clusters, while individuals from the southern portion of the Brazilian range were mostly composed of an admixture of two different clusters. To further explore
genetic structure, we also conducted separate runs of InStruct for each region separately. For Brazilian populations these analyses provide no additional insight; however, within the Caribbean individuals from a single population from Cuba (C5) clustered with Jamaican populations (Figure 4.5).

Our analyses of population differentiation using $F_{st}$ corresponded with patterns obtained from InStruct. Pairwise $F_{st}$ values, in which members of a pair are populations from different regions, were substantially higher than when both populations were sampled from a single region (data not shown). When pairs are restricted to comparisons within regions, mean pairwise $F_{st}$ values were highest among Caribbean populations and lowest among trimorphic populations from northern Brazil (Figure 4.6, Tukey-Kramer, $q^* = 2.512, P < 0.05$). Populations representing all three morph structures are represented in the southern portion of the Brazilian range. This region had intermediate $F_{st}$ values compared with the exclusively trimorphic populations from the northern portion of the Brazilian range and Caribbean populations. Population morph structure had a significant effect on $F_{st}$ values within regions defined by InStruct at $K = 3$ (Kruskal-Wallis test, $\chi^2 = 8.4295, P < 0.05$). Monomorphic populations exhibited significantly higher levels of differentiation than either dimorphic or trimorphic populations, which did not differ significantly from one another (Tukey-Kramer, $q^* = 2.512, P < 0.05$).

**Linkage disequilibrium**

Using the method of Cutter et al. (2006a) with a scattered sample we found evidence for reduced effective recombination ($\rho = 4Ner$) within loci from Caribbean versus Brazilian populations (Figure 4.7, $\rho_{BRAZIL} = 0.0100, \rho_{CARIBBEAN} = 0.0014$). Moreover, when $\rho$ was estimated using Hudson’s coalescent-based approach the breakdown of LD averaged across
all loci was also moderately higher for Brazil than for the Caribbean ($\rho_{BRAZIL} = 0.0160, \rho_{CARIBBEAN} = 0.0121$). However, there was no evidence for recombination within any locus in populations of *E. paniculata* using either method. Estimates of $\rho = 4N_e r$ were zero for all sampled populations. In many cases there were not enough SNPs at a given locus to use Hudson’s $\rho$ estimator. Mean LD among loci within populations was much lower than intralocus LD ($R^2_{total} = 0.140$) and slightly lower in Brazil than the Caribbean ($R^2_{Brazil} = 0.062, R^2_{Caribbean} = 0.115$).

**Coalescent simulations**

Our simulation results indicate that Caribbean populations exhibit lower effective population size than Brazilian populations (Figure 4.8A). The mode of the marginal posterior probability for $\theta_{CRB} = 0.0012$ (90% HPD = 0.0008-0.0018) was lower than for Brazil, $\theta_{BRA} = 0.0042$ (90% HPD = 0.0019-0.0049). These model-based estimates were both substantially lower than the estimated diversity of the ancestral population $\theta_{ANC} = 0.0125$ (90% HPD = 0.0082-0.0154). This suggests that both regions have experienced a population bottleneck, with the Caribbean experiencing a more extreme reduction in effective population size. Using estimates of the mutation rate from grasses, these values provide estimates of effective population sizes for the three populations of $N_{CRB} = 46,200$, $N_{BRA} = 161,500$ and $N_{ANC} = 481,000$. The model-based estimate of diversity in the Caribbean was marginally lower than the estimate made directly from our sequence data ($\theta_{\pi, Caribbean} = 0.0019$, SE = 0.0007) suggesting that the retention of ancestral polymorphism is inflating estimates of diversity in the Caribbean. Estimates of diversity from simulations and empirical estimates were similar for Brazilian populations ($\theta_{\pi, Brazil} = 0.0039$). Using the mode as a point estimate of the time
since divergence between Caribbean and Brazilian populations gave a colonization time for the Caribbean of 125,000 generations BP (90% HPD = 66,449-146,232; Figure 4.8B).

**Discussion**

Highly selfing populations should have an effective size 50% that of equivalent outcrossing populations under neutral and equilibrium expectations (Charlesworth et al. 1993a; Nordborg 2000). However, most inter-specific comparisons of nucleotide diversity in outcrossing versus selfing species indicate that selfing populations maintain considerably lower values than this prediction. This suggests that genetic hitchhiking and demographic differences between species are reducing diversity beyond the standard neutral expectations.

In contrast, a major finding of our study was that the average neutral diversity of highly selfing monomorphic populations of *E. paniculata* showed a roughly two-fold reduction relative to polymorphic populations. Hence, it is unnecessary to invoke genetic hitchhiking to explain differences in diversity among populations. Significantly, the census population size was a strong predictor of the amount of nucleotide diversity in populations, indicating an important role of demography in shaping patterns of diversity. Levels of population differentiation showed the expected increase for regions with higher selfing rates, and there was some evidence for higher levels of intra-locus linkage disequilibrium in the Caribbean, where selfing rates are the highest. We now consider how colonization processes and the demographic origins of populations are likely to play a role in structuring regional patterns of nucleotide polymorphism.
Population size, morph structure and nucleotide polymorphism

Our results demonstrate that in *E. paniculata* population size and its association with morph structure is a stronger predictor of silent nucleotide diversity than mating patterns alone. Monomorphic, highly selfing populations were considerably smaller in size, consistent with higher population turnover and frequent founder events. Partial correlations indicated that morph diversity had little residual effect on within-population diversity levels once we controlled for census size. This suggests that the effect of mating system on variation within populations is primarily via its interaction with differences in colonization and/or persistence of selfing versus outcrossing populations. Although the effects of population size on diversity is largely driven by the lowest population size classes, the majority of monomorphic populations fall within this range, highlighting the important interaction between selfing and population size on genetic diversity. Among the populations we sampled there are orders of magnitude differences in census population size (Table 4.1), while neutral models predict only twofold effects of the mating system. Therefore, it is perhaps not surprising that there is strong signature of the influence of census size on nucleotide diversity in our study.

Diversity estimates among populations within each of the morph structures varied considerably, and this heterogeneity was not fully explainable by morph structure or population size. For example, although on average trimorphic populations maintained significantly more nucleotide diversity than monomorphic populations, two large trimorphic populations (B187, B206) contained less variation than occurred in several of the most diverse monomorphic populations from Jamaica (J29) and Cuba (C1, C5). The heterogeneity in estimates of neutral diversity within each of the morph structure classes probably results from multiple causes. First, it is important to emphasize that direct marker-based mating-system estimates were not obtained from the populations used in our sample. Although our
previous work indicates that style-morph structure provides a reasonable predictor of mating patterns in *E. paniculata*, variation in outcrossing rate among populations occurs within each of the morph structure classes (see for example Figure 9.1 in Barrett et al. 1992). Second, as discussed above, aspects of demographic history associated with the species’ annual life history also likely plays a role in affecting diversity. Large-scale censuses of population size have demonstrated dramatic fluctuations from year-to-year and a turnover of populations consistent with frequent colonizing episodes and metapopulation dynamics (Husband and Barrett 1998). These local ecological processes are likely to have an important influence on levels of genetic diversity. Finally, nucleotide variation within populations of a particular morph structure also will be affected by regional patterns of diversity and colonization history. The fragmentation of the Brazilian range and colonization of the Caribbean from South America have both likely affected the pool of diversity available in any given region (and see below).

We detected no significant difference in nucleotide diversity between trimorphic and dimorphic populations of *E. paniculata*. Several features of dimorphic populations led us to predict reduced diversity compared to trimorphic populations. The vast majority of dimorphic populations of *E. paniculata* are missing the S-morph, including the seven populations sampled in this study. Theoretical and empirical evidence indicates that genetic drift and founder events play a prominent role in the origins of dimorphic populations (Barrett et al. 1989; Husband and Barrett 1992a;b). These stochastic processes could potentially erode diversity and, indeed, comparisons of allozyme variation between trimorphic and dimorphic populations support this prediction (Husband and Barrett 1993). In addition, dimorphic populations possess mixed mating because of the occurrence of self-pollinating variants of the M-morph. Estimated selfing rates from dimorphic populations are
variable but in some cases can be considerable. These considerations led us to predict that
dimorphic populations should have lower diversity than the more outcrossing trimorphic
populations.

Why did we not observe the severe loss of diversity within our selfing and partially
selfing populations seen in interspecific comparisons? If there was insufficient time to
recover equilibrium since the evolution of dimorphism and monomorphism, many of the
genetic consequences of selfing (such as reduced $N_e$, increased LD or reduced efficacy of
selection) would be less apparent. For example, selfing populations that evolved less than $4N$
generations ago can still retain ancestral variation and ancient, short-range linkage
disequilibrium may be unexpectedly low in such selfing populations (Tang et al. 2007). Additionally, there may have been insufficient time for positive and negative selection to
erode diversity through genetic hitchhiking. Thus the origin of stylar dimorphism and
monomorphism less than $4N$ generations ago from diverse ancestral source populations
probably explains the retention of significant amounts of ancestral polymorphism. Our
coalescent simulation results are consistent with Caribbean populations having diverged from
Brazilian populations less than $4N$ generations ago (see below).

Dimorphic and monomorphic populations in N.E. Brazil maintained large to
moderate amounts of the polymorphism found in trimorphic populations. In the case of the
two monomorphic populations in our Brazilian sample this amounted to 35.7% of the
diversity found in the southern portion of the range. The maintenance of this diversity
suggests that the spread of selfing in Brazilian dimorphic and monomorphic populations may
have been sufficiently gradual to allow recombination and segregation to capture a
significant portion of the neutral diversity of progenitor populations than would be expected
if selfing evolved by a founder event or through a very rapid selective sweep.
Timescales and demographic history

Previous studies of allozyme variation in *E. paniculata* suggested that Jamaican populations likely arose from at least two long-distance dispersal events (Husband and Barrett 1991). More recent molecular analyses of the Cuban populations investigated here (Chapter 3), in concert with the InStruct analysis presented here (Figure 4.5), indicate that they share a significant proportion of their nucleotide variation with Jamaica and may have descended from the same colonization event(s) from mainland South America, most likely Brazil. Interpreting these patterns of molecular variation can be assisted by the application of models of demographic history (e.g. Haddrill et al. 2005; Ometto et al. 2005; Wright et al. 2005; Ross-Ibarra et al. 2008; Nielsen et al. 2009). Our own coalescent simulations indicated that the time since colonization of the Caribbean by *E. paniculata* likely occurred approximately ~125,000 years before present. This estimate is of significance because *E. paniculata* is a weed of rice fields in Cuba and Jamaica raising the possibility that migration to the Caribbean from South America occurred in historic times through human introduction, as appears to be the case for the related *E. crassipes* (Water Hyacinth) also native to Brazil (Barrett and Forno 1982). However, our estimate of the date of Caribbean colonization casts serious doubt on this hypothesis. Instead, this analysis suggests that establishment in the Caribbean occurred much earlier, probably the result of natural long-distance dispersal events, presumably by birds. The predominance of selfing variants of the M-morph in Cuba and Jamaica, as well as the absence of the S-morph in this region is generally consistent with the hypothesis that the facility for autonomous self-pollination in selfing variants enabled establishment following long-distance dispersal.
The estimate of divergence of Brazilian and Caribbean populations should be treated cautiously, because it is based on using the mean mutation rate from grasses \(6.5 \times 10^{-9}\) mutations per site per generation and mutation rates in plants can vary by orders of magnitude (Gaut et al. 1996; Koch et al. 2000). Unfortunately, mutation rates are difficult to obtain and no estimates are available for species more closely related to *E. paniculata*. With these caveats in mind, we calculate that approximately \(2.6N\) generations have elapsed since the colonization of the Caribbean islands, based on the observed level of diversity and the time since divergence estimated from coalescent simulations. Standard neutral theory predicts that on average \(4N\) generations is required for all individuals in a sample to coalesce and recover equilibrium (Kimura and Ohta 1969). This highlights the potential significance of the retention of ancestral diversity and recombination events in maintaining variation within and among monomorphic Caribbean populations.

Our MIMAR results indicate a three- to fourfold lower \(N_e\) for populations from the Caribbean islands compared with Brazil, and a tenfold reduction in effective size relative to the ancestral population. This result suggests that, although the effective size of Caribbean populations is lower than those from Brazil, they do not appear to have been derived from a single colonizer, a result consistent with Husband and Barrett’s (1991) allozyme study. If this were the case, a much stronger signal of a genetic bottleneck would be evident in our data, as was observed, for example, associated with the evolution of selfing in *Capsella* (Foxe et al. 2009; Guo et al. 2009). This finding is more consistent with multiple colonization events or multiple colonizers allowing for the maintenance of a larger fraction of the diversity from founding populations.

Caribbean populations of *E. paniculata* as a whole may have not yet recovered equilibrium, but may be close to approaching this point. Evidence for the recovery of neutral
diversity in Caribbean populations is that a substantial fraction of polymorphism (22.4%) is unique to populations from this region. Three Jamaican populations and the Caribbean as a whole show a significantly negative Tajima’s $D$, and the remaining populations generally show a trend towards a negative Tajima’s $D$. This suggests ongoing population expansion after colonization, perhaps associated with the species weedy habit.

MIMAR simulations also implicate a threefold reduction in effective size in the Brazilian populations relative to the ancestor. It is possible that this result is an artifact of population subdivision in the ancestral population (Becquet and Przeworski 2009). However, this could also result from range fragmentation in Brazil with an historical bottleneck contributing to an erosion of diversity in outcrossing Brazilian populations. It is noteworthy that populations from Brazil, particularly in the southern portion of the range, show a significant negative Tajima’s $D$, consistent with population expansion during recovery from a bottleneck.

**Comparing diversity in outcrossing and selfing populations**

The retention of a moderate fraction of ancestral polymorphism in selfing populations of *E. paniculata* contrasts with several recent interspecific studies comparing molecular diversity in related selfing and outcrossing species. For example, in *Solanum* self-compatible species have between 4-40 times less variation than the least variable self-incompatible species (Baudry et al. 2001), selfing *Mimulus nasutus* has approximately seven-fold less variation compared with its outcrossing relative *M. guttatus* (Sweigart and Willis 2003), and *Capsella rubella* has 100 to 1500-fold reduction in effective population size compared with *C. grandiflora* (Foxe et al. 2009). In this latter case it has been proposed that a speciation event was associated with the transition from outcrossing in *C. grandiflora* to selfing in *C.*
rubella. This may have resulted from a rapid breakdown in self-incompatibility (Foxe et al. 2009; Guo et al. 2009). The transition to selfing in Arabidopsis thaliana appears to have been more complex with multiple independent losses of self-incompatibility (Boggs et al. 2009), and a ~4-fold reduction in diversity relative to its closest self-incompatible relative A. lyrata (Ramos-Onsins et al. 2004; Nordborg et al. 2005). However, these two Arabidopsis species are believed to have been diverging for ~5 million years (Koch et al. 2000) and demographic bottlenecks in A. lyrata complicate the use of these two species for understanding the influence of mating system on molecular diversity.

Our studies of intraspecific variation in mating system in E. paniculata allow for molecular population genetic comparisons with fewer confounding effects introduced by independent evolutionary histories and ecological and life-history differences between species. Nevertheless, in wide-ranging species capable of long-distance dispersal, such as E. paniculata, evolutionary history and the demographic origins of selfing populations also play an important role in determining patterns of regional diversity and these need to be taken into account when considering the influence of mating patterns on genetic diversity.
Table 4.1. Populations samples and details of *Eichhornia paniculata* used in this study.

<table>
<thead>
<tr>
<th>Pop. Code</th>
<th>Nearest City</th>
<th>Morph structure</th>
<th>Pop. Size</th>
<th>No. of sequences</th>
<th>Freq. of selfer</th>
<th>Morph diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>B177</td>
<td>Anadia, Alagoas Brazil</td>
<td>dimorphic</td>
<td>500</td>
<td>12</td>
<td>0.25</td>
<td>0.74</td>
</tr>
<tr>
<td>B180</td>
<td>Vicosá, Alagoas Brazil</td>
<td>dimorphic</td>
<td>100</td>
<td>6</td>
<td>0.09</td>
<td>0.73</td>
</tr>
<tr>
<td>B182</td>
<td>Brejao, Pernambuco Brazil</td>
<td>monomorphic</td>
<td>500</td>
<td>12</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>B183</td>
<td>Corrente, Pernambuco Brazil</td>
<td>monomorphic</td>
<td>100</td>
<td>12</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>B184</td>
<td>Corrente, Pernambuco Brazil</td>
<td>dimorphic</td>
<td>500</td>
<td>12</td>
<td>0.26</td>
<td>0.53</td>
</tr>
<tr>
<td>B185</td>
<td>Sao Jose da Laje, Alagoas Brazil</td>
<td>dimorphic</td>
<td>750</td>
<td>6</td>
<td>0.13</td>
<td>0.73</td>
</tr>
<tr>
<td>B186</td>
<td>Garanhuns, Pernambuco Brazil</td>
<td>dimorphic</td>
<td>750</td>
<td>12</td>
<td>0.22</td>
<td>0.74</td>
</tr>
<tr>
<td>B187</td>
<td>Garanhuns, Pernambuco Brazil</td>
<td>trimorphic</td>
<td>1750</td>
<td>8</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>B192</td>
<td>Cupira, Pernambuco Brazil</td>
<td>trimorphic</td>
<td>8000</td>
<td>12</td>
<td>0.11</td>
<td>0.81</td>
</tr>
<tr>
<td>B202</td>
<td>Quixada, Ceara Brazil</td>
<td>trimorphic</td>
<td>1500</td>
<td>6</td>
<td>0.00</td>
<td>0.90</td>
</tr>
<tr>
<td>B206</td>
<td>Pirajana, Ceara Brazil</td>
<td>trimorphic</td>
<td>1500</td>
<td>6</td>
<td>0.00</td>
<td>0.89</td>
</tr>
<tr>
<td>B207</td>
<td>Choro, Ceara Brazil</td>
<td>trimorphic</td>
<td>2500</td>
<td>12</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>B210</td>
<td>Caninde, Ceara Brazil</td>
<td>trimorphic</td>
<td>3000</td>
<td>12</td>
<td>0.00</td>
<td>0.95</td>
</tr>
<tr>
<td>B211</td>
<td>Forteleza, Ceara Brazil</td>
<td>trimorphic</td>
<td>700</td>
<td>12</td>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>C1</td>
<td>Yara, Granma Cuba</td>
<td>monomorphic</td>
<td>500</td>
<td>5</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C2</td>
<td>Manzanillo, Granma Cuba</td>
<td>dimorphic</td>
<td>800</td>
<td>11</td>
<td>0.85</td>
<td>0.38</td>
</tr>
<tr>
<td>C3</td>
<td>Chorrera, Granma Cuba</td>
<td>monomorphic</td>
<td>120</td>
<td>7</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C4</td>
<td>Baracoa, Guantánamo Cuba</td>
<td>monomorphic</td>
<td>10</td>
<td>4</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C5</td>
<td>Camalote, Camagüey Cuba</td>
<td>monomorphic</td>
<td>600</td>
<td>12</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J28</td>
<td>Treasure Beach, St. Elizabeth Jamaica</td>
<td>dimorphic</td>
<td>200</td>
<td>10</td>
<td>0.97</td>
<td>0.08</td>
</tr>
<tr>
<td>J29</td>
<td>Fullerswood, St. Elizabeth Jamaica</td>
<td>monomorphic</td>
<td>28</td>
<td>12</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J30</td>
<td>Cataboo, St. Elizabeth Jamaica</td>
<td>monomorphic</td>
<td>20</td>
<td>4</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J31</td>
<td>Slipe, St. Elizabeth Jamaica</td>
<td>monomorphic</td>
<td>24</td>
<td>4</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J32</td>
<td>Little London, Westmoreland Jamaica</td>
<td>monomorphic</td>
<td>25</td>
<td>10</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J33</td>
<td>Georges Plain, Westmoreland Jamaica</td>
<td>monomorphic</td>
<td>unknown</td>
<td>1.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

---

a Code used throughout this paper to identify collection locations.
b City nearest to the population.
c Dimorphic populations contain selfing variants of the M-morph in varying frequencies; monomorphic populations are composed exclusively of this form.
d Estimate of the census size of populations at time of collection.
e Number of individuals that were sequenced for this study.
f Frequency of the modified selfing variant in populations.
g Morph diversity is a measure of evenness of the three floral morphs normalized to one. Populations with even ratios of all three morphs have a diversity of one, monomorphic populations have a diversity of zero (see Barrett et al. 1989 for further details).
### Table 4.2. Summary statistics of polymorphism for the 10 EST-derived nuclear loci sequenced in Eichhornia paniculata for this study

<table>
<thead>
<tr>
<th>Locus</th>
<th>Length total/coding</th>
<th>$\theta_W$</th>
<th>$\theta_\pi$</th>
<th>Tajima’s $D$</th>
<th>$S_{Rep}$</th>
<th>$S_{Sil}$</th>
<th>$\theta_N : \theta_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP0001</td>
<td>530 / 0</td>
<td>0.0093</td>
<td>0.0043</td>
<td>-1.285</td>
<td>-</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>EP0141</td>
<td>825/ 126</td>
<td>0.0167</td>
<td>0.0122</td>
<td>-0.754</td>
<td>2</td>
<td>60</td>
<td>0.192</td>
</tr>
<tr>
<td>EP0143</td>
<td>846 / 279</td>
<td>0.0040</td>
<td>0.0016</td>
<td>-1.451</td>
<td>7</td>
<td>18</td>
<td>0.538</td>
</tr>
<tr>
<td>EP0144</td>
<td>822 / 0</td>
<td>0.0150</td>
<td>0.0047</td>
<td>-1.821</td>
<td>-</td>
<td>33</td>
<td>-</td>
</tr>
<tr>
<td>EP0188</td>
<td>383 / 262</td>
<td>0.0019</td>
<td>0.0018</td>
<td>-0.033</td>
<td>2</td>
<td>4</td>
<td>0.332</td>
</tr>
<tr>
<td>EP0222</td>
<td>481 / 444</td>
<td>0.0021</td>
<td>0.0007</td>
<td>-1.200</td>
<td>9</td>
<td>6</td>
<td>0.445</td>
</tr>
<tr>
<td>EP0267</td>
<td>561 / 0</td>
<td>0.0068</td>
<td>0.0048</td>
<td>-0.696</td>
<td>-</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>EP0285</td>
<td>827 / 580</td>
<td>0.0101</td>
<td>0.0097</td>
<td>-0.094</td>
<td>10</td>
<td>41</td>
<td>0.610</td>
</tr>
<tr>
<td>EP0314</td>
<td>752 / 457</td>
<td>0.0102</td>
<td>0.0096</td>
<td>-0.155</td>
<td>6</td>
<td>32</td>
<td>0.218</td>
</tr>
<tr>
<td>EP0317</td>
<td>651 / 189</td>
<td>0.0070</td>
<td>0.0025</td>
<td>-1.542</td>
<td>1</td>
<td>18</td>
<td>0.111</td>
</tr>
</tbody>
</table>

1 EST-derived nuclear loci.
2 Total length of the loci and the length of coding sequence in base pairs.
3 Silent nucleotide polymorphism ($\theta_W$).
4 Silent nucleotide diversity ($\theta_\pi$).
5 Tajima’s $D$ test of neutrality.
6 Number of segregating replacement sites from coding sites ($S_{rep}$).
7 Number of silent segregating sites from synonymous and non-coding regions ($S_{sil}$).
8 Ratio of non-synonymous to synonymous polymorphism.
Table 4.3. Tajima’s $D$ and $\theta_W$ values for populations of *Eichhornia paniculata*. Multi-locus estimates of $\theta_W$ were calculated using a maximum-likelihood method based on the number of segregating sites as implemented by Wright et al. (2003b). Multi-locus estimates of Tajima’s $D$ were calculated using the software program HKA (Wakeley and Hey 1997). Significant deviations from neutrality are inferred in populations where the observed Tajima’s $D$ is higher or lower than 95% of simulations. Regions/populations that deviate significantly from neutrality are indicated with an asterisk.

<table>
<thead>
<tr>
<th>Population</th>
<th>Region</th>
<th>$\theta_W$</th>
<th>Obs. mean</th>
<th>Sim. mean</th>
<th>% lower</th>
<th>% higher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil*</td>
<td>-</td>
<td>0.00561</td>
<td>-1.028</td>
<td>-0.594</td>
<td>2.3</td>
<td>97.7</td>
</tr>
<tr>
<td>North</td>
<td>-</td>
<td>0.00247</td>
<td>-0.135</td>
<td>0.142</td>
<td>19.7</td>
<td>80.3</td>
</tr>
<tr>
<td>South*</td>
<td>-</td>
<td>0.00461</td>
<td>-1.159</td>
<td>-0.246</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Caribbean*</td>
<td>-</td>
<td>0.00173</td>
<td>-0.598</td>
<td>0.249</td>
<td>0.4</td>
<td>99.6</td>
</tr>
<tr>
<td>Cuba</td>
<td>-</td>
<td>0.00139</td>
<td>-0.554</td>
<td>-0.046</td>
<td>6.5</td>
<td>93.6</td>
</tr>
<tr>
<td>Jamaica</td>
<td>-</td>
<td>0.00157</td>
<td>-0.403</td>
<td>-0.032</td>
<td>13.9</td>
<td>86.1</td>
</tr>
<tr>
<td>B202*</td>
<td>North</td>
<td>0.00069</td>
<td>0.977</td>
<td>-0.028</td>
<td>98.6</td>
<td>1.4</td>
</tr>
<tr>
<td>B206</td>
<td>North</td>
<td>0.00071</td>
<td>-0.107</td>
<td>-0.031</td>
<td>43.6</td>
<td>56.5</td>
</tr>
<tr>
<td>B207</td>
<td>North</td>
<td>0.00275</td>
<td>0.378</td>
<td>-0.074</td>
<td>89.1</td>
<td>10.9</td>
</tr>
<tr>
<td>B210</td>
<td>North</td>
<td>0.00363</td>
<td>-0.573</td>
<td>-0.140</td>
<td>10.7</td>
<td>89.3</td>
</tr>
<tr>
<td>B211*</td>
<td>North</td>
<td>0.00339</td>
<td>-0.759</td>
<td>-0.011</td>
<td>1.8</td>
<td>98.2</td>
</tr>
<tr>
<td>B177*</td>
<td>South</td>
<td>0.00541</td>
<td>-0.774</td>
<td>-0.049</td>
<td>1.1</td>
<td>98.9</td>
</tr>
<tr>
<td>B180</td>
<td>South</td>
<td>0.00221</td>
<td>-0.339</td>
<td>-0.001</td>
<td>16.4</td>
<td>83.6</td>
</tr>
<tr>
<td>B182</td>
<td>South</td>
<td>0.00205</td>
<td>-0.399</td>
<td>-0.054</td>
<td>15.1</td>
<td>84.9</td>
</tr>
<tr>
<td>B183*</td>
<td>South</td>
<td>0.00171</td>
<td>-0.726</td>
<td>-0.017</td>
<td>2.1</td>
<td>97.9</td>
</tr>
<tr>
<td>B184</td>
<td>South</td>
<td>0.00159</td>
<td>0.515</td>
<td>-0.052</td>
<td>93.8</td>
<td>6.2</td>
</tr>
<tr>
<td>B185*</td>
<td>South</td>
<td>0.00235</td>
<td>-0.883</td>
<td>-0.083</td>
<td>1.3</td>
<td>98.7</td>
</tr>
<tr>
<td>B186*</td>
<td>South</td>
<td>0.00347</td>
<td>-0.785</td>
<td>-0.056</td>
<td>1.3</td>
<td>98.7</td>
</tr>
<tr>
<td>B187*</td>
<td>South</td>
<td>0.00229</td>
<td>-1.010</td>
<td>-0.019</td>
<td>0.3</td>
<td>99.7</td>
</tr>
<tr>
<td>B192*</td>
<td>South</td>
<td>0.00295</td>
<td>-0.729</td>
<td>-0.051</td>
<td>1.5</td>
<td>98.5</td>
</tr>
<tr>
<td>C1</td>
<td>Cuba</td>
<td>0.00221</td>
<td>-0.290</td>
<td>-0.273</td>
<td>48.9</td>
<td>51.1</td>
</tr>
<tr>
<td>C2</td>
<td>Cuba</td>
<td>0.00175</td>
<td>-0.093</td>
<td>-0.082</td>
<td>49.8</td>
<td>50.2</td>
</tr>
<tr>
<td>C3</td>
<td>Cuba</td>
<td>0.00139</td>
<td>0.004</td>
<td>-0.277</td>
<td>75.7</td>
<td>24.4</td>
</tr>
<tr>
<td>C4</td>
<td>Cuba</td>
<td>0.00007</td>
<td>0.334</td>
<td>0.036</td>
<td>52.8</td>
<td>47.2</td>
</tr>
<tr>
<td>C5</td>
<td>Cuba</td>
<td>0.00111</td>
<td>-0.556</td>
<td>0.018</td>
<td>5.7</td>
<td>94.4</td>
</tr>
<tr>
<td>J28</td>
<td>Jamaica</td>
<td>0.00147</td>
<td>-0.674</td>
<td>-0.019</td>
<td>5.9</td>
<td>94.1</td>
</tr>
<tr>
<td>J29*</td>
<td>Jamaica</td>
<td>0.00253</td>
<td>-1.873</td>
<td>-0.041</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>J30*</td>
<td>Jamaica</td>
<td>0.00129</td>
<td>-1.470</td>
<td>0.113</td>
<td>0.1</td>
<td>99.9</td>
</tr>
<tr>
<td>J31</td>
<td>Jamaica</td>
<td>0.00041</td>
<td>0.000</td>
<td>-0.020</td>
<td>50.2</td>
<td>49.8</td>
</tr>
<tr>
<td>J32</td>
<td>Jamaica</td>
<td>0.00017</td>
<td>0.015</td>
<td>-0.045</td>
<td>55.1</td>
<td>44.9</td>
</tr>
<tr>
<td>J33</td>
<td>Jamaica</td>
<td>0.00025</td>
<td>-0.754</td>
<td>-0.020</td>
<td>14.4</td>
<td>85.6</td>
</tr>
</tbody>
</table>
Table 4.4. The statistical relations between population size, the frequency of selfing variants and nucleotide diversity ($\theta_\pi$) in 25 populations of *Eichhornia paniculata* sampled from Brazil and the Caribbean. Above the diagonal are Kendall’s partial rank correlation coefficients and below are Kendall’s $\tau$ estimates. Significant correlations are indicated with asterisks.

<table>
<thead>
<tr>
<th></th>
<th>Population size</th>
<th>Selfing variant frequency</th>
<th>Nucleotide diversity ($\theta_\pi$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>-</td>
<td>-0.520</td>
<td>0.105</td>
</tr>
<tr>
<td>Selfing variant</td>
<td>***0.629</td>
<td>-</td>
<td>0.033</td>
</tr>
<tr>
<td>frequency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide Diversity</td>
<td>**0.512</td>
<td>*0.449</td>
<td>-</td>
</tr>
<tr>
<td>($\theta_\pi$)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $P < 0.005$, ** $P < 0.001$, *** $P < 0.0001$
Figure 4.1. *Eichhornia paniculata* populations sampled for this study. Inset maps show locations of populations from N.E. Brazil, Cuba and Jamaica.
Figure 4.2. Estimates of within-population nucleotide diversity ($\theta_\pi$) for 25 populations of *Eichhornia paniculata* based on the mean diversity of all 10 EST derived nuclear loci.
Figure 4.3. Joint maximum likelihood surface for nucleotide polymorphism ($\theta_W$) for combined Caribbean and Brazilian populations. The reference line at relative ln likelihood = 1.92 represents the credibility interval that determines significantly different estimates based on the $\chi^2$ approximation.
Figure 4.4. Relation between census population size and nucleotide diversity in 25 populations of *E. paniculata* from Brazil and the Caribbean (Kendall’s $\tau = 0.512$, $P < 0.001$). Population size is plotted on a log scale.
Figure 4.5. (A) Range-wide genetic structure of *Eichhornia paniculata* based on analyses conducted using InStruct with different numbers of clusters ($K = 2, 3, 9$). (B) Results for the Caribbean sample of populations from Jamaica and Cuba run separately ($K = 2$). Each thin bar represents a single individual, which maybe partitioned into $K$ colors depending on the estimated multi-locus membership in each of $K$ clusters, where each color represents the posterior probability of that individual belonging to a cluster. The best fitting model $K = 9$ is indicated with an asterisk.
Figure 4.6. Mean $F_{st}$ for pairwise population comparisons of *Eichhornia paniculata* within geographical regions. The values were averaged across populations within each of the four regions and also for trimorphic, dimorphic and monomorphic populations. Letters indicate significant differences using the Tukey-Kramer HSD. Error bars represent ±1 SE.
Figure 4.7. Decay of linkage disequilibrium ($R^2$) with distance (bp) in a scattered sample of *Eichhornia paniculata* from (A) Brazil and (B) Cuba and Jamaica combined. The line on each plot is fit to Weir and Hill’s (1986) equation for the expected decay of LD with distance to infer effective recombination rate, $\rho = 4N_e r$ ($\rho_{\text{BRAZIL}} = 0.0100$, $\rho_{\text{CARIBBEAN}} = 0.0014$).
Figure 4.8. Marginal posterior probability distributions from coalescent simulations of the demographic history of *Eichhornia paniculata*. (A) Simulated estimates of diversity ($\theta = 4N_e\mu$) in populations from Brazil, the Caribbean, and their common ancestor represented by dotted, dashed and solid lines, respectively. (B) Time, (in annual generations), since divergence between Brazilian and Caribbean populations estimated using the average mutation rate of grass *ADH* sequences $\mu = 6.5 \times 10^{-9}$. All curves were estimated using the program MIMAR (Becquet and Przeworski 2007), which implements coalescent simulations in a Bayesian framework. The mode for each distribution is labeled above each curve.
CHAPTER FIVE

DE NOVO SEQUENCE ASSEMBLY AND CHARACTERIZATION OF THE FLORAL TRANSCRIPTOME IN CROSS- AND SELF-FERTILIZING PLANTS

This chapter resulted from a collaboration with Mathieu Siol and Spencer C. H. Barrett. Mathieu Siol contributed to the computational data analysis, ideas and writing of the manuscript. Spencer C. H. Barrett contributed ideas and the writing of the manuscript which is currently under review.

Summary

The shift from cross-fertilization to predominant self-fertilization is among the most common evolutionary transitions in the reproductive biology of flowering plants. Increased inbreeding has important consequences for floral morphology, population genetic structure and genome evolution. The transition to selfing is usually characterized by a marked reduction in flower size and the loss of traits involved in pollinator attraction and the avoidance of self-fertilization. Here, we use short-read sequencing to assemble, de novo, the floral transcriptomes of three genotypes of *Eichhornia paniculata* including an outcrosser, two genotypes from independently derived selfers, and the sister species *E. paradoxa*. By sequencing mRNA from tissues sampled at various stages of flower development, our goal was to sequence and assemble the floral transcriptome and identify patterns of gene expression. Our 24 Mbp assembly resulted in ~27,000 contigs that averaged ~900bp in length. All four genotypes had highly correlated gene expression but the three *E. paniculata* genotypes were more correlated with one another than each was with *E. paradoxa*. Our analysis identified 269 genes associated with floral development, 22 of which were differentially expressed in selfing lineages of *E. paniculata* relative to the outcrosser. Many of the differentially expressed genes affect floral traits commonly altered in selfing plants.
and represent a set of potential candidate genes for investigating the evolution of the selfing syndrome. Our study is among the first to demonstrate the use of Illumina short read sequencing for \textit{de novo} transcriptome assembly in non-model species, and to implement this technology for comparing floral transcriptomes in outcrossing and selfing plants.

\section*{Introduction}

Among the most prevalent evolutionary transitions in plants is the shift from cross-fertilization to predominant self-fertilization among numerous herbaceous angiosperm lineages (Stebbins 1974). This change in mating system has important consequences for many aspects of the biology of selfing taxa including population genetic structure, colonizing ability, genome evolution and the morphology of flowers (Baker 1955; Lloyd 1965; Ornduff 1969; Charlesworth and Wright 2001). The loss of floral mechanisms that reduce the incidence of self-fertilization results in high rates of autogamous selfing, leading to the evolution of the ‘selfing syndrome’ (Morgan and Barrett 1989; Ritland and Ritland 1989; Armbruster et al. 2002). Although studies of the causes and consequences of cross- and self-fertilization in flowering plants have a long and venerable history, beginning with Darwin’s seminal work (Darwin 1876), relatively little is know about the underlying molecular changes that accompany the transition from outcrossing to selfing (Fishman et al. 2002; Tang et al. 2007) and genomic analyses of related outcrossing and selfing plants are in their infancy.

Recent technological advances in DNA sequencing technology have removed several limitations associated with gathering large amounts of genomic data from non-model organisms (Shendure and Ji 2008), providing opportunities for detailed investigation into the
genomics of mating-system variation. Although assembling large eukaryotic genomes, de novo, may not yet be practical sets of expressed genes or transcriptomes present a viable and attractive alternative to population genetic analyses of whole genome sequences. Transcriptomes represent a fraction of the total genome in size, contain fewer repetitive elements, and by selecting specific tissues they can be enriched for genes relevant to the particular aim of the research. In addition, if the RNA sample is not normalized the relative abundance of different reads has been shown to accurately reflect the expression level of transcripts in the tissue (Mortazavi et al. 2008; Nagalakshmi et al. 2008). Despite these potential advantages there remain a number of challenges for de novo transcriptome assembly, including gene duplication or paralogy, heterozygosity and alternative splicing, each of which require careful consideration.

There are relatively few studies to date involving de novo transcriptome assembly in non-model organisms (Table 5.1). So far all have used the Roche 454 GS platform (currently 200-400 bp /read, 2-4 x10^8 bp/run) which has the advantage of longer reads but produces a fraction of the total amount of sequence produced per instrument per run compared with the Illumina GAII platform (currently 38-100 bp/read, 10-20 x10^9 bp/run). Therefore, to maximize the coverage for rare transcripts, cDNA samples are typically normalized (Table 5.1). As a result, these studies were not able to estimate expression levels of different ESTs, as this requires deep sequencing of non-normalized cDNA. Further, due to lower sequencing depth, many transcripts are represented by a single read and others by very few reads. This can create problems in accurately distinguishing SNPs from errors, and in retrieving orthologous transcripts for sequence comparisons across experiments or species. Therefore new methods are required to generate and assemble large datasets, many of which currently consist of substantially shorter reads.
Here, we present de novo floral transcriptome assemblies using short read sequencing of four individual plants that vary in floral morphology and mating system. The samples include three individuals of neotropical *Eichhornia paniculata* (Pontederiaceae), including two from independently derived selfing populations and the third an outcrosser. *Eichhornia paniculata* is an annual diploid that has been the subject of detailed studies over the past two decades on the ecology and genetics of mating-system variation (reviewed in Chapter 3). Populations of *E. paniculata* are largely concentrated in northeastern Brazil, with smaller foci in Jamaica and Cuba and isolated localities in Nicaragua and Mexico. Populations in Brazil are largely outcrossing and possess the sexual polymorphism tristyly, which promotes cross-pollination among the three floral morphs (reviewed in Barrett et al. 1992). Morphological, genetic and biogeographical evidence indicates that tristyly in *E. paniculata* has broken down on multiple occasions resulting in independently derived selfing populations (Husband and Barrett 1993; Fenster and Barrett 1994; Chapter 3). Populations from Jamaica are largely composed of selfing variants of the mid-styled morph (M-morph) in which short level stamens are elongated to a position adjacent to mid-level stigmas resulting in autonomous self-fertilization. In contrast, plants from Mexico and Nicaragua are selfing variants of the long-styled morph (L-morph) with a different arrangement of their sexual organs (see Figure 3.2). Although both variants possess the selfing syndrome, comparisons of molecular variation at 10 EST-derived nuclear loci indicate a high level of differentiation consistent with their separate origins from different outcrossing ancestors (see Figure 3.3). Our analysis included both of these the selfing variants, an individual of an outcrossing L-morph from northeastern Brazil, and a selfing individual of *E. paradoxa*, the sister species of *E. paniculata* (Kohn et al. 1996; Barrett and Graham 1997). We included *E. paradoxa* to serve as an additional selfing phenotype and also as an outgroup for future studies of molecular
evolution. The approaches we describe demonstrate the utility of short-read sequencing for quantifying variation in gene expression among related samples.

**Methods**

*Sampling and RNA preparation*

We selected the four plants used in our study from glasshouse collections maintained at the University of Toronto. The plants were originally obtained by germinating open-pollinated seed collected from field populations at the following localities: outcrossing L-morph B211, Fortaleza, Ceará, N.E. Brazil; selfing M-morph J16, Georges Plain, Westmoreland, Jamaica; selfing L-morph N1, Rio Las Lajas, Rivas, Nicaragua; *E. paradoxa*, Patos, Paraiba, N.E. Brazil. We collected fresh tissue from different stages of bud and flower maturation to sample as many of the genes expressed during development as possible. Flower buds were classified into six sizes (< 3mm, 3-5mm, 5-7mm, 7-10mm, >10mm, open flower), with multiple buds for each stage and each stage represented equally among all individuals. Following bud and flower removal samples were immediately placed in liquid nitrogen to avoid RNA degradation. We extracted RNA from pooled bud samples from each individual using the Invitrogen (Carlsbad, CA) Trizol high salt precipitation extraction protocol. We visualized samples on a gel to provide an initial assessment of quality and quantified using a spectrophotometer.

*Sequencing*

We provided 5 µg total RNA to the Center for Analysis of Genome Evolution & Function (CAGEF) at the University of Toronto (Toronto, ON) for sequencing. The RNA
was sequenced using the Illumina (San Diego, CA) mRNA-Seq, paired-end protocol on a Genome Analyzer, GAII, for 40 cycles. This resulted in an average of ~38.9 x 10^6 total reads per sample or ~1.55 x 10^9 bp of sequence per sample (Table 5.2). It is common for the quality of bases from the 3’ end of Illumina reads to drop in quality, we therefore trimmed the 3’ end of reads when the Phred quality score dropped below $Q = 20$ (or 0.01 probability of error) for two consecutive bases. In addition we also trimmed all 5’ and 3’ stretches of ambiguous ‘N’ nucleotides.

*De novo assembly*

We performed *de novo* assembly on each sample separately using software packages designed for short read sequences including Abyss (Simpson et al. 2009), Edena (Hernandez et al. 2008), Velvet (Zerbino and Birney 2008) and Oases (D. R. Zerbino, European Bioinformatics Institute). To choose the optimal parameters for each method, we used a combination of BLASTx searches of the NCBI non-redundant protein database (NR), and summary statistics of the assemblies ($N50$, longest contig, number of contigs, proportion of reads assembled). Consideration of the summary statistics led us to finally choose Oases, which generated the longest assembled ESTs, with the best hits to NR in terms of low E-values. Oases is a program designed as an extension of Velvet, specifically released for assembly of transcriptome sequences. Unlike the other software mentioned above, Oases handles the uneven coverage of contigs due to variation in expression levels of the transcripts in the sample. An undesirable consequence of the type of algorithm used by Oases is a tendency to generate identical or near-identical contigs. We attempted to remove these to reduce redundancy in the dataset by comparing each transcriptome assembly to itself using
BLAST (Altschul et al. 1990; Altschul et al. 1997). Any pair of contigs that were > 98% identical over 95% of the length of the shorter contig were collapsed into a single contig.

Consensus transcriptome generation

To create a reference transcriptome we conducted a ‘four-way’ reciprocal BLAST (all pairwise comparisons) to identify all orthologous sequences. The goal here was to identify sequences that may not show similarity to other known proteins or ESTs, but which are expressed in more than one sample. This procedure allowed us to confirm a large proportion of our transcripts without having to rely on comparative searches to distantly related species. In addition, we were able to generate longer consensus sequences when one of the reciprocal best BLAST sequences was longer than the others. This was implemented using a custom Biopython script (Cock et al. 2009) and BLAST.

We compared each of the four individual redundancy-reduced transcriptome assemblies to each other using BLASTn (default parameters without low complexity filter). Reciprocal best BLAST hits found in more than two samples were then placed into groups and aligned using Muscle (Edgar 2004) to generate a consensus sequence. We defined a number of criteria to identify ortholog sequences including minimum alignment length (200bp), minimum sequence identity (90%), and minimum alignment proportion (> 80% of shorter sequence). This last criterion was used to avoid alternatively spliced transcripts or incompletely aligned contigs being collapsed in an alignment. After generating the consensus sequences with reciprocal BLAST we identified unaligned sequences that aligned well to the ortholog groups, but may not have been > 200bp. These sequences were incorporated into the consensus only when the contig extended the length of the consensus sequence and had > 95 % identity over > 50bp with no unalignable segments.
Due to low coverage or repetitive elements within coding loci it is possible that separate contigs are fragments of a single protein. To reduce fragmentation and recover longer coding sequences we aligned each contig to all unique *Oryza sativa* (another monocotyledon) proteins using BLASTx. We used *O. sativa* because it is the closest related plant for which an extensive set of protein sequences is available. This allowed us to identify consensus sequences that probably belong to the same protein and assemble them into a single contig. We aligned sequences that were potentially from the same protein enabling an elongated consensus to be generated. Only for a small number of contigs were found to potentially be fragments of longer ESTs (~1.6%) and all of the alignments made in Sequencher 4.7 were verified manually to ensure that no gaps, or mismatches were introduced.

After we assembled the consensus of all potential orthologs we identified sequences that were not included in these groups but had homologs in other species (hereafter referred to as singletons). We compared each singleton against NR and those over the size threshold of 1000bp and with a strong BLASTx hit (expectation or *E*-value < 1 x 10^{-15}) were included in the reference sequence along with all potential orthologs identified with our reciprocal BLAST scheme. To ensure that there was no remaining redundancy of transcripts in the consensus we used the same technique to reduce redundancy, as outlined above.

*Genotype calling and SNP detection*

For each sample we mapped the original short reads, trimmed using the base qualities as outlined above, to the consensus transcriptome using the Burrows-Wheeler Aligner (BWA) version 0.5.7 (Li and Durbin 2009). By mapping the reads back to the consensus transcriptome we were able to more accurately estimate coverage and use counts of
nucleotides at each position to call genotypes. BWA allowed us to vary the number of mismatches between the reference and aligned reads. We tested a number of parameter values for alignment and all analyses presented here use a value of $n = 0.05$, where $n$ is the fraction of missing alignments given 2% uniform base error rate. To generate a final sequence for each individual we included all loci where coverage was on average greater than five fold across the locus. Furthermore, within loci only sites where coverage exceeded five were used to call genotypes, the other sites were marked as ambiguous.

To generate genotypes for each sample we used the genotype calling method implemented in the software Maq (Li et al. 2008). This method takes into account the counts of different bases at each site as well as the quality of each base and the mapping quality of the sequence read. Maq uses a Bayesian statistical model to compare the inferred genotype to the original reference. To call genotypes we used a threshold ‘consensus quality’ cutoff of $Q > 13$ ($P = 0.05$), where $Q$ is the Phred-scaled probability that the consensus genotype call is wrong (Ewing and Green 1998; Ewing et al. 1998; Li et al. 2008). Sites for which we could not determine the consensus with at least this level of confidence were marked as ambiguous.

To detect potential errors in read mapping we assumed that selfing genotypes were largely homozygous and therefore the presence of heterozygous sites in multiple selfing genotypes may indicate errors in read mapping (see Discussion). These loci were excluded from downstream analyses. True heterozygosity was therefore estimated at loci where there was no evidence of read mapping errors. We used a subset of the identified loci that were shared between all four samples to assess the number of SNPs between pairs of genotypes. We also calculated nucleotide polymorphism values, $\theta_w$ (Watterson 1975) for the three *E. paniculata* sequences.
**Measurement of gene expression**

In addition to generating a consensus sequence, we also used abundance of reads derived from each locus to estimate gene expression. We calculated the number of fragments per kilobase per million fragments mapped (FPKM) with the program Cuffdiff from the package Cufflinks v 0.83 (Trapnell et al. 2010). This program estimates confidence intervals around expression estimates of each transcript using a Bayesian inference method and will identify significant differences in expression after correcting for multiple tests. To compare expression differences among selfing genotypes we identified loci in which both *E. paniculata* selfing genotypes, or all three selfing genotypes (including *E. paradoxa*) differed in expression relative to the outcrossing genotype of *E. paniculata* from Brazil. To estimate the overall similarity in expression we calculated the correlation of FPKM among pairs of plants after log transformation so that the data fitted a normal distribution. To test for significant differences among correlation coefficients we bootstrapped the data (10,000 replicates) to estimate 95% confidence intervals.

**Functional annotation**

To functionally annotate each gene and to assess the quality of our assembly we used the Gene Ontology (GO) based annotation suite Blast2GO v2.4.2 (Conesa et al. 2005; Götz et al. 2008). Blast2GO allowed us to identify similarity of the sequences in our reference transcriptome to known and predicted proteins, and to assign each of the sequences GO terms that were associated with the proteins found by BLASTx. We searched all 26,994 sequences against the non-redundant protein database (NR) with maximum *E*-value = $1 \times 10^{-15}$. Blast2GO assigned GO terms using a pro-Similarity-Hit-Filter of 15, an annotation cut-off of
55 and a GO weight of 5. We conducted enrichment analyses to test for an excess or paucity of gene classes (based on GO terms) in test sets relative to the whole reference transcriptome. These included: 1) genes absent in the two selfing *E. paniculata* genotypes (Jamaica and Nicaragua); 2) genes absent in the outcrossing genotype (Brazil), and 3) genes with low expression in the two selfing *E. paniculata* genotypes, and 4) genes with high expression in the two selfing *E. paniculata* genotypes. Additionally, we repeated these contrasts with *E. paradoxa* included in all sets with the Jamaican and Nicaraguan selfers. However, the results of these analyses were not informative because there were no over represented gene classes in each group and the results are therefore not presented here. All comparisons were implemented in Blast2GO, which uses a Fisher’s exact test to determine significance after controlling for multiple tests with a false discovery rate (FDR) = 0.05.

*Assessment of the accuracy of EST assembly*

To assess the quality of our assembly we compared the ESTs assembled in our consensus transcriptome with a set of 217 unique ESTs sequenced from *E. paniculata* with Sanger sequencing. These ‘Sanger-ESTs’ were sampled from a cDNA library generated from leaf and floral tissue (details in Chapter 4). We assembled and aligned forward and reverse strands of each EST using Sequencher 4.7 and edited chromatographs and alignments manually. We compared the two sets of ESTs to identify conflicts, in which pairs from each EST set share a significant portion of their sequence but cannot be aligned over another overlapping portion. This could be an indication of our Illumina ESTs having been assembled incorrectly, creating chimeric sequences of distinct transcripts. Because the set of Sanger ESTs was not exhaustive, we attempted to identify well-assembled ESTs from the
subset with conflicting alignments by comparing them to NR protein database to determine whether there is a known protein that covered the full length of the Illumina EST.

Results

Assembly and consensus transcriptome generation

*De novo* assemblies of each sample using Oases resulted in an average of 56,791 (50,581 - 61,922) contigs per sample, totaling approximately 21.3 Mbp of sequence for each individual. Many of the sequences were small, resulting in an $N50$ size of 611bp and mean contig size of 374bp (Figure 5.1). After removing highly similar sequences and contigs that were shorter than 100bp, the mean number of contigs per sample was 44,614, totaling on average 17.6 Mbp/sample.

The four-way reciprocal BLAST scheme returned consensus sequences of all contigs present in at least two samples. This resulted in a nearly two-fold reduction in the total number of contigs to 22,630, along with an increase in the $N50$ to 807bp and only a slight decrease in the total amount of sequence in the transcriptome, relative to each Oases assembly individually (15.9Mbp). We attempted to improve our consensus transcriptome by incorporating the remaining contigs into the consensus and by joining contigs that were fragments of the same protein. These steps had only a minor effect on our consensus, increasing the $N50$ to 819bp, the total length to 16.0 Mbp and decreasing the total number of contigs to 22,282.

BLAST identified all the unincorporated contigs that had similarity to known proteins. 5624 contigs over 1000bp had BLASTx hits with $E < 1 \times 10^{-15}$ to known proteins and were added into the consensus transcriptome. After we removed all remaining redundancy, the final consensus transcriptome consisted of 26,994 contigs representing 23.9Mbp of sequence with an $N50$ of 1129bp and a mean contig size of 884 bp. The final consensus
transcriptome resulted in a nearly two-fold increase in the \( N50 \), was comprised of \(~30,000\) few contigs than any of the original assemblies, and represented a slightly larger total transcriptome length (Table 5.3).

**Genotype calling and SNP detection**

Using our consensus transcriptome as a reference, we mapped the original short sequence reads for each sample with the software bwa-0.5.7. On average there was 22.0 reads covering each position in the reference with a standard deviation of 32.0 (Table 5.4). We generated, on average, 23,543 contigs for each sample. 18,063 of the 26,994 original contigs are found in all four samples and only 139 of the reference contigs are not recovered in any of the samples, likely due to low coverage. Few loci were unique to any single *E. paniculata* genotype, 5254 sequences were shared by all *E. paniculata* samples and not found in *E. paradoxa*. Moreover, 1392 loci were unique to *E. paradoxa*.

We identified heterozygous loci and potential read mapping errors as loci with one or more bases called as heterozygotes. Assuming that selfing genotypes are largely homozygous, the presence of heterozygous sites in multiple samples of selfers may indicate errors in read mapping (see Discussion). We identified 15,962 loci where there was no evidence of read mapping errors, 8469 loci with some evidence for read mapping errors, and 2563 loci were expressed in either one or zero selfers, precluding the application of this test. Within the loci for which we do not have evidence for read mapping errors, the number of heterozygous loci was highest in the outcrossing *E. paniculata* genotype (4979) compared to the two selfing genotypes from Jamaica (1659) and Nicaragua (895). *E. paradoxa* had an intermediate number (3994) of heterozygous loci. To detect the number of SNPs between pairs of genotypes, we selected a conservative set of 5,011 loci (4.2 Mbp) that were
expressed in all four individuals and were homozygous in all selfers (Table 5.5). The outcrossing Brazilian and selfing Nicaraguan genotypes had the fewest divergent sites (36,998). Intraspecific variation in *E. paniculata* was substantially lower than divergence of each *E. paniculata* sample was to *E. paradoxa*. (Table 5.5).

*Functional annotation*

23,476 of 26,994 contigs (86.97%) had similarity to known proteins in the NCBI NR database, with a cutoff of $E < 1 \times 10^{-10}$. 6329 loci had alignments which covered more than 80% of the top protein hits and 10,323 of the query sequences were at least 80% covered by their best protein hit (Figure 5.2). Blast2GO assigned a functional annotation to 21,779 of the loci (80.68%). Within the broad GO category ‘cellular components’ over a third of the sequences are localized to the plastid, 23.4% to the mitochondrion and 17% to the nucleus (Figure 5.3). A number of other cellular components make up the remaining 25.4 % of the annotated loci. Within the broad GO category ‘biological process’ the two most common type of genes were those involved in cellular (32.5%) and metabolic (31.4%) processes (Figure 5.3). 812 genes identified are involved in reproductive processes including flower development (269 genes) and pollination (60 genes).

*Gene expression*

Using FPKM to measure gene expression, we found significant correlation in expression among our samples (Figure 5.4). As expected, the correlation of each of the three *E. paniculata* samples with *E. paradoxa* was lower ($r$ from 0.60 - 0.63) compared with the correlation of *E. paniculata* genotypes with one another. The two independently derived selfing genotypes were slightly more correlated ($r_{JAM-NIC} = 0.93$), but not significantly more
so than either was to the outcrossing genotype from Brazil ($r_{BRA-JAM} = 0.91$, $r_{BRA-NIC} = 0.92$). There were 147 genes that were significantly up-regulated in all three selfing genotypes compared with the outcrosser, 12 of which were involved in reproduction. A similar number of genes (134) were down-regulated in the selfers relative to the outcrosser, 10 of these genes were involved in pollination or flower development (Table 5.6).

**Gene ontology enrichment tests**

We investigated whether there was an excess or paucity of particular gene classes that were differentially expressed in the two selfing lineages of *E. paniculata* compared to the outcrosser. In the up-regulated genes in *E. paniculata* selfers there were 146 GO categories that were significantly enriched. However, using a two-tailed Fisher’s exact test we found that only genes involved in photosynthesis (photosynthesis GO:0015979, thylakoid GO:0009579, plastid GO:0009536) were significantly over-represented and the remaining 145 classes of genes were under-represented. In the genes that were expressed at a lower level or completely absent in the selfers, 12 of 106 and 8 of 62 GO classes were significantly over-represented. Significantly, of the 12 classes of genes that were over-represented among the down-regulated genes in the two *E. paniculata* selfers, five are related to the regulation of cellular structure and development (regulation of cell morphogenesis GO:0022604, regulation of cellular component organization GO:0051128, regulation of developmental process GO:0050793, regulation of anatomical structure morphogenesis GO:0022603, regulation of cell size GO:0008361). In the *E. paniculata* selfers, the genes involved in flower development were significantly under represented in both higher and lower expression genes.
Assessment of EST assembly accuracy

Comparisons of our consensus transcriptome ESTs to a set of 217 Sanger sequenced ESTs revealed 11 Illumina ESTs (5.1%) that conflicted with their best Sanger EST match. These 11 ESTs each had regions that could not be aligned with the full length of the best BLAST hit from the Sanger ESTs and also did not have full-length BLAST hits to known proteins in NR. However, all 11 had highly significant hits to proteins in NR. None of the 11 ESTs appeared to be chimeric assemblies of distinct proteins that may have resulted from errors during assembly. Moreover, 9 of the 11 ESTs had sequence flanking their protein-coding region with significant similarity to known 5’ and 3’ untranslated regions from other monocot genomes.

Discussion

We assembled ~24 Mbp of transcriptome sequence in each of four individuals of two species of *Eichhornia*. The data represents an important genetic resource of nearly 27,000 transcripts, many of which are common to all four samples. Further, using read mapping to a consensus transcriptome we have generated statistically informed genotypes for each individual in our study. By choosing to extract RNA from buds and mature flowers we were able to recover many genes involved in reproduction and floral development, some of which are likely to provide future insights on genetic changes to floral traits governing mating-system variation. We now compare our assembly and analysis with previous *de novo* transcriptome sequencing projects and briefly review some of the challenges and interpretations specific to our assembly. In addition, we also discuss the utility of short-read
sequencing for characterizing genetic changes in transcriptomes and the expression level of different loci.

Assembly and consensus transcriptome

Our study represents one of the first published attempts at a de novo transcriptome assembly using short-read (Illumina GAII) sequencing. Although we are aware of 10 studies using next generation transcriptome sequencing, all have used some form of the Roche 454 platform (Table 5.1). As previously mentioned, this platform has the advantage of longer reads but at the expense of less sequence data per run. Longer reads may be critical for resolving assembly challenges associated with repetitive elements, such as gene duplications, allelic differences and alternative splicing (reviewed in (Pop 2009)). However, despite using shorter reads, our assembly is comparable to other published transcriptomes, which, on average, generated 27,684 contigs (12,883 - 71,384), a number similar to our own. In addition, despite the longer read lengths (average = 230 bp) of the Roche 454 platform, the mean contig size in previous transcriptome studies is ~ 400bp (197 - 526), whereas our assembly generated contigs that were over twice as long averaging 884 bp. One of the reasons that this is the case is because the Illumina generates greater depth of sequencing ensuring more complete coverage of the transcriptome. On average previous studies generated 109 Mbp of raw sequence whereas our study generated approximately 1,567 Mbp of sequence per sample.

Although summaries of the distribution of contig lengths are informative, the ultimate goal of transcriptome assembly is not long sequences, but accurate assembly of full-length sequences. However, it is difficult to assess the success of an assembly without a priori knowledge of the transcriptome. One metric that may be informative is the proportion of
contigs that have significant similarity to known proteins. It is difficult to compare this measure across studies because each reports slightly different results using different BLAST parameters and databases. However, nearly 87% of our contigs had matches in NR and this value is as high or higher than all other comparable statistics reported in other de novo assemblies. Another useful metric is the proportion of the contig and its corresponding BLAST hit that align (Figure 5.2). 7273 (26.9%) contigs cover greater than 75% of their best BLAST hit and 12,659 (47%) contigs are fully covered by their best BLAST hit. This means that although we assembled a large number of full-length proteins, many of the contigs appear to be fragments of larger proteins. One explanation is that gene duplication or alternative splicing creates repetitive elements in the assembly and these cannot be resolved. Although we found a substantial fraction of our ESTs that had conflicts with Sanger sequenced ESTs (5.1%), this is likely an overestimate of assembly error because some of these conflicts could result from paralogy or alternative splicing. There was no evidence that any of the 11 conflicting ESTs were chimeric assemblies of two or more proteins and most of the conflicts (9 of 11) appear to be the result of misaligned UTR flanking the coding region. It is possible that some of the discrepancies between the ESTs we assembled and Sanger ESTs or known proteins resulted from real biological differences in the sequence.

**Genotype calling and SNP detection**

One of the major challenges in dealing with very short sequence reads is that they must be assembled into longer contigs based on overlap with other reads. The algorithms used in many de novo assemblers, including Velvet (Zerbino and Birney 2008), may misinterpret small differences between alleles (SNPs) or gene duplicates as sequencing errors. If this occurs they can be ‘collapsed’ or purged from the assembly. Although the
program Oases has begun to address this problem for transcriptome data, we chose to use read mapping to a consensus transcriptome because it allowed us to use allele frequencies at each site to statistically determine the genotype of each individual (Li et al. 2008). From this approach we obtained on average more than 20 reads for each position to inform genotype calls. This allowed us to generate sequences for ~23,500 loci/genotype, 18,000 of which were found in all samples. An additional benefit of using this approach is that we were also able to identify heterozygous loci and potential read mapping errors. Of the 26,994 consensus sequences, 8712 heterozygous loci were identified. As expected, the fraction of heterozygous loci was highest in the outcrossing genotype of *E. paniculata* from Brazil. The selfing genotypes from Jamaica and Nicaragua appear to retain some residual heterozygosity despite their predominantly autogamous mating systems. We also found evidence of read mapping errors in 8469 loci where more than one selfer appeared heterozygous. Possible explanations for these read-mapping errors include, sequencing errors, alternative splicing and most likely paralogy. With our current data and the available methods we have no way of determining their relative contribution to read mapping errors.

Sequencing of multiple paralogous transcripts will generate short reads that are similar or identical to many other reads derived from different loci. As a result, when there has not been sufficient divergence between gene copies, reads may be erroneously mapped to the reference sequence. One possible explanation for false heterozygosity is that repetitive elements, for example, conserved motifs in gene families, may be difficult or impossible to assemble into long contigs using current technology. As a result, a fraction of the original short reads may not have been assembled by Oases, or included in our consensus transcriptome. However, if they share enough similarity with a paralog in the transcriptome they may be incorrectly mapped. This could explain, in part, why the original Oases
assemblies contain so many short contigs (< 100bp, see Figure 5.1). If there is divergence between paralogous loci, incorrectly mapped reads may create a signature similar to heterozygosity. For future analyses it is critical that potentially paralogous sequences are identified because evolutionary inferences from genes that are not orthologous are misleading. Although our approach does not eliminate paralogous sequences it has provided a useful method for detecting single copy transcripts.

Using a conservative set of 5011 transcripts for which there was no evidence of paralogy, based on homozygosity in all three selfers and presence in all four genotypes, we determined the number of SNPs between each pair of genotypes. As expected, *E. paniculata* samples were more differentiated from *E. paradoxa* than with one another (Table 5.5). The patterns of divergence among *E. paniculata* genotypes reflect relationships previously reported (see Figure 3.3). Specifically, the Nicaraguan selfing genotype is more similar to the outcrossing Brazilian genotype than it is to the selfing Jamaican genotype, despite more similar biogeographical origins and mating systems. This suggests that the Nicaraguan population is more closely related, or more recently derived from the Brazilian population. Further, when we calculate nucleotide polymorphism across 5011 sequences for the three *E. paniculata* genotypes the value we obtained ($\theta_W = 0.0104$) is comparable to our previously published species-wide estimate of total diversity based on 10 nuclear-derived EST loci assayed in samples of 225 individuals from 26 populations, $\theta_W = 0.0101$ (Chapter 4). This evidence supports the validity of our SNP detection method.

*Expression and enrichment*

There was a weak trend indicating that the selfing genotypes of *E. paniculata* were more correlated; however, this difference was not significant using bootstrapping to generate
95% confidence intervals. All three genotypes of *E. paniculata* retain highly correlated gene expression despite phenotypic divergence and geographic isolation. The slight elevation in correlated expression between the two selfing genotypes of *E. paniculata* may be caused by a small number of genes that are differentially expressed in both selfing genotypes (see below), although overall patterns of expression during flower development appear to remain largely conserved.

Enrichment tests of our annotated transcriptomes demonstrated that genes that were differentially expressed in selfers exhibit a paucity of particular gene classes. This can be interpreted as the conservation of expression of these gene functions, which are rarely differentially expressed. Of the 313 GO categories found to be significantly enriched among all differentially expressed genes only 21 were found to be over-abundant, and 20 of these 21 categories were over-represented in genes absent or expressed at lower levels in the two selfers. Therefore, it appears that many of the differences common to the selfing lineages of *E. paniculata* are associated with reductions of gene expression in floral tissue. This may be related to the convergence of floral traits in these two lineages, both of which have much smaller, less pigmented flowers, with reduced pollen production compared to the outcrossing genotype. 96 of the 108 gene classes that are expressed at a low level in the selfers are under-represented but 5 of the 12 over-represented GO categories were associated with the regulation of cell development and structure. Selfing flowers commonly display modified stamen positions and floral instability including twisted, fused or missing perianth parts, whereas outcrossing plants rarely display these floral modifications (Richards and Barrett 1992). It is possible that the changes in gene expression we have documented influence the regulation of cell growth and division and are responsible for changes to floral morphology.
that characterize selfing populations. If so, these regulatory loci could be used as a set of
candidate genes to investigate aspects of the evolution of the selfing syndrome.

Differentially expressed floral genes in selfers

By sampling different stages of floral development up to and including anthesis we
were able to sequence and annotate a large number of florally expressed genes. In total 812
genes with the GO annotation ‘reproduction’ were identified, which is a large fraction of the
number reported for *Arabidopsis thaliana* in which 1184 genes for reproduction have been
documented (Swarbreck et al. 2008). The lower number of genes in our annotation is not
unexpected because we did not include tissue from reproductive stages after flowering, such
as fruit and seed development. Within the GO category ‘reproduction’ we found 269 genes
involved in floral development, similar to the number that has been annotated in *A. thaliana*
(323 genes). Of particular significance are the floral development genes that are differentially
expressed in the three selfing genotypes (Table 5.6), several of which affect structures that
are modified in the selfers. Anther development and filament elongation are influenced by
*AFB2* (Cecchetti et al. 2008), *ARF6, ARF8* (Nagpal et al. 2005), *BAM1* (Hord et al. 2006),
*GiD1c* (Cheng et al. 2004) *myb24* (Mandaokar and Browne 2009), *PGP1* (Cecchetti et al.
2008) and *PTR1* (Komarova et al. 2008) genes in *A. thaliana* and pollen maturation and
pollen tube growth are altered by *AFB2, CSDL3* (Bernal et al. 2008), *CER1* (Aarts et al. 1995),
*POP2* (Hiscock and Allen 2008), *myb24* (Mandaokar and Browne 2009) and *TIR1* (Cecchetti
et al. 2008). *ERL1* plays an important roles in normal anther lobe formation and anther cell
differentiation (Hord et al. 2008) and mutants in the *ERECTA* gene family have reduced
lateral organ size and abnormal flower development, including defects in petal polar
expansion and carpel elongation (Shpak et al. 2004). We also identified a number of genes
involved in flowering time including \textit{DCL1} (Schmitz et al. 2007), \textit{PGI1} (Yu et al. 2000), \textit{SHK1} (Wang et al. 2007), \textit{SUF4} (Kim et al. 2006). Lastly, all of the differentially expressed genes that effect ovule development were significantly up-regulated, including \textit{ARF6}, \textit{ARF8} (Nagpal et al. 2005), \textit{OVA9} (Berg et al. 2005), \textit{PEP} (Ripoll et al. 2006). Significantly, most of the candidate genes discussed above, cause alterations to attractive structures (perianth) and male function (stamen development) consistent with the relaxation of selection for showy flowers, reduced allocation to pollen production and the loss of herkogamy through filament elongation of stamens. In contrast, the requirement for functional ovules to maintain seed fertility in selfers may explain the apparent absence of changes in gene expression to female traits.

We have shown that short-read sequencing can be used to characterize the transcriptomes of multiple individuals for use in comparative studies. We were able to assemble as many contigs as other sequencing methods by \textit{de novo} assembly, but they were on average substantially longer. By comparing sequences from related individuals we generated a consensus transcriptome. This allowed us to make SNP genotype calls and provided a method for detecting paralogous sequences. Discerning among copies of paralogous sequences presents a major challenge to the current technology and requires either technological or analytical solutions to discern among different members of gene families or duplicates. However, despite these complications we annotated > 80% of contigs and identified thousands of putative orthologs, many of which are differentially expressed. We identified 22 genes that were differentially expressed in selfers and which have developmental functions that suggest a role in the evolution of the selfing syndrome. This sequencing effort has generated a valuable resource of coding DNA for a non-model species. The transcriptome sequences will help in future studies of changes in the genetic architecture
involved in the transition from outcrossing to selfing and also for identifying the genes controlling heterostyly.
Table 5.1. Sequencing and assembly statistics from previously published *de novo* transcriptome assemblies.

<table>
<thead>
<tr>
<th>Species</th>
<th>Total raw sequence</th>
<th>Mean read length</th>
<th>No. of reads</th>
<th>No. of contigs</th>
<th>Mean contig length</th>
<th>Reads per contig</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Eucalyptus grandis</em> (Novaes et al. 2008)</td>
<td>148.4 Mbp</td>
<td>145 bp</td>
<td>1,024,251</td>
<td>71,384</td>
<td>247 bp</td>
<td>12.7</td>
</tr>
<tr>
<td><em>Melitaea cinxia</em> (Vera et al. 2008)</td>
<td>66.9 Mbp</td>
<td>110 bp</td>
<td>608,053</td>
<td>48,354</td>
<td>197 bp</td>
<td>12.6</td>
</tr>
<tr>
<td><em>Acropora millepora</em> (Meyer et al. 2009)</td>
<td>146.3 Mbp</td>
<td>232 bp</td>
<td>628,649</td>
<td>44,444</td>
<td>440 bp</td>
<td>5.0</td>
</tr>
<tr>
<td><em>Sarcophaga crassipalpis</em> (Hahn et al. 2009)</td>
<td>49.9 Mbp</td>
<td>241 bp</td>
<td>207,110</td>
<td>20,995</td>
<td>332 bp</td>
<td>8.6</td>
</tr>
<tr>
<td><em>Zoarces viviparous</em> (Kristiansson et al. 2009)</td>
<td>-</td>
<td>-</td>
<td>400,000</td>
<td>36,110</td>
<td>395 bp</td>
<td>8.0</td>
</tr>
<tr>
<td><em>Castanea dentata</em> (Barakat et al. 2009)</td>
<td>81.1 Mbp</td>
<td>223 bp</td>
<td>418,923</td>
<td>12,883</td>
<td>273 bp</td>
<td>-</td>
</tr>
<tr>
<td><em>Castanea mollissima</em> (Barakat et al. 2009)</td>
<td>144.1 Mbp</td>
<td>222 bp</td>
<td>724,088</td>
<td>15,085</td>
<td>330 bp</td>
<td>-</td>
</tr>
<tr>
<td><em>Pinus contorta</em> (Parchman et al. 2010)</td>
<td>179.5 Mbp</td>
<td>306 bp</td>
<td>586,732</td>
<td>63,657</td>
<td>452 bp</td>
<td>3.1</td>
</tr>
<tr>
<td>10 bird species&lt;sup&gt;2&lt;/sup&gt; (Künstner et al. 2010)</td>
<td>109.1 Mbp</td>
<td>229 bp</td>
<td>525,752</td>
<td>20,631</td>
<td>453 bp</td>
<td>9</td>
</tr>
<tr>
<td><em>Amphiholophus astorquiit</em> (Elmer et al. 2010)</td>
<td>60.7 Mbp</td>
<td>202 bp</td>
<td>300,610</td>
<td>24,174</td>
<td>300 bp</td>
<td>7.57</td>
</tr>
<tr>
<td><em>Amphiholophus zaliosus</em> (Elmer et al. 2010)</td>
<td>54.1 Mbp</td>
<td>206 bp</td>
<td>262,494</td>
<td>21,382</td>
<td>299 bp</td>
<td>7.39</td>
</tr>
<tr>
<td><em>Panax quinquefolius</em> (Sun et al. 2010)</td>
<td>89.6 Mbp</td>
<td>427 bp</td>
<td>209,747</td>
<td>16,592</td>
<td>526 bp</td>
<td>8</td>
</tr>
</tbody>
</table>

<sup>1</sup> In many of these studies, analyses after contig assembly included reads not incorporated into contigs. Here, we only list the number of contigs for brevity.

<sup>2</sup> In Künstner et al. (2010) the transcriptome of 10 bird species is analyzed, we present averages of the data provided in the original article.

<sup>3</sup> Reads per contig is slightly different from reads/bp. Reads/bp can be roughly estimated from this data as Reads/contig * (Read length / Contig length).
Table 5.2. Sequencing statistics for this study. The number of reads and total amount of sequence generated for each sample. Values are shown for both raw sequence and the sequence after trimming from reads low quality bases.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total reads</th>
<th>Total raw bases</th>
<th>Reads after trimming</th>
<th>Bases after trimming</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. paradoxa</em> - Brazil</td>
<td>39,198,840</td>
<td>1.57 x 10^9</td>
<td>24,539,860</td>
<td>812,538,461</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Brazil</td>
<td>38,374,454</td>
<td>1.54 x 10^9</td>
<td>24,796,504</td>
<td>827,426,981</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Jamaica</td>
<td>39,047,450</td>
<td>1.56 x 10^9</td>
<td>23,555,014</td>
<td>764,543,073</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Nicaragua</td>
<td>38,980,224</td>
<td>1.56 x 10^9</td>
<td>22,323,972</td>
<td>740,370,472</td>
</tr>
</tbody>
</table>
Table 5.3. Summary statistics for reference transcriptome through progressive stages of assembly. For the first stage all statistics are the average across the four independently sequenced genotypes.

<table>
<thead>
<tr>
<th>Assembly stage</th>
<th>N50</th>
<th>Mean contig size</th>
<th>Total length</th>
<th>No. of contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oases assembly</td>
<td>611</td>
<td>374.1</td>
<td>21,311,238</td>
<td>56,971</td>
</tr>
<tr>
<td>Reciprocal BLAST</td>
<td>807</td>
<td>703.5</td>
<td>15,919,812</td>
<td>22,630</td>
</tr>
<tr>
<td>Contig elongation</td>
<td>812</td>
<td>709.9</td>
<td>16,065,194</td>
<td>22,630</td>
</tr>
<tr>
<td>Reduced fragmentation</td>
<td>819</td>
<td>717.8</td>
<td>15,995,086</td>
<td>22,282</td>
</tr>
<tr>
<td>Final (including singletons)(^1)</td>
<td>1,129</td>
<td>884.3</td>
<td>23,869,762</td>
<td>26,994</td>
</tr>
</tbody>
</table>

\(^1\) Singletons over 1000bp with BLASTx hits where \(e < 1 \times 10^{-15}\) are included in the final consensus
Table 5.4. Results of read mapping using the original trimmed short reads for each of the four *Eichhornia* samples against the reference transcriptome.

<table>
<thead>
<tr>
<th>Sample</th>
<th>No. of loci</th>
<th>Mean coverage (per bp ± S.D.)</th>
<th>No. of heterozygous loci</th>
<th>Total sequence (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. paradoxa</em> - Brazil</td>
<td>20,653</td>
<td>21.8 ± 33.7</td>
<td>3,994</td>
<td>17,308,982</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Brazil</td>
<td>24,849</td>
<td>24.3 ± 34.1</td>
<td>4,979</td>
<td>21,362,493</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Jamaica</td>
<td>24,527</td>
<td>21.7 ± 32.1</td>
<td>1,659</td>
<td>20,943,070</td>
</tr>
<tr>
<td><em>E. paniculata</em> - Nicaragua</td>
<td>24,142</td>
<td>20.5 ± 27.9</td>
<td>895</td>
<td>20,603,010</td>
</tr>
</tbody>
</table>
Table 5.5. Number of single nucleotide polymorphisms (SNPs) between pairs of *Eichhornia* genotypes. Values reflect the polymorphism statistics for 5011 loci that were present in all four samples and homozygous in the three selfing genotypes. Values below the diagonal show the total number of SNPs and values above the diagonal show the number SNPs per bp (4,207,280bp total)

<table>
<thead>
<tr>
<th>Sample</th>
<th><em>E. paradoxa</em></th>
<th><em>E. paniculata</em></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Brazil</td>
<td>Jamaica</td>
<td>Nicaragua</td>
<td></td>
</tr>
<tr>
<td><em>E. paradoxa</em> - Brazil</td>
<td>-</td>
<td>0.048</td>
<td>0.048</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td><em>E. paniculata</em> - Brazil</td>
<td>202,687</td>
<td>-</td>
<td>0.014</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td><em>E. paniculata</em> - Jamaica</td>
<td>200,576</td>
<td>58,410</td>
<td>-</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td><em>E. paniculata</em> - Nicaragua</td>
<td>195,967</td>
<td>36,998</td>
<td>51,409</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.6. Pollen and flower development genes that were differentially expressed in both independently derived *Eichhornia paniculata* selfing genotypes and were also identified from selfing *Eichhornia paradoxa*.

<table>
<thead>
<tr>
<th>Homolog name</th>
<th>Fold expression change&lt;sup&gt;1&lt;/sup&gt;</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibberellin receptor (<em>GID1c</em>)</td>
<td>0.51****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peptide transport protein (<em>PTR1</em>)</td>
<td>0.62****</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ERECTA</em>-like 1 (<em>ERL1</em>)</td>
<td>0.64*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>DICER</em>-like 1 dsrna-specific nuclease (<em>DCL1</em>)</td>
<td>0.77**</td>
<td>Differentially expressed in</td>
<td></td>
</tr>
<tr>
<td>Exocyst complex component (<em>SEC5</em>)</td>
<td>1.29*</td>
<td><em>E. paniculata</em> selfers</td>
<td></td>
</tr>
<tr>
<td>SHK1 binding protein 1 (<em>SKB1</em>)</td>
<td>1.38**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>REBELOTE</em> (<em>RBL</em>)</td>
<td>1.64**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cellulose synthase (<em>CSLD3</em>)</td>
<td>0.55***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pollen-pistil incompatibility 2 (<em>POP2</em>)</td>
<td>0.62****</td>
<td>Differentially expressed in</td>
<td></td>
</tr>
<tr>
<td><em>ECERIFERUM</em> 1 (<em>CER1</em>)</td>
<td>0.70***</td>
<td>all selfers</td>
<td></td>
</tr>
<tr>
<td>Auxin signaling F-box 2 (<em>AFB2</em>)</td>
<td>0.71****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transport inhibitor response 1 (<em>TIR1</em>)</td>
<td>0.89*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta-amylase (<em>BAM1</em>)</td>
<td>0.99****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auxin response factor 8 (<em>ARF8</em>)</td>
<td>1.08****</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>PEPPER</em> nucleic acid binding protein (<em>PEP</em>)</td>
<td>1.15**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suppressor of <em>FRIGIDA4</em> (<em>SUF4</em>)</td>
<td>1.18*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphoglucose isomerase (<em>PGI1</em>)</td>
<td>1.35**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Description</td>
<td>Ratio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auxin response factor 6 (ARF6)</td>
<td>1.50**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory particle non-atpase 10 (RPN10)</td>
<td>1.50*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphatidylglycerolphosphate synthase 1 (PGP1)</td>
<td>1.60*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>myb transcription factor (myb24)</td>
<td>1.79****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutaminyl-tRNA synthetase (OVA9)</td>
<td>1.93****</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$

1 Measured as the mean ratio of expression (FPKM) in each selfer relative to the outcrossing genotype
Figure 5.1. Histogram of the frequency of different contigs sizes in transcriptome assemblies of *Eichhornia* samples. Blue bars represent the distribution of contig sizes for the initial *de novo* Oases assemblies and the red bars represent the final consensus transcriptome.
Figure 5.2. Proportion of each of the assembled sequences that is aligned to its top BLASTx protein hit versus the proportion of the top hit which is covered by the assembled query sequence. Dense clusters of points along the top of the figure represent loci entirely aligned to their respective protein hits and points along the right are genes fully covering their best BLAST protein hit.
Figure 5.3. Distribution of genes in the transcriptome assembly assigned to broad GO categories, cellular components (blue) and biological processes (red). Percentages indicate the proportion of sequences assigned within each subcategory.
Figure 5.4. Pairwise correlations of gene expression between the four genotypes: *Eichhornia paradoxa*, and *Eichhornia paniculata* - Brazil, Jamaica and Nicaragua. Above the diagonal are the correlation coefficients for data plotted below the diagonal. Each pair below the diagonal is plotted on a log scale of expression level, measured in FPKM. Every correlation is significant at $P < 0.00001$. 

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Brazil outcrosser</th>
<th>Jamaica selper</th>
<th>Nicaragua selper</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. paradoxa</em></td>
<td>$r = 0.63$</td>
<td>$r = 0.60$</td>
<td>$r = 0.61$</td>
</tr>
<tr>
<td></td>
<td>$r = 0.91$</td>
<td></td>
<td>$r = 0.92$</td>
</tr>
<tr>
<td></td>
<td>$r = 0.93$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER SIX

GENOMIC CONSEQUENCES OF TRANSITIONS FROM CROSS- TO SELF- FERTILIZATION ON THE EFFICACY OF SELECTION IN THREE INDEPENDENTLY DERIVED SELFING PLANTS

This chapter resulted from a collaboration Mathieu Siol and Spencer C. H. Barrett. All authors contributed to the writing and ideas presented in this chapter. Mathieu Siol conducted analyses of protein sequence evolution.

Summary

Transitions from cross-fertilization to self-fertilization reduce the effective population size. As a result, slightly deleterious or advantageous mutations can be rendered effectively neutral because of the increased importance of genetic drift over natural selection. The effect of drift is predicted to reduce selective constraints on protein sequences and cause relaxed selection for biased codon usage. Here, we investigate patterns of nucleotide variation in protein-coding genes to assess the effect of inbreeding on the accumulation of deleterious mutations in three independently evolved selfing lineages of annual plants. Using high throughput sequencing, we assembled the floral transcriptomes of four genotypes of Eichhornia (Pontederiaceae); these included one outcrosser and two independently derived selfers of E. paniculata and E. paradoxa, a selfing outgroup. This dataset included ~8000 orthologous sequences totalling ~3.5 Mb of coding DNA. Several tests of selection provided results that were consistent with the occurrence of purifying selection across the genome. There was a small but significant elevation in the proportion of non-synonymous to synonymous changes in the two E. paniculata selfers relative to the outcrosser. Further, two lines of evidence supported a reduction in selection for biased codon usage in selfing lineages. Measurements of differentiation in codon usage between high vs. low expression
genes indicated significantly lower codon usage in the three selfers. Second, estimates of the selection coefficient for codon usage bias conducted in a phylogenetic framework also supported relaxed selection on synonymous sites in selfing genotypes. Our findings, although based on limited sampling of genotypes, suggest a reduced efficacy of selection in selfers, especially with respect to weakly selected synonymous changes that affect codon usage.

Introduction

Transitions from cross- to self-fertilization can have profound effects on population genetic structure and patterns of molecular evolution across the genome (reviewed in Chapter 2). Most importantly, homozygosity increases with more intense selfing thus decreasing effective population size \( N_e \) and reducing opportunities for crossing over between heterozygous sites and increasing linkage disequilibrium among loci (Charlesworth et al. 1993a; Nordborg 2000). Effective population size is also reduced by the effects of genetic hitchhiking, including selective sweeps of beneficial mutations and background selection against deleterious mutations (reviewed in Charlesworth and Wright 2001). Linkage among weakly selected sites with opposing selective forces can also interfere with the ability of selection to act efficiently (McVean and Charlesworth 2000).

Estimates of \( N_e \) in selfers are often lower than the expected two-fold decrease based on selfing alone. This result presumably occurs because life-history characteristics associated with selfing often promote population subdivision, isolation, and frequent genetic bottlenecks (Charlesworth and Pannell 2001; Ingvarsson 2002; Foxe et al. 2009; Guo et al. 2009; Chapter 4). Thus, both genetic and demographic processes operating in selfing populations should lead to a decrease in the efficacy of natural selection and an increase in the fixation of slightly deleterious mutations with important consequences for genome evolution. The
accumulation of deleterious mutations may also be an important factor in causing species extinction (Lynch et al. 1995) and could explain the lack of persistence of selfing lineages (reviewed in Takebayashi and Morrell 2001; Igic et al. 2008; Goldberg et al. 2010). However, the extent to which these theoretical predictions on the reduced efficacy of selection in selfing populations occur is unclear.

The efficacy of selection depends on the product of $N_e$ and the selection coefficient $(s)$. As a result of the reduction in $N_e$ due to selfing, a higher rate of fixation of slightly deleterious mutations and a lower rate of fixation for advantageous mutations is expected. Diverse methods have been developed to detect the footprint of natural selection at the molecular level (reviewed in Nielsen 2005; Siol et al. 2010) and one common approach is to quantify the ratio of mutations at non-synonymous sites ($d_N$) versus synonymous or silent sites ($d_S$); hereafter $\omega$. Because selection acts primarily on proteins and not DNA sequences, synonymous changes are selectively neutral thus enabling measurement of the degree of selective constraint on amino acid sequences. Under neutrality $\omega$ is expected to equal 1, whereas $\omega$ less than or greater than 1 indicates purifying selection or positive selection, respectively. Empirically, the vast majority of functional proteins that have been examined exhibit $\omega$ values much less than one indicating that most protein sequences are subject to purifying selection. However, in selfers a reduction in the efficacy of selection may result in elevation of this value as a result of the accumulation of deleterious mutations.

Although predictions concerning the effect of selfing on levels of polymorphism have been well documented (Hamrick and Godt 1996; Nybom 2004), evidence for a reduction of selection efficacy in selfing populations of diverse plants and animals is equivocal. The only study that we are aware of that has attributed an accumulation of deleterious mutations to selfing was a comparison of the selfing plant Arabidopsis thaliana with Drosophila.
melanogaster (Bustamante et al. 2002). Other studies (Wright et al. 2002; Glémin et al. 2006; Cutter et al. 2008; Haudry et al. 2008; Chapter 2; Escobar et al. 2010) focusing on closely related outcrossers and selfers have failed to detect a clear signal of reduced selection efficacy leading to several hypotheses to explain the apparent lack of signal in the molecular data. First, the genomic distribution of selection coefficients is still poorly known (but see Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), and if there are few weakly selected mutations very little effect of the mating system is predicted (Glémin 2007). Another explanation involves the amount of time that has elapsed since the transition from outcrossing to selfing, which in some instances may be too short for substantial changes to have occurred at the genome level (Chapter 2). Finally, few comparisons have been made using genomic data involving many loci and it is possible that because of the stochasticity and slowness of mutation accumulation much larger samples are necessary to detect such effects.

The reduced efficacy of selection is also predicted to diminish the signal of biased codon use. Proportional usage of synonymous codons often differs between high and low expression genes due to selection for higher translational efficiency and accuracy (Bulmer 1991; Duret and Mouchiroud 1999). However, the strength of selection on synonymous codon usage is expected to be weak relative to purifying selection against amino acid altering mutations. If this is true even slight reductions in $N_e$, such as those associated with transitions to selfing, are predicted to reduce the efficacy of selection to the extent that these mutations become effectively neutral (Akashi 1995; Kreitman and Antezana 1999). Thus, in principle the reduced efficacy of selection can be inferred from an increase in the frequency of substitutions (or polymorphisms) for unpreferred codons, or by less differentiation of codon usage between high and low expression genes. However, because codon bias is eroded by
genetic drift this process will occur rather slowly and be dependent on the rate of mutation (Marais et al. 2004). Therefore, to detect the reduced efficacy of selection based on codon usage it may be necessary to include either relatively old selfing lineages, which may be difficult given the relatively recent origin of most selfers, or to obtain data from a very large number of loci, to detect the small changes that are likely in codon usage.

Empirical tests of these predictions have included selfing and outcrossing species of *Arabidopsis* (Wright et al. 2002), *Caenorhabditis* (Cutter et al. 2006b; 2008) and members of the Tritaceae (Haudry et al. 2008). The results of these studies are mixed, with no evidence of reduced codon bias in selfing species of *Arabidopsis* and Tritaceae and only a slight reduction in codon bias in selfing *C. briggsae* relative to outcrossing *C. sp. 5*. Two explanations have been proposed for the lack of a strong effect of selfing on codon usage. First, as mentioned above, many selfing lineages are thought to be of recent origin (e.g. Cutter et al. 2008), and there simply may not have been sufficient time for enough mutations to have drifted to fixation. Second, in interspecific comparisons confounding effects of demography (such as bottlenecks or population structure) can distort conclusions about the effect of selfing on codon bias (Wright et al. 2002). Ideally, predictions require contrasts between conspecific selfing and outcrossing lineages with few confounding effects; however, this may not be feasible over long evolutionary time spans. The approach we use in this study is to contrast both inter and intraspecific selfing lineages of different ages and to use a large number of loci in an effort to detect changes in the efficacy of selection on genomes.

Here we investigate patterns of molecular evolution in the floral transcriptomes of three independently derived selfing lineages relative to an outcrossing genotype in *Eichhornia* (Pontederiaceae), a neotropical genus of aquatic plants. Our samples include three individuals of *E. paniculata*, an annual diploid that has been the subject of detailed
studies on the ecology and genetics of mating-system variation over the past two decades (reviewed in Chapter 3). Populations of *E. paniculata* are largely concentrated in N.E. Brazil, with smaller foci in Jamaica and Cuba and isolated localities in Nicaragua and Mexico. Populations in Brazil are mostly outcrossing and possess the sexual polymorphism tristyly, which promotes cross-pollination among the three floral morphs (reviewed in Barrett et al. 1992). Morphological, genetic and biogeographical evidence indicates that in *E. paniculata* tristyly has broken down on multiple occasions resulting in independently derived selfing populations (Husband and Barrett 1993; Fenster and Barrett 1994; and see Chapter 3). Populations from Jamaica are largely composed of selfing variants of the mid-styled morph (M-morph) in which stamens are elongated to a position adjacent to mid-level stigmas resulting in the autonomous self-fertilization of flowers. In contrast, plants from Mexico and Nicaragua are selfing variants of the long-styled morph (L-morph) with a different arrangement of their sexual organs (see Figure 3.2). Although both variants possess the selfing syndrome, comparisons of molecular variation at 10 EST-derived nuclear loci indicated a high level of genetic differentiation consistent with their separate origins from different outcrossing ancestors (see Figure 3.3). Our analysis included an individual of both selfing variants and an individual of an outcrossing L-morph from N.E. Brazil, the likely centre of origin of the species. We also included a selfing individual of *E. paradoxa*, the sister species of *E. paniculata* (e.g. Graham et al. 2002), to serve as a potentially more ancient selfing phenotype, and as an outgroup for inferences on the ancestral DNA sequence in our samples. Most populations of *E. paradoxa* are predominantly selfing although a tristylos population is known from Brazil indicating that, in common with *E. paniculata*, selfing has likely arisen from the evolutionary breakdown of tristyly (see Barrett 1988).
We used high-throughput DNA sequencing technology to generate a set of approximately 8000 orthologous ESTs from the floral transcriptomes of the four *Eichhornia* genotypes. Using this data set we investigated the molecular evolution of protein-coding genes to address the following specific questions predicted by the hypothesis of reduced efficacy of selection in selfers: (1) Is there evidence for a relaxation of purifying selection against non-synonymous mutations in selfing lineages? (2) What evidence is there that positive selection plays an important role in the floral transcriptomes of the four samples? (3) Can we detect the signature of selection of biased codon usage in our samples, and if so, does this vary among lineages based on their mating systems?

**Methods**

*Samples, tissue preparation and sequencing*

We selected the four plants from glasshouse collections maintained at the University of Toronto. These were originally obtained by germinating open-pollinated seed collected from field populations at the following localities: outcrossing L-morph B211, Fortaleza, Ceará, N.E. Brazil; selfing M-morph J16, Georges Plain, Westmoreland, Jamaica; selfing L-morph N1, Rio Las Lajas, Rivas, Nicaragua; *E. paradoxa*, Patos, Paraíba, N.E. Brazil. Plant material for each genotype was collected from multiple stages of floral development and sequences were generated using the paired-end mRNAseq protocol for the Illumina Solexa GAII. We assembled transcripts using a custom pipeline that generated a consensus transcriptome and used this consensus as a reference for read mapping. Expression for each transcript from each genotype was measured as the number of sequence reads mapped per kilobase per million reads mapped (FPKM) using the program Cuffdiff from the package
Cufflinks v 0.83 (Trapnell et al. 2010). For full details of de novo transcriptome assembly see Chapter 5.

**EST analysis, filtering and alignments**

To assign open reading frames and repair potential indels inserted during transcript assembly we used the pipeline prot4EST (Wasmuth and Blaxter 2004). We started from an extended list of 10,263 single-copy ESTs that were assembled in all four genotypes. Nucleotide sequences were aligned on the basis of their translated amino-acid sequences with the software Muscle (Edgar 2004) using the backalign function of the seqlib/egglib package (Stephan De Mita and Mathieu Siol, unpublished). We removed alignments with frame shifts and the remaining alignments were confirmed visually. Our final dataset was comprised of 7914 alignments totalling 4,965,756 bp of sequence and these were used for all subsequent analyses. Sequences have been submitted to GenBank under accession numbers XXX.

**Calculation of non-synonymous and synonymous mutations**

We employed several approaches to measure selection on protein-coding mutations. We conducted analyses of the rate of non-synonymous to synonymous substitution in each genotype with the software package PAML v4.4 (Yang 2007). The package includes the program codeml, which uses a maximum likelihood approach to estimate substitution ratio (\( \omega \)) across the branches of a tree under various models of sequence evolution. We ran four models allowing for variation in \( \omega \) among branches. The first model, ‘single \( \omega \)’, fixes a single \( \omega \) value for all the branches of the tree; the second model, “relaxed selection”, estimates one ratio for the outcrosser and one ratio for the three selfers, and model ‘three \( \omega \)’ estimates three ratios, one for the outgroup (*E. paradoxa*), one for the two *E. paniculata* selfers and one for
the Brazilian *E. paniculata* outcrosser. Finally the last ‘full model’ estimates four \( \omega \) values, one for each genotype in our study. We set the basal branch leading to the clade including the outcrosser to have the same \( \omega \) value as the terminal branch to the outcrosser because it is probable that the common ancestor to this group was tristylos and outcrossing (see Kohn et al. 1996). To alleviate convergence issues, four replicate runs were performed for each model and the run with the highest likelihood was retained for each model. We used likelihood-ratio tests to assess whether more complex models provided significantly better fits than simpler models. Significance was assessed using a \( \chi^2 \) approximation. We ran each of these models on all loci from each genotype concatenated and also on each gene separately. All analyses were based on the concatenated tree constructed using PhyML v3.0 (Guindon and Gascuel 2003) with a general time reversible (GTR) substitution model. We did this because there were often too few informative sites to infer the underlying topology from a single locus. Furthermore, the concatenated tree is consistent with previously reported relationships between Brazilian, Jamaican and Nicaraguan populations (see Chapter 3).

We used an alternative way to investigate an increase in the proportion of non-synonymous changes with fewer underlying assumptions by counting the number of non-synonymous and synonymous derived mutations in each *E. paniculata* genotype, using the outgroup *E. paradoxa* to infer the ancestral state. We then tested for significant differences in the proportion of derived non-synonymous to synonymous mutations in each individual using a \( \chi^2 \) homogeneity test with two degrees of freedom.

**MK tests of selection**

The McDonald-Kreitman (MK) test is a robust method to detect selection using polymorphism and divergence data and has been widely used to assess the strength of
purifying selection and to detect positive selection (Smith and Eyre-Walker 2002; Sella et al. 2009). We constructed a contingency table with columns for divergent and polymorphic sites and rows for non-synonymous and synonymous sites. Significant differences in the ratio of the two rows were then tested using the Fisher’s exact test. We calculated the MK tables with the Bio++ library (Dutheil et al. 2006) using a Python wrapper implemented in the seqlib/egglib package (Stephan De Mita and Mathieu Siol, unpublished). We applied the MK test to each polymorphic locus in E. paniculata. The MK test using individual loci had very low power when there were few polymorphisms; we therefore created two sets of loci, one in which MK tables containing zeros were discarded, and another more restrictive set in which we set a threshold such that the marginal sums of each contingency table was greater or equal to five. Given the large number of tests required, we used a multiple testing correction from the R package QVALUE (Storey and Tibshirani 2003). In addition to this simple application of the test, we used the software MK-Test 2.0 (Welch 2006) which implements multi-locus MK tests in a maximum likelihood framework to contrast four models. The models were: 1) the proportion of new neutral mutations \( (f) \) is fixed to one at all loci (no purifying selection) and the proportion of adaptive substitutions \( (\alpha) \) is fixed to zero; 2) a single value of \( f \) is fit to all loci and \( \alpha \) is fixed to zero; 3) \( f \) is fixed to one and \( \alpha \) takes a single value at all loci, and 4) both \( f \) and \( \alpha \) are fit to a single value across all loci. Each estimation used maximum likelihood and we assessed the relative support for the different models using a “second order” Aikaike information criterion (AICc, see Welch 2006).

Assessing selection for codon bias

To measure codon bias we calculated the frequency of optimal codon usage \( (F_{\text{op}}) \) for each gene from all four genotypes. Because optimal codons can vary among species we first
identified these codons based on the assumption that codon bias is strongest in highly expressed genes (e.g. Gouy and Gautier 1982; Duret and Mouchiroud 1999). We therefore identified optimal codons as those which differed in their usage between high versus low expression genes using the method of Duret and Mouchiroud (1999). We first calculated the relative synonymous codon usage (RSCU) for each codon in all genes using the program codonW (J. Peden http://codonw.sourceforge.net). From this we contrasted RCSU in high expression (top 10th percentile, 185,546 codons) versus low expression (bottom 10th percentile, 260,695 codons) as $\Delta$RSCU = RSCU$_{\text{high}}$ - RSCU$_{\text{low}}$ (Duret and Mouchiroud 1999). We measured expression as the number of sequence reads per kilobase per million reads mapped (FPKM, Trapnell et al. 2010). Statistical departure of $\Delta$RSCU from zero was tested for each codon using a one-way analysis of variance (ANOVA) conducted in JMP v8.0 and codons with significantly positive values were inferred to be putative optimal codons. We calculated $F_{op}$ in codonW with a customized optimal codons table. We repeated this procedure for each of the four genotypes to calculate genotype-specific tables. We also defined consensus optimal codons as those that were optimal in all four genotypes and limited this set to include only one codon per amino acid. All analyses presented here are based on the consensus optimal codons, but the results were qualitatively similar using genotype specific estimates.

Both selection and the background base composition can influence estimates of $F_{op}$. We assessed the role of these factors, genotype and mating system to determine their influence on codon usage bias in our four samples. We constructed ANOVA models in JMP v8.0 where $F_{op}$ was a function of base composition (measured as GC content at silent sites, GC3s), gene length (bp), gene expression (FPKM), genotype and mating system (represented as a categorical variable; selfing or outcrossing). Genotype and mating system could not be
included simultaneously because there is only one outcrossing genotype (zero degrees of freedom) and they were therefore run in two separate models. Our models initially assessed the effects of all of these factors and their interactions. Terms were excluded from the model by backward elimination ($\alpha = 0.05$) if they did not explain a significant proportion of the variation in codon bias, and they were not involved in any significant higher order interactions. Although GC3s was found to be highly significant, we also present the results of an ANOVA that excluded GC3s because we found that all of our putatively optimal codons have guanine or cytosine in the third position, therefore GC3s was not independent of $F_{op}$.

The statistic $\Delta \overline{RSCU}_+$ proposed by Cutter et al. (2006b), measures the strength of selection for codon bias as the mean of all positive values of $\Delta RSCU$ for a given genotype. Because RSCU is independent of the amino acid content and $\Delta RSCU$ controls for base composition, $\Delta \overline{RSCU}_+$ provides a useful metric for comparing the degree of codon usage bias across genotypes and species. We calculated $\Delta \overline{RSCU}_+$ for each genotype and tested for significant effects of mating system and genotype using two separate one-way ANOVAs. We also compared the $\Delta \overline{RSCU}_+$ values for mating system (selfing versus outcrossing) and genotype using the Tukey-Kramer HSD. All values of $\Delta \overline{RSCU}_+$ were log transformed for these statistical tests so they fitted the normal distribution.

*Phylogenetic inference of selection on codon bias*

Codonbias (Nielsen et al. 2007) is an extension of PAML with a likelihood method for estimating codon bias parameters in multiple genotypes in a phylogenetic framework. They extend these methods by adding a parameter representing the selection coefficient ($S = N_e s$) against unpreferred codon usage, which is simultaneously optimized with mutation rates
(α) and \( \frac{d_N}{d_S} \) (ω) along each branch of the phylogenetic tree. To assess whether evidence exists for significant selection on codon bias in our four genotypes we constructed four models. (1) ‘No selection’ - selection on codon bias was constrained to \( S = 0 \), or no selection. This served as a null model to test subsequent selection models against. (2) ‘Equal selection’ - selection is estimated, but constrained to be constant across the tree. (3) ‘Full model’ - selection against unpreferred codons was allowed to vary freely across every branch of the tree. (4) ‘Relaxed selection model’ - one selfing \( S \) was estimated for all three terminal selfing branches and a second outcrossing \( S \) along the branches from the root to the Brazilian outcrossing genotype of \( E. paniculata \). For each of these models we also estimated α and ω but after initial trials, and based on the results of our PAML analyses, we fixed a single estimate for each parameter to reduce the number of parameters being jointly optimized. We also conducted each of these analyses on a restricted set of high expression genes (top 10th percentile) because we reasoned that if selection is strongest in these genes it is the most likely set of candidate genes to display a signal of relaxed selection on codon usage. To compare among all nested models (those with the same set of genes) we applied a likelihood ratio test under a \( \chi^2 \) approximation to determine whether increasingly complex models fitted our data significantly better.

Results

Our dataset for the four \( Eichhornia \) genotypes consisted of 7914 loci (average length = 627.5 bp) totalling 3,563,891 bp of aligned sequence. There were 21,944 polymorphic sites within \( E. paniculata \) and 105,964 divergent sites between \( E. paniculata \) and \( E. paradoxa \) (see Table 6.1). Mean non-synonymous polymorphism (\( \theta_w = 0.0024\% \)) was an order of magnitude lower than mean synonymous polymorphism (\( \theta_w = 0.016\% \)).
Selection on non-synonymous mutations

Using the set of concatenated genes (Table 6.2) there was strong statistical support for the “full model” in which each branch had a different $\omega$ value ($\chi^2 = 39.95, P < 0.001$). Estimated values did not provide overall support for an elevated $\omega$ in selfing lineages. Only on the Nicaraguan branch did we observe a slightly higher $\omega$ than on the outcrossing branch, consistent with weak relaxation of selection in this lineage. Assuming *E. paradoxa* represented the ancestral state, we calculated the number of derived synonymous and non-synonymous mutations in *E. paniculata* (Figure 6.1). The $\chi^2$ test of homogeneity was highly significant ($P < 0.001$) with a slightly higher ratio of non-synonymous to synonymous derived mutations in the selfing genotypes of Nicaragua (0.484) and Jamaica (0.437) compared to the outcrosser from Brazil (0.422). Although the magnitude of these differences is not very large the trend suggests a slight elevation of $\omega$ in both *E. paniculata* selfers.

In contrast to the analyses on concatenated genes, most genes (6871 out of the 7881 for which all four models could be run) supported a model with a single $\omega$. This probably reflects insufficient power to accurately estimate the parameters of more complex models with so few variable sites within *E. paniculata*. For more complex models, the distribution of $\omega$ values was strongly bimodal (values much greater than one or nearly zero, data not shown) indicating that for many loci the estimate of $\omega$ is unreliable (see discussion). Furthermore, evidence for $\omega$ values >1 conflicts with the MK tests presented below which show no evidence for positive selection. As a result, we do not discuss the results of analyses of individual loci.
MK tests of selection

We conducted MK test on two sets of genes: 1) where only MK tables with non-zero entries occurred, and 2) where only MK tables for which the marginal sums were greater than or equal to five occurred. The first more liberal set comprised 2407 genes, 65 of which were significantly different from neutral evolution; 11 genes supported positive selection with an excess of non-synonymous divergence and 54 with an excess of non-synonymous polymorphism, supporting purifying selection. The second more conservative set comprised 1046 genes, 50 of which were significant; 10 and 40 supporting positive and purifying selection, respectively. After statistical correction for multiple testing using the QVALUE package in R (Storey and Tibshirani 2003), none of the loci in either set were significantly different from expectations based on neutral evolution. It is notable that the distribution of $P$-values is highly deviant from the expected uniform distribution (under the null assumption of no selection) with many $P$-values of 1. Summing over all genes there was a highly significant signal ($P=4.10^{-31}$ and $P=5.10^{-5}$, respectively, for the liberal and conservative sets) consistent with the action of purifying selection ($P_N/P_S > D_N/D_S$). This pattern is frequently observed when genes are pooled (Charlesworth and Eyre-Walker 2006; Charlesworth and Charlesworth 2010).

We also conducted multi-locus MK tests with MK-Test 2.0 (Welch 2006). Table 6.3 presents the results of the four models we fitted for the second more conservative set of loci (the results are qualitatively similar for first set and are not shown). The two models that estimated a single value of $f$ (the proportion of new mutations that are neutral) for all loci were better supported and had similar AICc values (AICc = 27137.98 for the model with only $f$ being estimated). The model that fitted a non-zero proportion of adaptive substitutions was marginally better (AICc = 27123.75) and estimated a negative value of $\alpha$ ($\alpha = -0.1039$).
Under the neutral theory, $1-f$ is a measure of constraint on proteins. Therefore, results from these multi-locus MK tests provide evidence for the action of purifying selection ($1-f \sim 0.8$) across the genome.

**Assessing selection for codon bias**

Codon usage across all four genotypes was highly non-random. Summary statistics for high and low expression genes as well as all loci combined are presented in Table 6.4. GC content at 4-fold degenerate sites in all genes (GC3s) was similar across all samples ($\text{GC3s}_{E.\text{paniculata}} = 0.480$, $\text{GC3s}_{E.\text{paradoxa}} = 0.483$) but was significantly different in high versus low expression genes ($\text{GC3s}_{\text{high}} = 0.583$, $\text{GC3s}_{\text{low}} = 0.465$, Tukey-Kramer HSD $q^* = 2.34 P < 0.0001$). As a result, we found that values of RSCU for each codon in each genotype differed between high and low expression genes. Using $\Delta$RSCU analysis we identified 24 putative optimal codons (those with significantly positive $\Delta$RSCU values, Figure 6.2) that were found in all four genotypes representing 18 different amino acids. All 24 of these putatively optimal codons had guanine or cytosine in the third position. Each of the degenerate amino acids was represented by at least one putatively optimal codon. The six amino acids for which we have more than one optimal codon include alanine, leucine, proline, serine, threonine and valine. In each of these amino acids with two preferred codons, the codon with a greater $\Delta$RSCU value terminated with a cytosine rather than a guanine. This pattern occurs despite the higher proportion of guanine in synonymous third positions sites across all genes.

We used these 18 putatively optimal codons to calculate codon bias in each gene of each genotype. We found that GC3s had a highly significant effect on measures of codon bias ($F_{\text{op}}$) ($F_{6,31813} = 370016.2$, $P < 0.0001$) explaining 88.9% of the variance in $F_{\text{op}}$. In addition, both length and gene expression had significant effects on $F_{\text{op}}$, but only explained
0.011% and 0.008% of the variance in $F_{op}$, respectively. There was no significant effect of either mating system or genotype on the frequency of optimal codons. However, there was strong differentiation in both GC3s and $F_{op}$ in high versus low expression genes, demonstrating that selection for codon bias is stronger in up-regulated genes.

To compare bias in codon usage in each genotype we also used the statistic $\Delta \text{RSCU}^+$ (Figure 6.3). Although no difference was detected in $F_{op}$ among genotypes, there was a significant effect of mating system on $\Delta \text{RSCU}^+$ ($F_{1,94} = 5.78$, $P < 0.05$). The mean for all three selfers combined ($\Delta \text{RSCU}^+ \text{Selfers} = 0.24$) was significantly lower than for the Brazilian outcrosser ($\Delta \text{RSCU}^+ \text{Brazil} = 0.29$) ($t = 1.99$, $P < 0.01$). This pattern may have been largely driven by the significant effect of genotype ($F_{3,92} = 3.02$, $P < 0.05$) due to the lower $\Delta \text{RSCU}^+ \text{Nicaragua} (0.22)$ relative to the corresponding value of $\Delta \text{RSCU}^+ \text{Brazil} = 0.29$ from Brazil (Figure 6.3. Tukey-Kramer HSD $q^* = 2.62$, $P < 0.01$).

**Phylogenetic inference of selection on codon bias**

We estimated the selection coefficient ($S = N_e s$) for optimal codon usage of the four genotypes across the tree under a number of different evolutionary scenarios (see table 6.5). The hypothesis of ‘no selection’ was rejected with a high degree of certainty ($\chi^2 = 1702.0$, 1 DF, $P < 0.0001$) when compared to the more complex model of ‘equal selection’, where all lineages experience the same selection against unpreferred synonymous mutations ($S = 0.082$). However, the ‘full model’, in which each branch of the tree had an independently estimated selection coefficient, explained the data significantly better than either the ‘equal selection’ or ‘relaxed selection’ models ($\chi^2 = 5564.5$, 52 DF, $P < 0.0001$). In the ‘full model’, selection for codon usage bias was over 70% higher in the outcrossing genotype from Brazil.
(\(S_{\text{Brazil}} = 0.034\)) compared with the Nicaraguan selfing genotype (\(S_{\text{Nicaragua}} = 0.020\)) and substantially higher than selection in the Jamaica lineage (\(S_{\text{Jamaica}} = -0.01\)) (see Figure 6.4). However, selection in \(E.\) paradoxa was highest by an order of magnitude (\(S_{E.\) paradoxa} = 0.434).

We also ran the same four models using only loci from the top 10\(^{th}\) percentile of gene expression (Table 6.5). High expression genes are expected to be under the strongest selective pressure for optimal codon usage. In this category of genes, we again found strong evidence for biased codon usage in the ‘equal selection’ model (\(S = 0.431\)), which fitted significantly better than the ‘no selection’ model (\(\chi^2 = 3221.5, 1\) DF, \(P < 0.0001\)). The strength of selection estimated across the tree, using only one selection coefficient was approximately five-fold larger in the highly expressed genes than in the full set of genes. Similar to above, the ‘full model’ fitted significantly better than both the ‘equal selection’ and ‘relaxed selection’ models (\(\chi^2 = 8200.5, 52\) DF, \(P < 0.0001\)). The strength of selection inferred was highest in the outcrossing genotype (\(S_{\text{Brazil}} = 0.602\)), and lowest in the Nicaraguan and Jamaican selfing genotypes of \(E.\) paniculata (\(S_{\text{Nicaragua}} = 0.013, S_{\text{Jamaica}} = 0.015\)) (Figure 6.4). Unlike the full set of genes, the strength of selection in \(E.\) paradoxa (\(S_{E.\) paradoxa} = 0.192) was substantially lower than the outcrosser but still substantially higher than either of the \(E.\) paniculata selfers.

**Discussion**

Large numbers of loci are required to detect significant shifts in patterns of selection across the genome accompanying the transition from outcrossing to selfing. This is largely because the mutational process is inherently noisy and many selfing populations are recently derived. As a result, few studies have detected a signature of the effects of a transition to
selfing on the efficacy of natural selection ($N_s$). In this study, we sequenced ~8000 genes from three independently derived selfing lineages and one outcrosser from *Eichhornia*, a plant genus illustrating multiple independent shifts from outcrossing to selfing. Below we discuss the implications of our results in the context of selective constraints on protein sequences and the underlying codon usage during mating system transitions.

Our study was based on a limited number of genotypes, in part because of the large number of genes that we were interested in sampling. As a result, accurate estimates of polymorphism were not possible and this limited the analyses that we could conduct. However, our sampling scheme involving many loci provides an accurate reflection of the deeper coalescent history of each population because by definition the coalescent histories of unlinked loci are independent. Therefore, the specific identity of individuals sampled in each population is of less importance. In the case of the outcrossing genotype, this was sampled from a region of Brazil dominated by outcrossing populations, with little evidence of population structure and higher levels of gene flow than occur elsewhere in the range, where selfing populations occur more commonly (Husband and Barrett 1992a; Chapter 4). As a result, it seems reasonable to assume, given that lack of any obvious population subdivision, that the loci sequenced in this individual provide a reasonable approximation of the larger metapopulation in this region, and thus are representative of outcrossing genotypes in general.

*Mating system and protein evolution*

The results from PAML analyses failed to detect a clear genomic signature from a reduction of selection efficacy in selfing lineages at the protein level. Several factors may explain the lack of signal (see below), but it should be noted that regardless of the model
studied this pattern has proved surprisingly elusive and difficult to demonstrate (Cutter et al. 2006b; Escobar et al. 2010). Our analyses should be interpreted with caution since an important assumption underlying the $d_N/d_s$ approach to detect selection was likely violated, namely that differences among the lineages compared represent independent substitutions (see Kryazhimskiy and Plotkin 2008). Specifically, the relation between selection and $\omega$ is different if mutations do not represent independent fixation events along the lineages but instead are segregating polymorphisms within a species. Although this problem is particularly important for assessing evidence for positive selection, it may be less acute if one is mostly concerned with deleterious mutations. Indeed, the shape of the $d_N/d_s$ ratio curve as a function of the scaled selection coefficient for within-species samples and truly divergent lineages agree quite well for deleterious mutations whereas they differ widely for those that are advantageous (see Figure 1 in Kryazhimskiy and Plotkin 2008). The extent to which this remains true under a wide range of recombination values requires further investigation. In our study we specifically aimed at testing for an increase of the $d_N/d_s$ ratio in selfing lineages as a consequence of a reduction of $N_e$ causing nearly-neutral slightly deleterious mutations to rise to fixation.

We detected a highly significant, albeit slight, increase in the proportion of non-synonymous changes in the two selfing lineages of *E. paniculata* in comparison with the outcrossing genotype (see Figure 6.1). Although this finding is consistent with relaxed selection in the selfing genotypes, it obviously cannot be considered definitive because only a single outcrossing genotype was represented in our study. Among the ~8000 loci, the Nicaraguan and Jamaican genotypes had 526 and 248 more non-synonymous mutations, respectively, than were inferred in the outcrossing genotype. Interpretation of these values

must take into account that not all the changes observed are necessarily fixed within the respective populations. A recent study of polymorphism in a derived selfer, *Capsella rubella*, detected an excess of rare non-synonymous mutations (SI Wright, unpublished data). If a similar pattern occurs in the Jamaican and Nicaraguan genotypes, some portion of the observed non-synonymous mutations could be due to polymorphism. Future sampling of more individuals from both selfing and outcrossing populations could allow evaluation of the generality of our findings and permit a more in depth understanding of selection on the protein sequence.

**MK tests and evidence for purifying selection**

Single- and multi-locus MK tests provided evidence for purifying selection across all loci but did not support positive selection for any loci. The MK test is not sensitive to deviations from neutrality in single genes and has relatively low power when applied to individual genes. However, when we pooled across many genes we detected a significant signal of purifying selection. Under the assumptions of the standard neutral theory, $1-f$ represents a quantification of the level of constraint experienced at the protein level. In our case, this would imply that ~80% of the sites were under purifying selection. A similar pattern has been observed in *Arabidopsis thaliana* and *A. lyrata* (see Foxe et al. 2008; Slotte et al. 2010). Interestingly, like *E. paniculata*, both species have been shown to have relatively small $N_e$, and this has been suggested as the cause of reduced purifying selection and the occurrence of adaptive substitutions (Wright and Andolfatto 2008).

We found no support for adaptive evolution at the genomic level in *E. paniculata*, a result consistent with results for most plant species studied to date (reviewed in Siol et al. 2010, but see Ingvarsson 2008; Slotte et al. 2010). The MK test is mostly robust to
demographic influences on selection models because the synonymous and non-synonymous sites are interspersed throughout the genome and should be affected more or less equivalently by bottlenecks and other stochastic forces. However, a number of factors could bias inferences (Eyre-Walker 2002; Sella et al. 2009); for example, a fraction of non-synonymous mutations might be weakly rather than strongly selected, leading to an under-estimation of both $f$ and $\alpha$. This problem is usually dealt with by removing rare polymorphisms from the analysis. However, because in our case polymorphism estimates are based on only three sequences it is likely that such rare polymorphisms have not been sampled. Thus, the main limitation in our application of the MK test is that estimation of polymorphism lacks power and this may partly explain why we failed to detect positive selection on any locus.

*Mating system and selection for codon bias*

Among the four genotypes of *Eichhornia* included in this study we found evidence for bias in codon usage. Measures of both codon bias ($F_{op}$) and GC3s distinguished high versus low expression genes. Moreover, gene expression explained a small but highly significant fraction of the variance in $F_{op}$. This result can be interpreted as the result of selection causing codon usage bias. However, the small fraction of variance explained by differences in expression likely underestimates the role of selection, because of the very high correlation of GC3s with $F_{op}$. The strong differentiation of GC3s reflected in $\Delta$RSCU values (Figure 6.2) also provides evidence for the role of selection in shaping patterns of codon usage in high and low expression genes. Our estimates of $\Delta$RSCU + for the four genotypes (0.22 - 0.29) are comparable with those reported from other species including *Caenorhabditis* spp. (0.335- 0.498, Cutter et al. 2008), *Drosophila melanogaster* (0.28, Duret and
Mouchiroud 1999), *A. thaliana* (0.140, Duret and Mouchiroud 1999) and *Populus* spp. (0.097-0.193, Ingvarsson 2008).

Our analyses of $F_{op}$ provided no evidence for an effect of selfing on codon usage. Neither mating system nor genotype explained a significant fraction of the variation in $F_{op}$ in our ANOVA models. This is likely due, in part, to the very large proportion of the variance in $F_{op}$ explained by synonymous GC content. GC3s explained approximately 89% of the variance in the frequency of optimal codons. This is not surprising given that all 18 optimal codons have either a guanine (G) or cytosine (C) in 3rd positions, causing these measures to be highly correlated. It is common for optimal codons in other species to be dominated by GC ending codons (e.g. in *Caenorhabditis* 17 of 19 optimal codons identified ended with GC; Cutter et al. 2006b). Differences in base composition may aid in explaining why $F_{op}$ is slightly, although not significantly, higher in *E. paradoxa*, which had marginally higher GC content at synonymous sites than the three genotypes of *E. paniculata*. However, this measure estimates the bias in codon usage in all genes and if only the highly expressed genes experience selection for codon bias it may not detect subtle changes in selection on synonymous sites.

The degree of bias in codon usage in high versus low expression genes was partially explained by genotype and mating system (Figure 6.3). The measure $\Delta \text{RSCU} +$ (Cutter et al. 2006b) provides a useful way to quantify codon usage bias across species or genotypes. In addition to being useful for comparing genotypes with differing base composition, this measure also quantifies the differentiation in codon usage between high and low expression genes. As discussed above, we found a striking difference in the GC3s of high versus low expression genes. However, the four genotypes varied with respect to the degree of this difference. Although *E. paradoxa* had the highest GC3s in high expression genes, it also had
the largest GC3s in low expression genes. This pattern was reflected in values of $\Delta \overline{RSCU}$, where the outcrossing *E. paniculata* genotype from Brazil had a significantly higher $\Delta \overline{RSCU}$ estimate than the three selfing genotypes combined. Moreover, while there was no significant difference between the outcrosser and the selfing genotype from Jamaica and *E. paradoxa*, there was a significant difference between the outcrosser and the Nicaraguan genotype. This pattern is illustrated in Figure 6.2 in which the Nicaraguan genotype appears less polarized in $\Delta RSCU$ values between optimal and sub-optimal codons. The reduction in codon bias in the Nicaraguan population is in line with the finding of small changes in the ratio of non-synonymous to synonymous mutations discussed above.

These differences in codon usage are supported by maximum likelihood estimates of selection on optimal codon usage ($S$). Using the method of Nielsen et al. (2007), we found evidence for a reduction in $S$ in all selfing lineages. There was a clear signal of significant selection on codon usage in both analyses (high expression and all loci). This was indicated by the rejection of the ‘no selection’ model in both datasets. Using all loci, we found that the Brazilian outcrosser had a higher estimate of $S$ ($S_{Brazil} = 0.034$) than either the Jamaican ($S_{Jamaica} = -0.01$) or Nicaraguan ($S_{Nicaragua} = 0.020$) genotypes. This result is consistent with a reduced efficacy of selection for purging of mutations that cause unpreferred codons due to reductions in $N_e$. However, in the same analysis the predominantly selfing *E. paradoxa* has substantially stronger selection for codon bias ($S_{E. paradoxa} = 0.434$), and this can probably be explained by the elevated GC content found in this species. This analysis uses the optimal codons inferred in the $\Delta RSCU$ analysis, which all end with G or C. Therefore, strong selection would be required, from the time of the split between *E. paradoxa* and *E. paniculata* and their common ancestor, to explain this difference in GC content by selection for codon bias alone. However a number of other processes such as biased gene conversion
could result in the same pattern (Marais 2003), and we do not know the base composition of
the common ancestor. As expected, the analysis of high expression genes revealed stronger
selection for this category of genes. Moreover, it provided even stronger support for relaxed
selection on codon bias in the three selfers (e.g. $S_{\text{Brazil}} = 0.602$ compared with the selfing
genotypes $S_{\text{Nicaragua}} = 0.013$, $S_{\text{Jamaica}} = 0.015$ and $S_{E.\text{ paradoxa}} = 0.192$. This result was consistent
with those for $\Delta \text{RSCU}$, demonstrating that relaxation of selection is most evident in the
degree of differentiation of codon bias between high versus low expression genes.

The point estimates of $S$ for each genotype in the best fitting model for each analysis
were small and should be viewed with caution. In theory, selection is expected to be over-
powered by drift and the fate of mutations are effectively neutral when $N_eS < 1$ (Kimura
1964; Kimura and Ohta 1971). In all of our estimates the value of $N_eS$ is very small and may
be expected to have no effect. However, as with PAML (Yang 2007), the program codonbias
(Nielsen et al. 2007) assumes that differences amongst samples represent fixed substitutions
and including sites which are polymorphic may cause underestimation of $S$. This is because
recent slightly deleterious mutations not yet purged by selection may still be segregating in
populations. However, this effect should diminish the difference between outcrossing and
selfing lineages and therefore our estimates can be considered conservative. We have no
reason to expect more mutations to be segregating in the selfing populations sampled here,
including the Jamaican population, which has low levels of nucleotide diversity.
Furthermore, the analysis of multiple nuclear loci presented in Chapter 4 indicates that
contemporary gene flow from Brazil to the Caribbean is unlikely. Thus, our estimates seem
likely to reflect the relative levels of selection on codon bias along each of the lineages
examined here.
Factors influencing the efficacy of selection in selfers

One explanation for the stronger effect of selfing on codon bias relative to protein evolution is that the underlying distributions of fitness effects (DFE) of synonymous and non-synonymous mutations are likely to be very different. Most models that predict an increase in the accumulation of deleterious mutations in inbreeders rely on a large class of slightly deleterious mutations with additive fitness effects (Wang et al. 1999; Glémin 2007). However, if a large fraction of non-synonymous mutations are strongly deleterious so that $N_e s$ remains greater than one in selfing populations, the rate of fixation may not be elevated as predicted. Therefore, if the shape of the DFE for synonymous sites differs, such that a large portion of the mutations altering codon usage are very weakly selected, then only a small decline in $N_e$ due to selfing should result in $N_e s < 1$. Weak selection on codon bias seems likely given the relatively low estimates of $N_e s$ in other species using the same method (e.g. *Drosophila* spp. $N_e s = -0.37 – 1.74$, Nielsen et al. 2007; *Populus* spp, $N_e s = -0.232 – 0.720$, Ingvarsson 2008). These results are consistent with the weak evidence we found for the accumulation of non-synonymous mutations and the subtle reductions in selection for codon bias evident in our data.

Models which predict an accumulation of deleterious mutations in selfing populations generally require that mutations are additive in their effects (Glémin et al. 2006). However, if a sizable fraction of slightly deleterious mutations are recessive they will be more effectively eliminated in highly selfing, homozygous populations where they are exposed to selection, a process known as purging (Crnokrak and Barrett 2002; Schoen and Busch 2008). Barrett and Charlesworth (1991) reported differences in the intensity of inbreeding depression between selfing Jamaican and outcrossing Brazilian populations of *E. paniculata* after five generation of controlled selfing and crossing. They interpreted their results as being consistent with the
occurrence of natural purging of genetic load in the Jamaican population owing to a history of selfing. We may therefore expect that some mutations identified in our outcrossing genotype are polymorphic and that sheltered genetic load may have been purged during the colonization bottleneck associated with long-distance dispersal to the Caribbean by *E. paniculata*. If true this could lessen the signal of deleterious mutation accumulation in selfers. However, while it seems reasonable to assume that deleterious recessives make a substantial contribution to inbreeding depression in *E. paniculata*, we cannot quantify the impact of these mutations on our estimates of selection with the current data.

A number of other factors may further limit accumulation of deleterious mutations in selfers relative to outcrossers. Selfing rates vary considerably among Jamaican populations (e.g. Barrett et al. 1992) and even low levels of outcrossing are sufficient to allow recombination to lessen many of the predicted consequences of extreme inbreeding (e.g. Charlesworth and Charlesworth 2010; Platt et al. 2010). We have no estimate of the selfing rate or amount of nucleotide diversity in the Nicaraguan population from which our sample originated. However, it is possible that the significantly reduced value of $\Delta RSCU$ compared to the outcrosser, is a result of a more severe reduction of $N_e$ in this population than in the Jamaican population, which occurs on an island with many *E. paniculata* populations (see Figure 3 in Barrett et al. 1989). In contrast, only a few isolated populations of *E. paniculata* are reported from Nicaragua and it is possible that a much stronger bottleneck has accompanied colonization of Central America. Another explanation for the limited accumulation of deleterious mutations in *E. paniculata* is that the transition from outcrossing to selfing may have occurred relatively recently. In Chapter 4 we estimated that the colonization of Jamaica from Brazil occurred ~125,000 generations ago. Therefore, there may not have been sufficient time to reach equilibrium and for ancestral polymorphism to
coalesce within selfing populations. Insufficient time to detect the genomic effect of selfing has been raised by other authors in selfing-outcrossing comparisons (Wright et al. 2002; Cutter et al. 2008) and this may generally be true of selfing lineages which are often recently derived and short lived (reviewed in Takebayashi and Morrell 2001; Igie et al. 2008; Goldberg et al. 2010). Population genomic data from the selfing plant *Capsella rubella* indicates an increase in number of low frequency non-synonymous polymorphism relative to its outcrossing progenitor, *C. grandiflora* (SI Wright, unpublished data). Estimates from sequence data for this species pair imply a very recent origin of selfing (~20,000 years, Foxe et al. 2009). In recently derived selfers the signal of relaxed selection at non-synonymous sites across the genome may only be detectable with more extensive polymorphism data.
Table 6.1. Summary statistics from the dataset on synonymous and non-synonymous sites across all loci of *Eichhornia*. We display the number of aligned sites ($N_{sites}$), the number of polymorphic sites ($S$), the number of divergent sites ($K$) and Watterson’s polymorphism estimator ($\theta_w$).

<table>
<thead>
<tr>
<th>Site class</th>
<th>$N_{sites}$</th>
<th>$S$</th>
<th>$K$</th>
<th>$\theta_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>805936</td>
<td>14682</td>
<td>72383</td>
<td>0.0159 (0.0176)</td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>2757955</td>
<td>7262</td>
<td>33581</td>
<td>0.0024 (0.0042)</td>
</tr>
</tbody>
</table>
Table 6.2. Results from PAML tests of reduced selection efficacy in selfers of *Eichhornia* at the protein level with the genes concatenated. We have included the transition/transversion ratio (κ), the log-likelihood of the model (Ln(L)), the number of parameters of the model (np), the likelihood ratio statistic (LR) of each model with the next less complex model immediately above (e.g. the “three ω” model is tested by comparing the log-likelihood with the log-likelihood of the “relaxed selection” model), (**P < .0001**).

<table>
<thead>
<tr>
<th>Model</th>
<th>κ</th>
<th>Ln(L)</th>
<th>np</th>
<th>LR</th>
<th>ω</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single ω</td>
<td>2.449</td>
<td>-7711152.98</td>
<td>7</td>
<td>-</td>
<td>All branches: 0.174</td>
</tr>
<tr>
<td>Relaxed selection</td>
<td>2.449</td>
<td>-7711051.42</td>
<td>8</td>
<td>203.11***</td>
<td>Selfers: 0.170</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Outcrosser: 0.227</td>
</tr>
<tr>
<td>Three ω</td>
<td>2.450</td>
<td>-7710793.45</td>
<td>9</td>
<td>515.95***</td>
<td><em>E. paradoxa</em>: 0.163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>E. paniculata</em> selfers: 0.236</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Outcrosser: 0.218</td>
</tr>
<tr>
<td>Full model</td>
<td>2.450</td>
<td>-7710773.47</td>
<td>10</td>
<td>39.95***</td>
<td><em>E. paradoxa</em>: 0.163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Jamaica: 0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Nicaragua: 0.257</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Outcrosser: 0.218</td>
</tr>
</tbody>
</table>
Table 6.3. Results of the multi-locus MK tests for different models with or without purifying and positive selection on genotypes of *Eichhornia*. We have included the log-likelihood of each model ( Ln(L) ), number of parameters estimated ( np ), neutral diversity parameter estimated ( θ ) from the data, expected neutral divergence ( μt ) estimated from the data, fraction of new mutations that are neutral ( f ), fraction of adaptive nucleotide substitutions, ( α ) and the “second order” Aikaike criterion (AICc) (lower values indicate more support). Models are as follow, f=θ, α = 0 (no purifying selection, no adaptive substitutions); f= free, α = 0 (purifying selection, with a single value of f estimated from all loci, no adaptive substitutions); f= 0, α = free (no purifying selection, a single value of α estimated from all loci); f= free, α = free (purifying and adaptive substitutions allowed with one value estimated for both parameters for all loci).

<table>
<thead>
<tr>
<th>Model</th>
<th>Ln(L)</th>
<th>np</th>
<th>θ</th>
<th>μt</th>
<th>f</th>
<th>α</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>f= 0</td>
<td>-27180.05</td>
<td>2</td>
<td>0.0078</td>
<td>0.0225</td>
<td>-</td>
<td>-</td>
<td>54364.09</td>
</tr>
<tr>
<td>α = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f= free</td>
<td>-13565.99</td>
<td>3</td>
<td>0.0217</td>
<td>0.0625</td>
<td>0.1717</td>
<td>-</td>
<td>27137.98</td>
</tr>
<tr>
<td>α = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f= 0</td>
<td>-16850.01</td>
<td>3</td>
<td>0.0078</td>
<td>0.0819</td>
<td>-</td>
<td>-4.9774</td>
<td>33706.03</td>
</tr>
<tr>
<td>α = free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f= free</td>
<td>-13557.87</td>
<td>4</td>
<td>0.0211</td>
<td>0.0642</td>
<td>0.1847</td>
<td>-0.1039</td>
<td>27123.75</td>
</tr>
<tr>
<td>α = free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.4. Mean values of expression, gene length, frequency of optimal codon usage ($F_{op}$) and base composition, measured as GC-content at synonymous sites for genes (GC3s) sampled in genotypes of *Eichhornia*. For each of the four genotypes, the genes were divided into two categories based on expression level including the top (‘high’) and bottom (‘low’) and 10th percentile of expression. The value for all loci is summarized in the category, ‘total’.

<table>
<thead>
<tr>
<th>Expression category</th>
<th>Genotype</th>
<th>Expression (FPKM)</th>
<th>Length (bp)</th>
<th>$F_{op}$</th>
<th>GC3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td><em>E. paniculata</em>, Brazil</td>
<td>166.1</td>
<td>698.4</td>
<td>0.424</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Jamaica</td>
<td>155.8</td>
<td>698.4</td>
<td>0.424</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Nicaragua</td>
<td>162.2</td>
<td>698.4</td>
<td>0.424</td>
<td>0.581</td>
</tr>
<tr>
<td></td>
<td><em>E. paradoxa</em></td>
<td>154.8</td>
<td>698.4</td>
<td>0.427</td>
<td>0.586</td>
</tr>
<tr>
<td>low</td>
<td><em>E. paniculata</em>, Brazil</td>
<td>3.4</td>
<td>981.3</td>
<td>0.333</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Jamaica</td>
<td>3.6</td>
<td>981.3</td>
<td>0.334</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Nicaragua</td>
<td>3.7</td>
<td>981.3</td>
<td>0.334</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td><em>E. paradoxa</em></td>
<td>5.8</td>
<td>981.3</td>
<td>0.335</td>
<td>0.467</td>
</tr>
<tr>
<td>total</td>
<td><em>E. paniculata</em>, Brazil</td>
<td>43.2</td>
<td>628.0</td>
<td>0.339</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Jamaica</td>
<td>42.1</td>
<td>628.0</td>
<td>0.338</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td><em>E. paniculata</em>, Nicaragua</td>
<td>43.2</td>
<td>628.0</td>
<td>0.338</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td><em>E. paradoxa</em></td>
<td>43.0</td>
<td>628.0</td>
<td>0.341</td>
<td>0.483</td>
</tr>
</tbody>
</table>
Table 6.5. Maximum-likelihood estimates of selection coefficients on codon usage bias in *Eichhornia* genotypes from the program codonbias (Nielsen et al. 2007). The table includes the number of parameters and Ln(likelihood) score for each of eight models. We ran two separate datasets, one “all loci” included all 7914 loci from each individual concatenated, and the other, “high expression” included only the 10th percentile of high expression genes. We also include the number of parameters (np) for each model and the selection coefficient $S (= N_e s)$ for codon usage bias for each of the four genotypes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>np</th>
<th>Ln(L)</th>
<th>$S_{Brazil}$</th>
<th>$S_{Nicaragua}$</th>
<th>$S_{Jamaica}$</th>
<th>$S_{E. paradoxa}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>all loci</td>
<td>no selection</td>
<td>18</td>
<td>-3977428.1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>equal selection</td>
<td>19</td>
<td>-3927502.2</td>
<td>0.082</td>
<td>0.082</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>relaxed selection</td>
<td>32</td>
<td>-3926651.2</td>
<td>0.230</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>full model</td>
<td>84</td>
<td>-3923868.9</td>
<td>0.034</td>
<td>0.020</td>
<td>-0.010</td>
<td>0.434</td>
</tr>
<tr>
<td>high expression</td>
<td>no selection</td>
<td>18</td>
<td>-499114.0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>equal selection</td>
<td>19</td>
<td>-496624.4</td>
<td>0.431</td>
<td>0.431</td>
<td>0.431</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>relaxed selection</td>
<td>32</td>
<td>-495575.0</td>
<td>1.088</td>
<td>0.139</td>
<td>0.139</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>full model</td>
<td>84</td>
<td>-495013.7</td>
<td>0.602</td>
<td>0.013</td>
<td>0.015</td>
<td>0.192</td>
</tr>
</tbody>
</table>
Figure 6.1. Counts of synonymous (S) and non-synonymous (NS) derived mutation in selfing and outcrossing genotypes of *Eichhornia*. $P_N/P_S$ represents the ratio of the number of non-synonymous to synonymous derived mutations. The ancestral state is considered to be the nucleotide observed at each position in *E. paradoxa*. The colours along the branches reflect the putative model of changes in mating system. The branch lengths are not to scale.
Figure 6.2. Heat map of $\Delta$RSCU values for four genotypes of *Eichhornia* included in this study generated with the software CIMMiner (http://discover.nci.nih.gov/cimminer). Each column represents a single genotype. Each row corresponds to one of the 61 codons (excluding stop codons), where the value of $\Delta$RSCU for each genotype is indicated. Codons are ordered by the mean value of $\Delta$RSCU across all genotypes. Optimal codons identified as part of the consensus set of codons are marked with an asterisk.
Figure 6.3. Comparisons of codon bias, measured as $\Delta RSCU^+$, among each of the four *Eichhornia* genotypes, and the three selfing genotypes combined. $\Delta RSCU^+$ is the mean of all positive $\Delta RSCU$ values for a given genotype. The results of two separate tests of significance are marked with A and B for the difference in mean $\Delta RSCU^+$ between the outcrosser and the selfers combined ($t = 1.99, P < 0.01$). Bars indicate significant differences among individual genotypes using a Tukey-Kramer HSD ($q^* = 2.62, P < 0.01$) with error bars indicating the standard error of each estimate.
Figure 6.4. Schematic representation of estimates of the selection coefficient, $S$, for preferred codon usage in *Eichhornia* genotypes, using the maximum likelihood program codonbias (Nielsen et al. 2007). The values indicated for each terminal branch represent estimates from the best fitting models for the full set of loci ($S_{\text{all}}$) and highly expressed genes ($S_{\text{high}}$) in each genotype. Branches are colour coded to represent the floral morph structure of the populations from which the genotypes originated. Monomorphic populations are predominately selfing and trimorphic populations are largely outcrossing. Branch lengths are not to scale.
CHAPTER SEVEN

RECONCILING GENE AND GENOME DUPLICATION EVENTS: USING MULTIPLE NUCLEAR GENE FAMILIES TO INFER THE PHYLOGENY OF THE AQUATIC PLANT FAMILY PONTEDERIACEAE

This chapter resulted from a collaboration with Sean W. Graham and Spencer C. H. Barrett. Sean Graham contributed ideas, advice on data analysis and to the writing of the manuscript. Spencer C. H. Barrett contributed ideas and to the writing of the manuscript which is currently under review

Summary

Most plant phylogenetic inference has used DNA sequence data from the plastid genome. This genome represents a single genealogical sample with no recombination among genes, potentially limiting the resolution of evolutionary relationships in some contexts. In contrast, nuclear DNA is inherently more difficult to employ for phylogeny reconstruction because major mutational events in the genome, including polyploidization, gene duplication and gene extinction can result in homologous gene copies that are difficult to identify as orthologs or paralogs. Gene tree parsimony (GTP) can be used to infer the rooted species trees that are most compatible with one or more gene families that arise through duplication events, by fitting gene genealogies to species trees while simultaneously minimizing the estimated number of duplications needed to reconcile conflicts among them. Here, we use GTP for five nuclear gene families and a previously published plastid data set to reconstruct the phylogenetic backbone of the aquatic plant family Pontederiaceae. Plastid-based phylogenetic studies strongly supported extensive paraphyly in Eichhornia (one of the four major genera), but also depicted considerable ambiguity concerning the true root placement for the family. Our results indicate that species trees inferred from the nuclear genes (alone and in combination with the plastid data) are highly congruent with gene trees inferred from
plastid data alone. Consideration of optimal and suboptimal gene tree reconciliations place the root of the family at (or near) a branch leading to the rare and locally restricted \textit{Eichhornia meyeri}. We also explore methods to incorporate uncertainty in individual gene trees during reconciliation by considering their individual bootstrap profiles, and relate inferred excesses of gene duplication events on individual branches to whole-genome duplication events inferred for the same branches. This study increases our understanding of the phylogenetic history of Pontederiaceae and also demonstrates the utility of GTP for phylogenetic analysis.

\textbf{Introduction}

Phylogenetic inference provides an historical framework for understanding fundamental evolutionary processes such as speciation and adaptation. Phylogenetic reconstruction in plants has largely used data from the plastid (chloroplast) genome, although low copy nuclear genes are increasingly employed (e.g. Steele et al. 2008; Duarte et al. 2010; Regier et al. 2010), and the growing availability of large genomic data sets from multiple taxa has the potential to revolutionize the inference of plant phylogeny. Nonetheless, the plastid genome continues to be the molecule of choice for phylogenetic studies for a variety of reasons (e.g. Olmstead and Palmer 1994; Graham and Olmstead 2000). For example, it has a relatively conservative DNA substitution rate across a range of genes and a relatively conserved gene order in most land plants. It also has strong conservation in gene number, with generally only one gene copy per genome; plastid genes are occasionally lost but never transferred from other genomes. These properties reduce complications associated with primer design, sequence recovery and orthology assignment across divergent taxa. However, the entire plastid genome is a single, non-recombining linkage group, which means that
different plastid genes do not provide completely independent records of phylogenetic history (e.g. Doyle 1992; Maddison 1997).

The nuclear genome, in contrast, represents a potentially much larger source of information on phylogenetic relationships and individual genes can belong to multiple linkage groups. However, the relatively rapid evolution of nuclear genes, coupled with the overall fluidity of this genome in terms of gene copy number and order, make phylogenetic inferences more challenging. A particular difficulty is the uncertainty in the orthology of the characters used. Orthology of genes is almost always unambiguous for plastid data, but nuclear genes are usually part of small to large gene families that undergo repeated rounds of expansion and contraction in copy number due to gene duplication and extinction processes (e.g. Kellogg and Bennetzen 2004). Over time these processes commonly obscure orthology assessment. Incorrect assumptions about gene orthology, particularly mistaking paralogs for orthologs, can lead to conflicting gene trees and uncertain or distorted inferences of the overall species tree from gene tree data (Maddison 1997).

The increasingly inexpensive recovery of large-scale nuclear data sets through advances in DNA sequencing technology has encouraged the development of theoretical and analytical techniques for assessing gene orthology. Orthology assessment requires reconciling apparent conflicts within or among phylogenies inferred from multi-gene families, which is the basis of the gene tree parsimony (GTP) method (Goodman et al. 1979; Page and Charleston 1997; Slowinski et al. 1997). GTP considers both congruence and conflict among one or a collection of gene-tree genealogies, using both to infer an overall (rooted) species tree that minimizes the number of gene duplications, gene losses or deep coalescences. It does so by reconciling any conflicts among the genealogies considered. When considering multi-gene families, the reconciliation cost of a tree can be calculated
simply as the minimum number of gene duplications needed to reconcile the gene trees to the observed species tree. Gene losses can also be considered as part of the reconciliation cost, but this is not recommended where there is the possibility of incomplete sampling of all members of the gene family in each species (Page and Charleston 1997). The tree with the lowest reconciliation cost among and within gene trees is preferred as the best estimate of species phylogeny. GTP has been implemented in three programs GeneTree (Page 1998), Gtp (Sanderson and McMahon 2007) and DupTree (Wehe et al. 2008), and has been applied to data from a variety of taxonomic groups (e.g. Maier et al. 2001; Frajman et al. 2009; Holton and Pisani 2010). The advantage of GTP for phylogenetic inference using nuclear sequences is that no a priori knowledge about the orthology of gene copies is necessary to infer species trees; indeed, the reconciliation of conflict among gene genealogies provides core evidence for the overall pattern of species relationships.

Here, we use GTP to reconstruct the species phylogeny of the aquatic flowering plant family Pontederiaceae (Monocotyledoneae: Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009) using five nuclear gene trees and previously published plastid data (Graham et al. 1998; Graham et al. 2002). This small family is composed of four major genera (plus two to five widely recognized segregates), and 35-40 species, most of which are native to the New World tropics (Barrett 1978). Members of the family display a remarkable diversity of life history and reproductive strategies, ranging from highly clonal, long-lived taxa that inhabit permanent marshes and river systems, to exclusively sexual species that are annual and occur in ephemeral pools, ditches and rice fields. Linking these extremes are species with various combinations of sexual and asexual reproduction and a variety of floral strategies and mating systems. Evolutionary studies of the family over the past two decades have focused primarily on selected taxa that possess tristyous and homostyous reproductive
systems (reviewed in Barrett 1988; Barrett et al. 1992; Barrett 1993). Phylogenetic reconstructions using both morphological (Eckenwalder and Barrett 1986) and plastid sequence data (Graham and Barrett 1995; Kohn et al. 1996; Graham et al. 1998; Graham et al. 2002) have been employed to investigate character evolution and the systematic relationships of taxa within the family and its close relatives.

Four published datasets have been used to investigate systematic relationships and character evolution in Pontederiaceae. The first was based on morphological characters and the resulting trees were poorly supported, likely due to sampling error caused by the relatively small dataset (Eckenwalder and Barrett 1986). Later, plastid DNA restriction-site variation (Kohn et al. 1996) and DNA sequences from the plastid genes $rbcL$ and $nhdF$ (Graham and Barrett 1995; Graham et al. 1998; Graham et al. 2002) resulted in well supported trees that were mutually highly congruent, but incongruent with the poorly supported trees inferred from morphological data (Graham et al. 1998). The plastid genealogy strongly supported the monophyly of Pontederiaceae and three of the major genera ($Pontederia$, $Monochoria$, $Heteranthera$). In contrast, $Eichhornia$ was inferred to be non-monophyletic and to comprise four distinct lineages, two of which are polyploid. Our ability to investigate the polyploid origins of these species has been limited by the non-recombining nature of the plastid genome. In addition, there remains substantial ambiguity in the location of the root of the phylogeny (see Graham et al. 2002). Nuclear DNA sequence data may provide further insight into resolving these phylogenetic issues.

In this study, we present new data from five nuclear gene families recovered using primers designed from genes from an earlier EST study (Chapter 4). We used GTP on these nuclear genealogies to infer relationships among 14 species that we chose to encompass the broad phylogenetic backbone of Pontederiaceae. Our study addresses the following specific
issues: (1) How congruent is the species tree inferred by using GTP to reconcile the five nuclear gene families (either by themselves or in combination with plastid data) with previous phylogenetic estimates from plastid data alone? (2) Can GTP clarify our understanding of the placement of the root of Pontederiaceae, which is unclear from plastid data alone? (3) GTP assumes that the gene genealogies being reconciled have been inferred correctly, so how can we incorporate estimates of the uncertainty in gene tree inferences when inferring the species tree? (4) How do phylogenetic inferences of where the gene duplication events occurred correspond to inferences of past episodes of polyploidization in the family?

Methods

Taxon sampling

We selected species of Pontederiaceae distributed across major lineages, based on our current understanding of the phylogeny of the family from the plastid-based analysis of Graham et al. (2002). The exemplar species sampled for nuclear data were: *Eichhornia azurea*, *E. crassipes*, *E. meyeri*, *E. paniculata*, *E. paradoxa*, *Eichhornia* sp., *Heteranthera multiflora*, *H. seubertiana*, *H. zostericoflia*, *Hydrothrix gardneri*, *Monochoria hastata*, *M. korsakovii*, *Pontederia sagittata*, and *P. subovata*. Source information for these species is presented in Table 7.1.

Primer development and sequencing

Primers for nuclear loci were developed based on EST sequences derived from *E. paniculata* and *E. paradoxa* (for details of EST sequencing see Chapter 4). We identified ESTs with conserved DNA sequence in both *Oryza sativa* and *Zea mays*, under the
assumption that these sequences would be more similar across the species used in our study. We annotated the selected ESTs using BLAST2GO (Conesa et al. 2005; Götz et al. 2008), a program that uses BLASTx to identify all putative homologs in the non-redundant (NR) database at NCBI. We designed degenerate primers and used them to amplify gene regions from genomic DNA extractions. Six primer pairs that resulted in amplification across all samples were used. We cloned the products of each PCR reaction and sequenced both forward and reverse strands in 4-8 clones per amplification per species using an ABI 3730XL fluorescent-based capillary sequencer. We assembled forward and reverse sequence strands with Sequencher 4.7 and confirmed all genotypes manually. Sequences from each of the six amplifications were aligned using Muscle (Edgar 2004). We manually adjusted all alignments to ensure the most accurate alignment (e.g. Graham et al. 2000), and sequences with no similarity to any other sample were assumed to be cloning or PCR artifacts and were therefore discarded. Three of the six primer pairs amplified loci with conserved exons flanking a highly variable intron. In these cases, the introns were too divergent amongst the species to align reliably and so we excluded them from subsequent analyses. We also excluded near identical sequence reads (< 1% of sites were variable) derived from the same species to avoid including allelic variants or loci that only varied due to errors introduced during amplification, cloning, or sequencing. We completely excluded one candidate gene family that had too-limited recovery across the taxa that were sampled because it recovered clearly non-homologous fragments in different species, presumably due to low specificity of amplification. However, including homologous sequences from this gene family in the analysis has little or no effect on the results or conclusions (unpublished data).
Reconstructing genealogical and phylogenetic history

We generated maximum likelihood (ML) genealogies for each of the five alignments using the software Garli v1.0 (Zwickl 2006) with a general time reversible (GTR + I + \Gamma) model in which the substitution matrix, proportion of invariant sites (I) and shape of the gamma distribution (\Gamma, with four rate categories for the shape parameter \alpha) were estimated from the data. In addition, we generated 2000 bootstrap replicate genealogies for each alignment to assess support for the genealogies, and for use in later analyses.

To fit the best overall species tree for the five genealogies we used the program DupTree (Wehe et al. 2008). DupTree generates rooted species trees from multiple genealogies using GTP (Goodman et al. 1979; Page and Charleston 1997; Slowinski et al. 1997). It implements a standard rooted Subtree Pruning and Regrafting (rSPR) heuristic search to identify the optimal rooted species tree topology, which is the one that minimizes the number of inferred gene duplications necessary to reconcile each genealogy with this candidate species tree. DupTree minimizes only the number of duplications during reconciliation, rather than considering both duplications and losses as part of the reconciliation cost, because incomplete taxon and sequence sampling may be erroneously interpreted as gene copy losses (Page and Charleston 1997). We used the unrooted ML genealogies generated in Garli v1.0 as input for DupTree. Local search heuristics are not guaranteed to find the global optimum and so we re-ran DupTree 1000 times using random starting trees to increase the probability that all the shortest trees are found.

DupTree assumes that the gene genealogies used are fully resolved (bifurcating) and does not take into account uncertainty in each gene tree. We therefore attempted to assess the impact of uncertainty in genealogical inferences using two methods. In the first case (‘Bootstrap-1’), we evaluated the sensitivity of the inferred species tree to uncertainty in each
of the five genealogies, in turn. We ran 1000 iterations of DupTree where one of the five gene trees was represented by an individual tree inferred from a bootstrap pseudo-replicate, generated in Garli v1.0, and the best ML gene tree was used for the other four. We repeated this for each of the five gene trees and summarized the effect of each on the species tree by calculating the proportion of iterations that supported each clade pooled across results from the separate analyses of each locus. We computed this using the majority-rule consensus function in PAUP* (Swofford 2003). In the second method (‘Bootstrap-2’), we assessed the effect of uncertainty in genealogical inferences of all alignments simultaneously. We ran 1000 DupTree iterations in which we simultaneously sampled single ML bootstrap pseudo-replicate genealogies for each of the five gene families per iteration. We again summarized the results as the fraction of analyses that recovered each clade of interest across iterations.

We repeated all of these analyses considering the unrooted plastid-based genealogy of Graham et al. (2002) as an additional gene tree. We pruned taxa in that study that were not included in our study to maintain similar taxon sampling, and interpolated the position of two taxa sampled here but not there. Specifically, *Pontederia subovata* was interpolated as the sister group of *P. sagittata* in the plastid tree (representing an assumption of monophyly for this genus, and based on unpublished plastid sequence data), and *H. multiflora* was interpolated in the position of *H. reniformis* in the study of Graham et al. (2002), as these three *Heteranthera* species were predicted to be closely related to each other by Horn (1985).

**Rooting the phylogeny of Pontederiaceae**

The current plastid genealogy of Pontederiaceae is well resolved, but its exact rooting remains unclear (see Graham et al. 2002). We used the program GeneTree v1.3.0 (Page 1998) to calculate the reconciliation cost for re-rooting the overall species tree, estimated
from each of the five nuclear gene trees and the plastid data, on all of its sub-optimal branches. We used GeneTree to reconcile each of the five nuclear gene-tree genealogies in turn to the optimal species tree (GeneTree can only perform pair-wise reconciliations), repeating this for each of its 26 possible roots, and then summing the minimum number of duplications required to fit each of the genealogies for each re-rooting. GeneTree requires rooted genealogies as input for this analysis. To root the individual nuclear gene trees generated in Garli v1.0, we used the most common, shortest, rooted genealogy for each alignment that was estimated by running 1000 iterations of DupTree; rooted gene trees are part of the output from the tree reconciliation exercise. We used this method to examine suboptimal species tree roots, including those that are nearly optimal. Using GeneTree in this manner might be expected to recover the same best species tree as DupTree, as the two programs use the same underlying reconciliation principle.

**Gene duplication and polyploidy**

We assessed correspondence between instances of whole-genome duplications (including paleopolyploidy events inferred on internal branches) and the number and position of inferred gene duplications on one of the four species tree inferred from the nuclear and plastid gene trees (see below). We first mapped polyploidization or demi-polyploidization events onto the tree using the program ChromEvol v1.1 (Mayrose et al. 2010). ChromEvol defines polyploidization as a doubling of the haploid chromosome number while demi-polyploidization is an increase of $1.5 \times$ in the chromosome number. The program uses the current chromosome number of each species and the rooted species tree to reconstruct ancestral changes in chromosome number in a phylogeny. We estimated the parameters for each of eight models included with the program and chose the most likely model with the
best fit based on the Akaike Information Criterion. The model chosen estimates a constant rate of chromosome loss and gain, and constrains the rate of demi-polyploidization and polyploidization to be equal and constant across the tree. This allowed us to map changes in ploidy along each branch in the tree. We then used GeneTree to identify the location on the tree where gene duplications occurred. The number of gene duplications inferred for each branch was summed across all five nuclear gene-tree reconciliations and these were visually compared with the occurrence of partial or whole genome duplications.

**Results**

In total, we recovered 226 alignable sequences for the five primer pairs totaling 2524 bp of aligned nuclear sequence per taxon, after intron exclusion. Following the removal of nearly identical sequences from the same species, 108 ‘unique’ sequences were identified (mean 21.6 sequences/alignment). In this reduced set, there were 560 parsimony informative sites (Table 7.2). The bootstrap majority-consensus for each of the five unrooted genealogies that we reconstructed is presented in Figure 7.1. The proportion of branches resolved with greater than 50% support varies from 0.57-0.80, with a mean bootstrap value of 82.7% for resolved branches across the five genealogies. As expected in genealogies with paralogous sequences, the relations among species in each of the genealogies did not unambiguously reflect the known relationships or the taxonomy of the family.

**Nuclear-based species tree**

The rooted species tree inferred from the five genealogies is contrasted with the plastid topology of Graham et al. (2002; modified here with two interpolated taxa) in Figure 7.2. Using our ‘Bootstrap-1’ method to assess support for this tree, we found 10 of 12
branches were supported by greater than 70% of the iterations. On average, 73% of bootstrap replicates from each of the five genealogies supported a rooting of Pontederiaceae on the branch leading to *E. meyeri* as the best choice. The nuclear-based tree is similar to the current plastid topology with two exceptions. Firstly, *Hydrothrix gardneri* is nested within *Heteranthera s.s.*, whereas in the most recent plastid trees *H. gardneri* is depicted as the sister group of other *Heteranthera* species. Secondly, unlike the plastid-based tree the nuclear tree did not support the monophyly of *Pontederia*, but instead indicated that *P. subovata* was sister to *E. azurea* and *P. sagittata*, among taxa sampled here.

*Combined nuclear and plastid species tree*

We recovered four shortest species trees; one of the four trees matched the unrooted plastid genealogy in its underlying topology and we chose to use this tree for display purposes (Figure 7.3). All subsequent analyses that are presented here are based on this species tree; however, the results are consistent across all four trees. We assessed the effect of uncertainty in individual genealogies on the species tree using two methods that incorporate bootstrap pseudo-replicates from each gene family. In both cases, a high proportion of species tree inferences using individual or multiple bootstrapped gene trees recovered the same clades found using the species tree inference illustrated in Figure 7.3. This pattern remained when we bootstrapped one genealogy at a time, pooling results for each iteration (‘Bootstrap-1’) and when we used bootstrap replicates for all genealogies simultaneously for an iteration (‘Bootstrap-2’). In both cases, a high fraction of bootstrapped reconciliations supported a root for the species tree with *Eichhornia meyeri* as the sister group to the rest of the family (88% with ‘Bootstrap-1’ and 68% with ‘Bootstrap-2’). Most branches were recovered with at least 50% support from both bootstrap methods (excluding
one branch in *Heteranthera* and the monophyly of *Pontederia* with ‘Bootstrap-1’). To explore whether the plastid genealogy drove this result we also generated 10 random equiprobable rooted genealogies to replace the plastid tree and visually compared the resulting species tree inferred by DupTree with the topology of the random tree. The similarity and effect of the random tree on the inferred species tree was minimal confirming that the plastid genealogy was not driving the congruence of the combined nuclear-plastid tree with previous plastid phylogenies.

**Rooting the Pontederiaceae phylogeny**

We used the program GeneTree to calculate the reconciliation cost of every possible rooting of the Pontederiaceae phylogeny. The cost, expressed in total number of duplications required to reconcile all five genealogies varied from 63 to 47 across the 26 possible roots. The five most parsimonious possible roots are shown in Figure 7.4. As expected, because Duptree and GeneTree use the same underlying reconciliation method, both agree that the branch leading to *E. meyeri* is the lowest cost root for the species tree that reconciles all of the nuclear and plastid gene trees (Figure 7.3, 7.4). The branch leading to *Heteranthera* and the branch connecting *Heteranthera* and *E. meyeri* to the rest of the family were the next two lowest cost possibilities, with one or two additional duplication(s) required to reconcile the genealogies compared to the optimal root.

**Gene duplication and polyploidy**

To investigate the association of polyploidy with gene duplications, we mapped polyploidization events and gene duplications onto the species tree reconstructed from the
nuclear and plastid gene trees (Figure 7.5). We inferred four full genome duplications and one demi-polyploidization inside the crown clade of Pontederiaceae. Three of these polyploidization events occurred on the terminal branches leading to *E. azurea*, *E. crassipes* and *M. korsakovii*, respectively, and the fourth occurred on the stem lineage leading to *Heteranthera*. With GeneTree we inferred a total of 46 nuclear gene duplications distributed across 14 branches in the phylogeny, including 12 duplications on the stem lineage leading to the family (i.e., before the radiation of the extant species, assuming the current sampling represents the whole crown). All instances of polyploidization were associated with gene duplication events.

Discussion

A major finding of this study is that the species tree we obtained from five nuclear gene trees is broadly congruent with previous phylogenetic estimates of the Pontederiaceae using plastid data alone, and is reasonably robust relative to the uncertainty inferred in the underlying gene trees. We also recover a single best solution for the root of Pontederiaceae using these new data, although there are also several nearly optimal alternatives to this rooting. Below we contrast our results with previous phylogenetic studies and discuss the potential causes of the relatively minor topological differences that were revealed between nuclear- versus plastid-based species trees. We also consider the implications of polyploidy and gene duplication for phylogenetic inference using nuclear sequence data.

Nuclear-based species tree

Gene-tree reconciliation as currently implemented does not attempt to take account of uncertainty in the underlying gene trees – these are assumed to be correct. However,
individual gene trees are undoubtedly subject to stochastic error and it would be appropriate to consider bootstrap support for the individual branches that are being reconciled. We attempted to do this here by developing two different methods for incorporating the tree uncertainty captured by bootstrap analysis. The support for the species tree measured with our ‘Bootstrap-1’ method was quite strong (in this method we examined the effect of including in each iteration, for each of the gene families, one pseudoreplicate bootstrap tree along with the best trees for the four other multi-gene families). Most species-tree branches in this case were recovered in at least 70% of iterations. Not surprisingly given the potential for a greater diversity of tree topologies in each iteration, support values from our ‘Bootstrap-2’ method were substantially lower (in this case, each iteration included only bootstrap pseudo-replicate gene trees from all five multi-gene families). The real robustness estimate for tree reconciliation may lie between these two estimates, but until a better method is developed for taking account of tree uncertainty we offer this approach as a possible means for estimating the certainty of species tree inferences using GTP. Despite the overall similarity between the five gene tree nuclear reconciliation presented here and previous estimates of phylogeny using plastid data alone (Graham et al. 1998; 2002) there were two instances of incongruence (Figure 7.2). One of these concerns *Heteranthera* (including *Hydrothrix gardneri*) and the other the placement of *E. azurea* relative to the genus *Pontederia*.

First, our nuclear-based phylogeny was potentially incongruent with previous estimates of the phylogeny using plastid sequence data with respect to the placement of *H. seubertiana* relative to other species of *Heteranthera*. The nuclear-based species tree depicts *H. seubertiana* as the sister group of a clade comprising the remaining species of *Heteranthera*, defined broadly here to include *H. gardneri* (Figure 7.2), with weak to
moderate support from the two different bootstrap methods. In contrast, previously published plastid datasets strongly group *H. seubertiana* in a clade with *H. zosterifolia*, to the exclusion of the other two species sampled here. The earlier studies were also unable to robustly resolve the root of *Heteranthera* (in the largest study, Graham et al. 1998 recovered a basal trichotomy in *Heteranthera* between *H. gardneri*, *H. limosa-H. rotundifolia* and a clade of species that included *H. zosterifolia* and *H. seubertiana*).

One line of evidence supporting *H. gardneri* as sister to the rest of *Heteranthera* comes from analysis of chromosome number evolution. Reconstructions using ChromEvol (not shown) using the topology of *Heteranthera* in the nuclear-based species tree (Figure 7.2; left-hand tree) require two independent genome duplications in this clade rather than the single one inferred if *Hydrothrix* is assumed to have the plastid placement (Figure 7.3). Thus, a more parsimonious arrangement considering chromosome number is one that places the morphologically distinctive *H. gardneri* as sister to the rest of the *Heteranthera*.

The second discordant result between our nuclear-based species tree and published plastid-based trees (e.g. Graham et al. 1998) is that *Pontederia* is not supported as monophyletic (Figure 7.2: *P. subovata* is depicted as the sister group of *P. sagittata-E. azurea*). This result has only weak to moderate bootstrap support and so it may reflect stochastic error in one or more nuclear gene trees. Unpublished plastid sequence data from *ndhF* recovers the monophyly of *Pontederia*, placing *P. subovata* as sister to all other species of *Pontederia*. Our analysis with ChromEvol inferred a polyploidization event on the terminal branch leading to *E. azurea* (Figure 7.5), and it is possible that this genome duplication is the result of an ancient hybridization event between a species of *Pontederia* and the ancestor of *E. azurea*. However, if this were true it would not necessarily lead to recovery of the species tree inferred using nuclear data seen in Figure 7.2, as GTP assumes
strictly bifurcating gene trees and species trees, and is unable to accommodate discordance
due to hybridization events (Page and Charleston 1997). Therefore, the recovery of a
relationship here that is consistent with this type of polyploidization event would be at best a
coincidence and at worst an artifact. Nonetheless, it should be possible to resolve this issue
by additional sampling of taxa (more Pontederia species, and species of Eichhornia inferred
to be closely related to E. azurea in plastid-based analyses, such as E. diversifolia and E.
heterosperma).

**Combined nuclear and plastid species tree**

When we combined our five nuclear gene trees with the plastid gene tree of Graham
et al. (2002) in a joint GTP analysis we recovered a species tree that was completely
congruent with the previous plastid-based phylogeny (Figure 7.3). We also recovered three
additional trees with the same reconciliation cost. The four shortest trees each have
combinations of the two differences found in the nuclear-based tree discussed above (i.e.,
two locations of E. azurea × two topologies of Heteranthera). Gene duplications in the
plastid tree are extremely unlikely. In fact, none have been observed in the long course of
land-plant evolution, with the exception of duplications in the plastid inverted repeat region,
where duplicated genes do not diverge (Goulding et al. 1996). Nonetheless, the plastid tree is
worth including in reconciliations, in the same manner that a strictly single-copy nuclear
gene would be worth including in tree reconciliations, as it is an independent estimate of
species phylogeny and thus could help uncover additional gene-tree conflicts.

Corroboration of the general shape of the Pontederiaceae phylogeny supports
previous conclusions derived from plastid DNA sequence (e.g. Graham et al. 1998; Graham
et al. 2002). In particular, our results support the idea that the genus Eichhornia is highly
paraphyletic. Modifications to the taxonomy of the family to take account of our findings could either recognize distinct genera to accommodate these lineages (note that as *Eichhornia azurea* is the type species of the genus, only it and its close relatives could retain this name), combine one or more genera (*Pontederia* has nomenclatural priority in the family), or employ rank-free taxonomy (i.e. the draft Phylocode, available online at www.phylocode.org).

**Rooting the Pontederiaceae species tree**

Rooting the Pontederiaceae tree has remained a difficult problem because of the erosion in phylogenetic signal of divergent outgroups which may lead to artifactual or ambiguous rootings. Graham et al. (2002) used multiple outgroups to root the Pontederiaceae and reported two potential root locations for different optimality criteria, and a range of suboptimal rootings that could not be ruled out statistically (see Figure 5 in Graham et al. 2002). The optimal root based on parsimony split the family into *Heteranthera* (including *Hydrothrix*) versus all other species, whereas maximum likelihood favored a root on the branch leading to both *Heteranthera* and *E. meyeri*. More extensive sequence data may help resolve the root of the plastid-based tree (S.W. Graham, unpublished data) but may still be subject to the long branch connecting Pontederiaceae and its closest relatives. Our gene tree reconciliations did not include outgroup taxa but instead considered gene duplication evidence alone to root the family tree, which may be a useful alternative when the nearest sister group is distantly related to the ingroup (Mathews and Donoghue 1999). Our analyses support a rooting of the family that places *Eichhornia meyeri* as an isolated lineage sister to the remainder of the family (Figure 7.2-7.4). Five gene duplications support this rooting (Figure 7.5), and there is also weak to moderately strong support for this considering our
different bootstrap measures (Figure 7.2, 7.3). However, there are several root placements implied by the gene duplication data that are nearly as optimal; only two gene duplications separate the reconciliation costs of the top three root placements. The second and third most likely root placements are the optimal placements found for the best ML and MP trees for plastid-based phylogenies in Graham et al. (2002). Thus our results cannot be considered definitive, although the nearly congruent rooting arrangement with this earlier study is encouraging. Clearly, however, gene tree reconciliation of nuclear and plastid trees have provided new insight into the difficult problem of rooting the Pontederiaceae phylogeny and with larger sequence datasets and the addition of outgroup samples a solution to the problem may be achievable.

The finding that *Eichhornia meyeri* may be the sister group to the rest of Pontederiaceae is of particular interest for several reasons. First, the species is rather poorly known, having only been reported from a few localities in Northern Argentina, Paraguay and Brazil, sites that are largely associated with the wetlands of the Grand Chaco and Matto Grosso. *Eichhornia meyeri* most closely resembles *E. paniculata* morphologically, which it has often been confused with in the literature (Eckenwalder and Barrett 1986). Earlier studies of the floral biology of this species (Barrett 1988) demonstrated that the species was monomorphic for style length, with two sets of stamens positioned close to the stigmas. This arrangement results in high levels of autonomous self-pollination and resembles the semi-homostylous condition found in several tristylos species of *Eichhornia* in which tristyly has broken down resulting in the evolution of self-pollinating forms (reviewed in Barrett 1988). However, the results from this study support the view that the monomorphic floral condition of *E. meyeri* is likely plesiomorphic (preceding the evolution of tristyly in the family).

*Eichhornia meyeri* is self-compatible and the root position inferred here for Pontederiaceae,
with *E. meyeri* sister to the remainder of the family, would still be consistent with our earlier studies in which self-incompatibility is inferred to be plesiomorphic (e.g. Barrett and Graham 1997), in contrast to most other self-incompatible flowering-plant families. The rarity of *E. meyeri* and its restricted distribution may be associated with progressive extinctions and loss of its wetland habitats during its long evolutionary history. The species clearly deserves renewed conservation attention and additional study.

**Gene duplication and polyploidy**

Using contemporary chromosome numbers we inferred four polyploidization events and one demi-polyploidization in the crown clade of Pontederiaceae (Figure 7.5). These numbers are almost certainly underestimates for these nuclear gene families, because other species in the family not included here have polyploid chromosome numbers (e.g. *Pontederia* includes other diploid and tetraploid taxa; Eckenwalder and Barrett 1986), and *H. zostericola*, included in our study, has no chromosome number estimate. The polyploidization events we inferred may be one of the major causes of the 46 gene duplications revealed by our analysis, as twenty of the total gene duplications occur in polyploid lineages and 13 of these are located on terminal branches of polyploid species. However, 12 gene duplications are inferred outside the crown clade. One possible explanation for this large number of early gene duplications is that they represent paralogous gene copies that arose on the stem lineage, or before the split of Pontederiaceae from its sister group, Haemodoraceae (Saarela et al. 2008) Alternatively, the duplications could reflect incomplete sampling of all members of a gene family. For example, divergent gene copies for which we have not sampled other orthologous sequences may be interpreted as duplications at the base of a tree because no information exists to support placement
elsewhere. The number of gene duplications and their dispersion across the tree highlights the complexity of identifying orthologous sequences and the potential difficulties in using nuclear gene trees to infer species trees. Despite these complexities, the species tree inferred here is highly congruent with previous estimates based on plastid data alone, supporting the use of GTP as a method for mining further phylogenetic information from the nuclear genomes of these and other plants.

Conclusions

The aim of our study was to explore the utility of multiple nuclear gene trees for inferring the phylogenetic history of Pontederiaceae. Our investigations have demonstrated that the shape of the nuclear-based species tree is generally consistent with the plastid tree, and when we combined our nuclear gene trees with the plastid genealogy we recovered a species tree that was completely congruent with the previous phylogenetic estimates. Our analysis also provided new evidence supporting a root placement in which *E. meyeri* is the sister group of the rest of the family, although several other placements are nearly as optimal. Lastly, by modeling the evolution of chromosome number we showed that polyploidy could be responsible for a sizeable fraction of gene duplications in our gene trees. The history of gene and genome duplication complicates the relationships among homologous nuclear loci. However, using the phylogenetic signal present in multiple genealogies can provide a valuable method for inferring the relations among species by reconciling conflicts and clarifying the identity of orthologous versus paralogous gene copies. This approach will become increasingly relevant as the number of large-scale nuclear genome sequencing projects burgeons in future years.
Table 7.1. Species from the family Pontederiaceae chosen for this study. Voucher and Genbank accession numbers for samples and DNA sequences used.

<table>
<thead>
<tr>
<th>Species</th>
<th>Voucher¹</th>
<th>GenBank Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Eichhornia azurea</em> (Swartz) Kunth</td>
<td>-na-</td>
<td>---</td>
</tr>
<tr>
<td><em>E. crassipes</em> (Mart.) Solms-Laub.</td>
<td>Barrett 814</td>
<td>---</td>
</tr>
<tr>
<td><em>E. meyeri</em> Schulz</td>
<td>Barrett 1409</td>
<td>---</td>
</tr>
<tr>
<td><em>E. paniculata</em> (Spreng.) Solms-Laub.</td>
<td>Barrett 1401</td>
<td>---</td>
</tr>
<tr>
<td><em>E. paradoxa</em> (Mart.) Solms-Laub.</td>
<td>Barrett 1402</td>
<td>---</td>
</tr>
<tr>
<td>²<em>Eichhornia sp.</em></td>
<td>Barrett and Shore 1399</td>
<td>---</td>
</tr>
<tr>
<td><em>Heteranthera reniformis</em></td>
<td>-na-</td>
<td>---</td>
</tr>
<tr>
<td><em>H. seubertiana</em></td>
<td>Barret 1412</td>
<td>---</td>
</tr>
<tr>
<td><em>H. zosterifolia</em> Mart.</td>
<td>Barrett 1413</td>
<td>---</td>
</tr>
<tr>
<td><em>Hydrothrix gardneri</em> Hook f.</td>
<td>Barrett 1414</td>
<td>---</td>
</tr>
<tr>
<td><em>Monochoria hastata</em> (L.) Solms-Laub.</td>
<td>Barrett 1407</td>
<td>---</td>
</tr>
<tr>
<td><em>M. korsakovii</em> Reg. and Maack</td>
<td>Barrett 1415</td>
<td>---</td>
</tr>
<tr>
<td><em>Pontederia sagittata</em> Presl</td>
<td>Barrett 1416</td>
<td>---</td>
</tr>
<tr>
<td><em>P. (Reussia) subovata</em></td>
<td>-na-</td>
<td>---</td>
</tr>
</tbody>
</table>

¹ Each voucher (deposited at the Green Plant Herbarium (TRT), Royal Ontario Museum, Toronto, Ontario, Canada) is a representative individual of material that was under cultivation at Toronto.

² An undescribed species of *Eichhornia* (referred to here as *Eichhornia sp.*.) was identified in Eckenwalder and Barrett (1986) as *E. paradoxa*. 
Table 7.2. Summary information about six ESTs used to design primers for amplifying nuclear regions in Pontederiaceae.

<table>
<thead>
<tr>
<th>EST label</th>
<th>BLAST protein description</th>
<th>No. seq.</th>
<th>Unique seq.</th>
<th>Aligned length (bp)</th>
<th>Parsimony informative sites</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>EX0014</td>
<td>Protein phosphatase</td>
<td>59</td>
<td>18</td>
<td>334</td>
<td>65</td>
<td>24</td>
</tr>
<tr>
<td>EX0042</td>
<td>Coatomer subunit epsilon</td>
<td>21</td>
<td>15</td>
<td>441</td>
<td>68</td>
<td>56</td>
</tr>
<tr>
<td>P0133</td>
<td>Nucleoside diphosphate kinase</td>
<td>39</td>
<td>22</td>
<td>255</td>
<td>55</td>
<td>19</td>
</tr>
<tr>
<td>P0249</td>
<td>Importin alpha</td>
<td>61</td>
<td>32</td>
<td>1040</td>
<td>271</td>
<td>69</td>
</tr>
<tr>
<td>P0263</td>
<td>Protochlorophyllide reductase precursor</td>
<td>46</td>
<td>21</td>
<td>454</td>
<td>101</td>
<td>59</td>
</tr>
<tr>
<td>P0508³</td>
<td>DNA-J like protein</td>
<td>25</td>
<td>10</td>
<td>567</td>
<td>101</td>
<td>63</td>
</tr>
</tbody>
</table>

¹ Protein descriptions are based on BLAST2GO annotations of EST translations
² Number of sequences included after removing nearly identical (>99% identical) sequences from the same species
³ This locus was excluded from all analyses presented in this paper
Figure 7.1. Maximum-likelihood reconstructions of each of five nuclear gene trees generated using Garli v1.0. Bootstrap support values from 500 replicates are shown on each branch. The trees correspond to those generated using the following primer sets: (A) EX0014, (B) EX0042, (C) P0133, (D) P0249 and (E) P0263. These unrooted gene trees are depicted with arbitrary roots for display purposes only.
Figure 7.2. (A) Rooted species tree for the Pontederiaceae inferred in DupTree using five nuclear genealogies. The first value was generated by sampling a bootstrap replicate for one of five gene trees in each iteration, with the remaining four were input as the best ML gene trees (‘Bootstrap-1’). Values in brackets were generated by randomly sampling bootstrapped gene trees for all five alignments for each iteration (‘Bootstrap-2’). (B) Unrooted plastid genealogy from Graham et al. (2002) with interpolated positions of *Pontederia subovata* and *Heteranthera multiflora* (see text).
Figure 7.3. One of four most parsimonious species tree based on simultaneous reconciliation of five nuclear gene trees and a plastid gene tree. Values on branches represent estimates of support made by considering bootstrapped gene trees during tree reconciliation, using our two bootstrap-based methods. The first value was generated by sampling a bootstrap replicate for one of five gene trees in each iteration, with the remaining four were input as the best ML gene tree (‘Bootstrap-1’). Values in brackets were generated by randomly sampling bootstrap genealogies for all five alignments for each iteration (‘Bootstrap-2’). The two branches marked with asterisks (*) collapse in the strict consensus of the four most parsimonious trees. The floral form of each species is indicated at branch tips, open symbols and closed symbols represent self-compatible or self-incompatible species respectively.
Figure 7.4. Species tree of the Pontederiaceae based on combined nuclear and plastid sequences presented as an unrooted cladogram showing reconciliation costs for root placements on each branch. Reconciliation costs are expressed in terms of the number of gene duplications necessary to reconcile all five nuclear gene trees to the species tree based on nuclear and plastid data, when rooted on the branch indicated. The lowest cost branch is marked with an asterisk (*) and the next two lowest costs root placements are bolded.
Figure 7.5. Species tree of Pontederiaceae inferred using five nuclear gene trees and the plastid gene tree. Polyploid branches were mapped using ChromEvol, with inferred shifts marked in red for full genome duplication and orange for demi-polyploidization. Values indicate the inferred location and number of gene duplications made when reconciling individual nuclear gene trees to the species tree. Branch tips are labeled with the species name and haploid chromosome number ($n$) from Eckenwalder and Barrett, (1986).
CHAPTER EIGHT
CONCLUDING DISCUSSION

In this thesis, I investigated the evolutionary genetics of mating-system variation in *E. paniculata*, specifically the molecular population genetic and genomic consequences of evolutionary transitions from outcrossing to selfing. I generated and analyzed large DNA sequence datasets to address a host of questions involving the impact of mating system, demographic history, selection and drift on patterns of nucleotide variation. In this final chapter of my thesis I summarize the main conclusions of each of the preceding six chapters and provide suggestions for future research on *Eichhornia* that might be profitably undertaken.

Summary of thesis chapters

**Chapter Two – Genomic consequences of outcrossing and selfing in plants.** In this chapter, I reviewed opportunities for selfing populations to avoid an irreversible decline in fitness towards extinction and the implications of selfing for several aspects of genome evolution. I discussed theories for how selfers may avoid fitness declines such as the mutation rates, recombination, larger census population size, reduced activity of selfish genetic elements and smaller genome sizes. I showed, using comparative data, that highly selfing plants have significantly smaller genomes relative to their outcrossing relatives, consistent with the reduced activity and spread of repetitive elements in inbred plants. I discussed opportunities for tests of theory as plant genomic data accumulate and argued that a genomic perspective on reproductive transitions in a phylogenetic context should provide important future insights into the diversity of reproductive systems in flowering plants.
Chapter Three – *Evolutionary pathways to self-fertilization in a tristyloous plant species*. In this chapter, I discussed how the morphological polymorphisms that characterize heterostyly provide opportunities to study the evolutionary pathways from outcrossing to selfing. I investigated the particular pathways by which selfing has evolved in tristyloous *Eichhornia paniculata* by providing new evidence based on morphology, DNA sequences and genetic analysis. The primary pathway from outcrossing to selfing involves the stochastic loss of the S-morph from trimorphic populations followed by the spread of selfing variants of the M-morph. However, the discovery of selfing variants of the L-morph in Central America indicates a secondary pathway and distinct origin for selfing in *E. paniculata*. Comparisons of multi-locus nucleotide sequences from 27 populations sampled from throughout the geographical range were consistent with multiple transitions to selfing. Genetic analysis of selfing variants of the L- and M-morphs demonstrated recessive control of the loss of herkogamy, although the number of factors appeared to differ between the forms. Early stages in the establishment of selfing involve developmental instability in the formation of flowers capable of autonomous self-pollination. I proposed that the relatively simple genetic control of the loss of herkogamy and frequent colonizing episodes often creates demographic conditions favouring transitions to selfing in *E. paniculata*.

Chapter Four – *Mating-system variation, demographic history and patterns of nucleotide diversity in the tristyloous plant Eichhornia paniculata*. In this chapter, I examined how variation in mating system influences $N_e$, genetic diversity and population differentiation. I sequenced 10 nuclear loci in 225 individuals from 25 populations sampled from throughout the geographic range and used coalescent simulations to investigate demographic history. Highly selfing populations exhibited moderate reductions in diversity, but there was no significant difference in variation between outcrossing and populations with a mixture of
outcrossing and selfing. Population size interacted strongly with mating system and explained more of the observed variation in diversity within populations. Bayesian structure analysis revealed strong regional clustering of populations and those that were selfing were highly differentiated based on an analysis of $F_{st}$. Regional samples revealed greater breakdown of linkage disequilibrium in Brazil than in selfing populations from the Caribbean. Coalescent simulations indicated a moderate bottleneck associated with colonization of the Caribbean from Brazil ~125,000 years before present. These results suggested that the recent multiple origins of selfing in *E. paniculata* from diverse outcrossing populations have resulted in higher diversity than would be expected under long-term equilibrium.

**Chapter Five** – De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. In this chapter, I used short-read sequencing for de novo assembly of the floral transcriptomes of four genotypes of *Eichhornia*, including one outcrosser and two independently derived selfing lineages of *E. paniculata*, and the sister species *E. paradoxa*. I assembled ~27,000 contigs representing nearly 24Mbp of coding sequence. All four genotypes had highly correlated gene expression and the floral transcriptomes of the three *E. paniculata* genotypes were more correlated with one another than with *E. paradoxa*. The transition to selfing in flowering plants is usually characterized by a marked reduction in flower size and the loss of traits involved in pollinator attraction and the avoidance of self-fertilization. My analysis identified 269 genes associated with floral development, 22 of which were differentially expressed in selfing lineages of *E. paniculata*, relative to the outcrosser. Many of the differentially expressed genes affect floral traits that are likely to be functionally associated with the transition to selfing and they therefore represent a valuable set of potential candidate genes for investigating the evolution
of the selfing syndrome. This study demonstrated the utility of short read sequencing for de novo transcriptome assembly in non-model plant species.

**Chapter Six** – *Genomic consequences of transitions from cross- to self-fertilization on the efficacy of selection in three independently derived selfing plants.* In this chapter, I investigated how transitions from cross-fertilization to self-fertilization influenced the efficacy of natural selection on protein-coding genes. The effect of drift is predicted to reduce selective constraints on protein sequences and cause relaxed selection for biased codon usage. I examined patterns of nucleotide variation in protein-coding genes to assess the effect of inbreeding on the accumulation of deleterious mutations in the same genotypes described in Chapter five, including one outcrosser, two independently derived selfers and *E. paradoxa*, a selfing outgroup. This dataset included ~8000 orthologous sequences totalling ~3.5 Mb of coding DNA. Tests of selection demonstrated the occurrence of purifying selection across the transcriptome. I found a small but significant elevation in the proportion of non-synonymous to synonymous changes in the two *E. paniculata* selfers, relative to the outcrosser. Two lines of evidence supported a reduction in selection for biased codon usage in selfing lineages. Measurements of differentiation in codon usage between high vs. low expression genes indicated significantly lower codon usage in the three selfers, compared to the outcrosser. Second, estimates of the selection coefficient for codon usage bias also supported relaxed selection on synonymous sites in the three selfing genotypes. These findings, although based on a limited sampling of genotypes, indicate a reduced efficacy of selection in selfers, especially with respect to weakly selected synonymous changes that affect codon usage.

**Chapter Seven** – *Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae.* In this
chapter, I revisited the phylogeny of Pontederiaceae, a small aquatic plant family. I expanded on previous phylogenetic inference with new evidence based on nuclear sequences from five gene families using gene tree parsimony (GTP). GTP inferred the rooted species trees that were most compatible with the five nuclear gene families through fitting each genealogy to species trees and by minimizing the number of duplications needed to reconcile conflicts among them. My results were congruent with earlier species tree estimates inferred using plastid data. I conducted new analyses that placed the root of the family at a branch leading to the rare and geographically restricted *Eichhornia meyeri*. I also developed bootstrap-like methods to incorporate uncertainty in individual genealogies into reconciliation. This study contributes to our understanding of the phylogenetic history of Pontederiaceae and has also demonstrated the utility of GTP for phylogenetic analysis.

**Future research**

My thesis has addressed a number of questions related to the consequences of mating-system variation for the population genetics of *E. paniculata*. In addition, I have generated valuable genetic resources that should aid future in-depth studies on the genetic basis of floral traits associated with the breakdown of tristyly and the evolutionary transition from outcrossing to selfing. Next I briefly outline potential future research on the evolutionary genetics of mating-system variation in *E. paniculata*.

**Architecture of the selfing syndrome**

The evolution of selfing from outcrossing is characterized by a series of morphological and physiological changes culminating in the selfing syndrome. However, which traits initiate increased selfing and which are accumulated after self-fertilization has
established is still poorly understood. *Eichhornia paniculata* presents an excellent opportunity to explore this problem because of the well-described breakdown of tristyly to semi-homostyly. In *E. paniculata*, reduction in herkogamy is the principal morphological change initiating the evolutionary transition to selfing. This modification has been shown to be under the control of recessive modifiers (Fenster and Barrett 1994; Vallejo-Marín and Barrett 2009; Chapter 3). As discussed in Chapter 3, there are two distinct pathways for the evolution of selfing involving modification of the L- or M-morph. The continuous distribution of anther-stigma distance observed in the F2 generation indicates that herkogamy is a polygenic trait in the L-morph (see Chapter 3). In contrast, there appear to be fewer genes of major effect controlling herkogamy in selfing variants of the M-morph. This difference could be because the genetic architecture underlying these transitions to selfing may be fundamentally different in the two cases. Evolution of the semi-homostylyous L-morph from Central American may have evolved by gradual selection on quantitative variation affecting mating system traits in the absence of the M-morph in this region. In contrast, elsewhere, shifts to selfing may be more rapid because of selection of major recessive genes in the M-morph. It is striking that the L-morph occurs in dimorphic populations containing selfing variants of the M-morph in N.E. Brazil, Jamaica and Cuba but is unmodified and incapable of autonomous self-pollination. Perhaps selfing is only possible in the L-morph in the absence of the M-morph, as proposed by Barrett (1996). Future studies using quantitative trait locus (QTL) mapping could be employed to dissect the genetic architecture of selfing in the semi-homostylyous variants of the L- and M-morph of *E. paniculata*. Selection of parental strains and markers directed by information gained through the transcriptomes sequences should allow the identification of genomic regions that influence traits associated with the selfing syndrome. The proportional contribution of genes...
of large versus small effect to the evolution of mating system remains poorly understood (but see Fishman et al. 2002), but this information is crucial for our understanding of the various stages in the evolution of the selfing syndrome.

In addition to investigations of the genetic architecture of herkogamy modification in *E. panculata*, QTLs that control other floral and inflorescence characteristics associated with stages in the breakdown of tristyly including small, unshowy flowers with less intense floral pigmentation, reduced pollen heteromorphism and pollen production, lower nectar volume and smaller inflorescences (see Barrett 1988; Morgan and Barrett 1989) could be investigated. These traits evolve gradually after the initial rapid selection for selfing variants with reduced herkogamy. A linkage map would provide a valuable resource for future genetic studies, especially with respect to identifying regions of the genome that may be subject to adaptive evolution with respect to selfing. *Eichhornia paniculata* is a good candidate for genetic mapping because rarely is it possible to observe the full spectrum of mating system and floral phenotype variation within a single species. Genetic mapping would be the first step to identifying the causal loci underlying changes associated with selfing and could help in the initial search for the genes governing the control of tristyly.

*Genetics of adaptation*

Understanding the genetic basis of adaptation is one of the central goals of ecological and evolutionary genetics. Recent technological and analytical innovations have resulted in considerable progress in this area as genes and gene regions underlying ecologically and evolutionary relevant traits are being identified (reviewed in Stinchcombe and Hoekstra 2008). This type of research addresses the following types of questions: what evolutionary and ecological processes affect patterns of variation in genes responsible for adaptation? Are
there multiple genetic mechanisms to create the same functional phenotype? In what follows I will outline two traits observed in *Eichhornia paniculata* that present intriguing candidates for further study: (1) The genes associated with the reduction of stigma-anther distance in selfers; (2) The genes controlling inheritance of the three floral morphs that characterize tristyly.

With the exception of the breakdown of sporophytic self-incompatibility in Brassicaceae (e.g. Tang et al. 2007) the sequence-level changes in genes that control selfing are not known. Reduction in herkogamy is the principal adaptation initiating the evolutionary transition to selfing in *E. paniculata* (Vallejo-Marín and Barrett 2009). Previous evidence from crosses between selfing variants of the M-morph from two separate Brazilian populations indicated that different recessive modifier loci are involved (Fenster and Barrett 1994). Identifying distinct alleles or genes that confer reduced herkogamy provides additional evidence for multiple independent origins of selfing in *Eichhornia paniculata*. Determining the patterns of sequence variation in these genes could provide valuable insights into the strength of selection on alleles governing the breakdown of tristyly. The identity of these loci may shed light on why stamen length seems so evolutionarily labile in *E. paniculata* and why there appears to be many possible genetic ways to modify herkogamy. If these genes were identified it would be possible to address a host of questions such as: (1) Are the modifiers regulatory elements which control expression or structural mutations to proteins? (2) Are the modifiers the results of new mutations in genes or supergenes involved in the genetic control of heterostyly, or are they part of the standing genetic variation of populations segregating at low levels in tristyloous populations? Answers to these questions could provide important insights into the genetic mechanisms governing the transition from outcrossing to selfing.
Although the inheritance of style morph in heterostylosus species is well established, the identity, number, and organization of genes regulating tristyly have not been determined (reviewed in Barrett and Shore 2008). The inheritance of tristyly in *E. paniculata*, as in other tristylosus species, is controlled by two linked epistatically interacting diallelic loci (reviewed in Chapter 1). However, whether each of these loci are supergenes involving a series of linked genes that function in combination, or instead are regulatory genes controlling the expression of phenotype by coordinating other genes is an open question. Current efforts are being made to identify the genetic basis of heterostyly using protein differences, differential gene expression and genetic localization (reviewed in Barrett and Shore 2008). To determine the genetic architecture of tristyly in *E. paniculata* requires a detailed genetic map of an F2 population derived from crosses made between different morphs. Depending on the resolution of the genetic map, markers that localize to regions associated with style morph identity could be used to probe a BAC library and the positive clones could then be sequenced. Molecular population genetic approaches could then be used to screen BAC sequence for regions subject to balancing selection, the signature expected when alleles at each of the two loci are maintained by negative frequency-dependent selection (Charlesworth 2006). If successful this approach would identify genes linked to the *S* or *M* loci. These candidate loci could then be filtered using bulk segregant analysis of mapping populations (Michelmore et al. 1991) or by association mapping (e.g. Atwell et al. 2010) of samples collected from natural populations. Identification of these genes would be a major advance in the ongoing research on heterostyly, informing previous theoretical investigations (Charlesworth 1979), and also in helping to elucidate the forces governing the evolutionary assembly of this complex sexual polymorphism. Heterostyly has been a model for genetic and evolutionary analysis since Darwin’s seminal work (reviewed in Barrett 1992;
Charlesworth and Charlesworth 2009) but as yet the specific genes governing the polymorphism have yet to be discovered.

*Population genomics of mating system*

The recent rapid advances in sequencing technology have made it possible to sequence orders of magnitude more DNA than was possible at the beginning of my Ph.D. research. In Chapter 6, I pointed out that our ability to infer patterns of selection across the genome in selfers and outcrossers of *E. paniculata* was limited by the absence of accurate estimates of polymorphism from populations. Deep sequencing would allow analysis of the fundamental forces shaping genome evolution including selection, drift, mutation, recombination and their interactions. For example, extensive population-level sampling of large tracts of genome sequence would enable accurate estimates of linkage disequilibrium (LD). The effective rate of recombination should be reduced in selfing populations (Nordborg 2000), and although we found some evidence for this effect in Chapter 4, the signal was weak and limited by the relative short length of the gene sequences used. Large segments of the genome possible with current sequencing technologies would be ideal to estimate the breakdown of LD in populations with variable mating patterns. Regions with reduced recombination should have depressed genetic diversity as a result of genetic hitchhiking (Maynard Smith and Haigh 1974; Kaplan et al. 1989). Testing these predictions with genome-wide polymorphism data would allow more rigorous investigations of the interactions among mating system, effective recombination rate and the efficacy of selection.

The generation of population genomic datasets would also allow investigation into the extent to which genomes are shaped by neutral processes, purifying selection, and adaptive evolution. The relative importance of these factors for genome evolution has been
debated since the development of the neutral theory (Kimura 1983). Recently, analytical
techniques for estimating the strength, direction and extent of selection on the genome have
been devised using patterns of polymorphism and divergence (McDonald and Kreitman,
1991), the site frequency spectrum (Tajima 1989; Fu and Li 1993; Fay and Wu 2000) and the
joint frequency spectrum (Li and Stephan 2006). These approaches should shed light on the
interaction of mating system and $N_e$ in affecting rates of adaptive evolution, deleterious
mutation and drift on the genome (Boyko et al. 2008; Eyre-Walker and Keightley 2009).
Moreover, population genomic data could be integrated with mapping experiments to
identify regions controlling functionally relevant traits. These analyses may confirm some of
the candidate genes that were differentially expressed in the selfing transcriptomes
characterized in Chapter 5. The relative importance of positive selection for autogamy,
changes in floral resource allocation, and relaxed selection are in explaining the loss of
outcrossing adaptations and convergence in phenotypes of the selfing variants is largely
unknown at this point.

Conclusion

The breakdown of tristyly to selfing in *E. paniculata* has been identified by Coyne et
al. (1997) as one of the very few known cases that meet several key conditions of Sewall
Wright’s shifting balance theory of evolution because of the joint interaction of genetic drift
and natural selection in promoting the shift to selfing. *Eichhornia paniculata* has been used
as a model system for investigating diverse questions in ecological genetics and reproductive
biology resulting in 37 journal articles over the past 25 years. Future population genomic
studies building on this thesis could provide further insights into the relative importance of
drift and selection in causing changes in mating systems. I hope that by initiating the first
molecular population genetic and genomic studies of *E. paniculata*, my thesis, and the publications that arise from it, will invigorate and broaden the past 25 years of research on this fascinating system.


Lohmueller, M.D. Adams, S. Schmidt, J.J. Sninsky, S.R. Sunyaev, T.J. White, R.


