MOLECULAR GENETICS AND POPULATION GENOMICS OF THE MAINTENANCE
AND BREAKDOWN OF THE FLORAL POLYMORPHISM TRISTYLY

by

Arunkumar Ramesh

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

© Copyright by Ramesh Arunkumar 2016
MOLECULAR GENETICS AND POPULATION GENOMICS OF THE MAINTENANCE
AND BREAKDOWN OF THE FLORAL POLYMORPHISM TRISTYLY

Ramesh Arunkumar

Doctor of Philosophy, 2016

Department of Ecology and Evolutionary Biology

University of Toronto

Abstract

Tristyly has evolved independently in several flowering plant families and functions to promote outcrossing by disassortative mating. The floral polymorphism is susceptible to evolutionary breakdown resulting in frequent transitions from outcrossing to predominant self-fertilization. By integrating experimental studies with analyses of genomic data and population genetic computer simulations, my Ph.D. thesis investigates the genetic and population genomic basis for the maintenance of this genetic polymorphism, and its breakdown to selfing in the annual aquatic Eichhornia paniculata (Pontederiaceae).

The transition from outcrossing to selfing in E. paniculata is associated with multiple, independent, long-distance colonization events from N.E. Brazil to the Caribbean and Central America, and a reduction in the effective population size of selfing lineages. The joint influence of these genetic and demographic factors was associated with several genomic consequences, including an increased frequency of effectively neutral mutations, more efficient removal of very strongly deleterious mutations, and the down-regulation of ~1000 genes. These findings suggest
that substantial differences in patterns of mutation and gene expression may accompany transitions from outcrossing to selfing in plants.

Genetic crosses and progeny tests conducted under glasshouse conditions demonstrated strong linkage between the $S$ and $M$ loci governing tristyly, with the loci separated by a map distance of 2.7 cM. Controlled crosses using semi-homostyous mid- and long-styled selfing variants from the Caribbean ($ssMM$) and Central America ($ssmm$) combined with genetic mapping of a backcross progeny indicated that style length and anther height localized to the center of linkage group 5. This finding strongly suggests that this region contains the $M$ locus. Population genomic data was consistent with the hypothesis that the $M$ locus is maintained by frequency-dependent selection. Finally, QTL mapping studies also demonstrated that the two semi-homostyous variants of $E. paniculata$ from the Caribbean and Central America were governed by different sets of modifier genes unlinked to the $M$ locus supporting their independent origins. My thesis research provides novel insights into the biology of tristyly, genetic architecture of reproductive adaptation, and the genomic consequences of mating system transitions. It provides a contemporary genomic perspective complementing the earlier seminal studies of Darwin and Fisher on tristyly.
Acknowledgements

I thank my supervisors Spencer C.H. Barrett and Stephen I. Wright whose mentorship was crucial in my personal development as a researcher. I would like to thank Stephen Wright for welcoming me as undergraduate into his research group based solely on my early interests in the study of genomics and evolution and, later, introducing me to the scientific method, shaping my research and career directions and being an extremely accessible and friendly mentor. I would like to thank Professor Spencer Barrett for imparting his infectious enthusiasm for plant sciences and evolutionary biology and his patient guidance in developing my verbal communication skills and for the many hours spent improving my writing. I also thank Spencer for introducing me to heterostyly and my study organism – *Eichhornia paniculata* – and for much advice ranging from the natural history of Pontederiaceae to academic professionalism. The plant material that was the basis of my thesis was collected from several field trips to N.E. Brazil, Jamaica and Cuba by Spencer Barrett, Willam Cole and Robert Ness and I thank Bruce Hall and Andrew Petrie for helping to maintain these collections in the glasshouses at the University of Toronto. I am also grateful to John Stinchcombe and Asher Cutter, members of my supervisory committee, who were both helpful in developing novel ideas and framing the broader significance of my research work. I would like to acknowledge Aneil Agrawal and Brian Charlesworth (University of Edinburgh) for proving extensive and constructive feedback in interpreting the results of my studies on the genomic consequences of self-fertilization. Finally, I thank Jeremiah Busch (Washing State University) for his willingness to serve on my Ph.D. defense committee.

I would like to acknowledge many colleagues in the Barrett, Wright, Agrawal, Stinchcombe and T. Sage research groups among others in the highly interactive Ecology and
Evolutionary Biology Department (EEB) at the University of Toronto for their friendship and numerous and varied contributions to making my time in the department as a graduate student an enjoyable experience. My fellow graduate students Andrew Simpson, Eddie Ho, Yuheng Huang, Christopher Balogh, David Timerman, Robert Williamson and Gavin Douglas assisted in maintaining my sanity as well as with plant care, analytic protocols and dissertation writing. More significantly, I cherish the strong personal ties I have developed with them during my tenure as a graduate student. I would like to thank William Cole for the extensive time and help he has devoted to plant care and attaining the necessary molecular equipment but, more importantly, being the unofficial mentor in several projects and a voice of support and comfort. I would like to thank Rob Ness, Emily Josephs and Wei Wang for assistance with numerous bioinformatics analyses and Khaled Hazzouri, Young Wha Lee and Epitawalage Niroshini for assisting with several ‘wet lab’ protocols. I would like to thank the numerous undergraduate students who provided me support in performing several of the projects in this thesis, particularly Teresa Maddison, Minkyu Kim, Allison Grey, Arielle Beaudry and Ramana Trivedi. Finally, I would like to thank my parents Chitra Sadasivan and Ramesh Kuppuswamy who assisted with the dissertation writing but also in countless other ways in support of my research ambitions.

The research conducted in this thesis was funded by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Spencer Barrett and Stephen Wright, and research funds from a Canada Research Chair to Spencer Barrett. I would also like to thank NSERC, the Ontario Ministry of Training, Colleges and Universities and the University of Toronto for student fellowships that supported me while I was a graduate student at the University of Toronto.
# Table of contents

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

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

Table of contents ............................................................................................................................vi

List of tables .....................................................................................................................................xii

List of figures ...................................................................................................................................xvi

List of appendices..........................................................................................................................xxiv

Chapter 1: General introduction.................................................................................................1

The transition from outcrossing to predominant selfing. .........................................................3

The effects of selfing on the genome............................................................................................5

Genetic architecture of heterostyly and its breakdown to selfing...........................................7

The study system *Eichhornia paniculata*....................................................................................9

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

Chapter 2: The demography and population genomics of evolutionary transitions to self-
fertilization in plants....................................................................................................................17

Summary........................................................................................................................................17

Introduction.....................................................................................................................................18

Selection of self-fertilization..........................................................................................................21

Modification to reproductive traits that promote selfing........................................................24
Molecular functional analyses.................................................................120

Calculating expression intensity............................................................120

Comparing nonsynonymous and synonymous diversity..........................121

Results......................................................................................................122

Allele-specific expression........................................................................122

Expression fold difference.......................................................................124

Relative proportions of nonsynonymous and synonymous mutations........126

Discussion..............................................................................................128

The evolution of selfing is accompanied by differences in cis-regulation .....128

Patterns of ASE and association to the evolution of the selfing syndrome.....131

Chapter 5: Genetic linkage between the S and M loci governing tristyly in *Eichhornia paniculata* (Pontederiaceae).........................................................................................142

Summary...............................................................................................142

Introduction...........................................................................................143

Materials and methods..........................................................................145

Test crosses............................................................................................145

Inferring genotypes and estimating linkage...........................................147

Results.....................................................................................................149

Discussion.............................................................................................151

Linkage between the S and M locus and sheltered load at the S locus........151
Consequences of linkage for the dynamics of style morph ratios in populations.................................................................152

Variation in linkage relations among tristylos species.................................153

Chapter 6: The genetic architecture of style morph variation in tristylos *Eichhornia paniculata* and modifiers associated with evolutionary transitions to self-fertilization.................................158

Summary..........................................................................................................................158

Introduction......................................................................................................................159

Materials and methods.....................................................................................................166

F1 and F2 backcross populations........................................................................166

Floral measurements........................................................................................................166

Genotyping-by-sequencing..................................................................................167

Composite interval mapping................................................................................168

Identifying genes within QTLs............................................................................168

Population genetic estimates................................................................................169

Bulk segregant analysis........................................................................................170

Candidate SNPs for style-length variation...........................................................171

Results..............................................................................................................................172

Phenotypic variation of the L' and M' progeny...................................................172

Genomic reference and genetic map......................................................................174

Interval mapping of the M locus............................................................................175
List of tables

Table 2S1. Survey of diversity measurements for outcrossing-selfing pairs.................................42

Table 2S2. Survey of Tajima’s D measurements and bottleneck signals for outcrossing-selfing pairs....................................................................................................................................45

Table 2S3. Survey of estimated age of selfing and selfing rate estimates for outcrossing-selfing pairs....................................................................................................................................48

Table 2S4. Literature sources for estimates reported in Tables S2.1-2.3......................................49

Table 3S1. The sample of *Eichhornia paniculata* individuals used for estimating the efficacy of selection in outcrossing and selfing populations.............................................................100

Table 3S2. Polymorphisms segregating in *Eichhornia paniculata* populations identified in aligned portions of EST-derived nuclear locus EP0314 from Barrett et al. (2009) and Ness et al. (2010) and the best matching contig from this study........................................102

Table 3S3. Nonsynonymous and synonymous diversity of simulated outcrossing and selfing populations under fixed dominance levels for all mutations..............................................105

Table 3S4. P-values from a two-sample t-test comparing \(N_e s^a\) categories of simulated outcrossing and selfing populations under fixed dominance levels for all mutations.........................105

Table 3S5. Two-way analysis of variance comparing the \(N_e s^a\) categories of simulated outcrossing and selfing populations under varying dominance for mutations.........................................106

Table 3S6. Counts of nonsynonymous and synonymous invariant sites and minor allele frequency spectra bins summed across loci for eight outcrossing and eight selfing populations of *Eichhornia paniculata*........................................................107
Table 3S7. Number of loci containing sites heterozygous across multiple selfing populations from the Caribbean..........................107

Table 3S8. $P$-values from randomization tests as applied in Keightley and Eyre-Walker (2007) comparing the $N_e < 1^a$ category between DFEs of eight outcrossing and eight selfing Eichhornia paniculata populations under various filtering conditions..................108

Table 4S1. Population codes and localities for parental genotypes of Eichhornia paniculata used in the study.................................................................137

Table 4S2. Number of sites homozygous for alternative SNPs in outcrossing and selfing parents and heterozygous in F1 progeny.........................................................138

Table 4S3. Functional classification for genes with posterior probability for allele-specific expression $>0.7$ sampled from floral buds and leaf tissue of F1 plants from crosses between selfing (maternal) and outcrossing (paternal) parents of Eichhornia paniculata.................................................................139

Table 4S4. $P$-values from comparisons of expression level differences between outcrossing and selfing Eichhornia paniculata parents for the set of genes showing allele-specific expression in the F1 offspring.................................................................141

Table 4S5. Diversity at nonsynonymous and synonymous sites for Eichhornia paniculata floral bud and leaf genes in outcrossing parents from Brazil and selfing parents from the Caribbean.................................................................141

Table 5.1. Locality and morph structure of the four trimorphic Eichhornia paniculata populations in N.E. Brazil from which the plants of the S-morph used in this study originated via subsequent crossing.................................................................145
Table 5.2. Expected segregation of style morphs in progeny from selfs and crosses involving short-styled (S) and long-styled (L) homozygous recessive individuals (sm/sm) under various genotypes for the S-morph parent...

Table 5.3. Observed segregation of style morphs in *Eichhornia paniculata* progeny from selfs and crosses involving short-styled (S) individuals suspected of being double heterozygotes (SM/sm or Sm/sM) and long-styled (L) individuals (sm/sm)...

Table 5.4. Studies on the inheritance of tristyly and degree of linkage between the S and M loci in three unrelated plant families...

Table 6S1. Summary metrics of genomic assembly for *Eichhornia paniculata*...

Table 6S2. Number of markers and genes within the interval on linkage group 5 that had high LOD scores following composite interval mapping of style length, and number of genes with candidate SNPs suggesting linkage to the style-morph locus in that interval in F2 *Eichhornia paniculata* progeny...

Table 6S3. Number of genes with candidate SNPs suggesting linkage to the style morph locus in each linkage group in *Eichhornia paniculata* F2 progeny...

Table 6S4. Percent of variation in height of anthers in the L’ semi-homostyle explained by each identified QTL in *Eichhornia paniculata* F2 progeny...

Table 6S5. Number of markers and genes within the interval on linkage group 2 that had high LOD scores following composite interval mapping of anthers in the L’ semi-homostyle of *Eichhornia paniculata* F2 progeny...
Table 6S6. Number of markers and genes found within the interval on linkage group 4 that had high LOD scores following composite interval mapping of the M' semi-homostyle anthers in *Eichhornia paniculata* F₂ progeny.................................................................210

Table 6S7. Number of markers within the interval on linkage group 3 that had high LOD scores following composite interval mapping of anther height in *Eichhornia paniculata* F₂ progeny.................................................................210
List of figures

Figure 1.1. The heterostyous floral polymorphisms: (a) distyly and (b) tristyly. L, M and S refer to the long-, mid- and short-styled morphs, respectively.....................................................2

Figure 1.2. Theoretical expectations of the genome-wide reduction in selection efficacy associated with the transition for outcrossing to predominant selfing........................................5

Figure 1.3. The study system *Eichhornia paniculata*.............................................................................11

Figure 1.4. Evolutionary pathways from cross-fertilization to self-fertilization in tristylos
*Eichhornia paniculata* via the origins of semi-homostyly.........................................................11

Figure 2.1. Expected reduction in genetic diversity in selfing compared to outcrossing populations with selfing (solid line), selfing and background selection (long dashed line), and with selfing, background selection and population bottlenecks (short dashed line)...30

Figure 2.2. The reduction in synonymous diversity ($\theta_w$) as a function of time since the transition to selfing, in units of $N$ generations, where $N$ is the effective population size of the ancestral outcrossing populations..........................................................................................................................32

Figure 2.3. Ratio of synonymous diversity ($\theta_s$) for related selfing and outcrossing populations and species in relation to: (a) the rate of selfing in populations with the higher degree of self-fertilization and (b) the estimated age (in years) of the transition to higher rates of selfing.................................................................................................................................36

Figure 2.4. Shared and unique sequence polymorphisms between outcrossing and selfing populations of *Eichhornia paniculata* sampled from different parts of the geographical range.................................................................................................................................39
Figure 3.1. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing \((t=1.00)\) and selfing \((t=0.02)\) populations under various fixed dominance coefficients \((h=0.2, 0.5, \text{or} \ 0.8)\) .......................................................................................................................... 69

Figure 3.2. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing populations of two census sizes \((N)\) under various fixed dominance coefficients ......................................................................................................................... 71

Figure 3.3. The type and effects of mutations accumulated by four outcrossing (black) and four selfing (grey) individuals over \(6N\) generations after a split from a common outcrossing ancestor ........................................................................................................................................................................... 73

Figure 3.4. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing \((t=1.00)\) and selfing \((t=0.02)\) populations when both dominance and selection coefficients were jointly varied ........................................................................................................................................................................... 74

Figure 3.5. Folded nonsynonymous and synonymous allele frequency spectra for outcrossing and selfing populations of Eichhornia paniculata .......................................................................................................................... 76

Figure 3.6. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing Eichhornia paniculata .......................................................................................................................... 77

Figure 3S1. Genome structure implemented in the simulations .......................................................................................................................... 86

Figure 3S2. Length distribution of all contigs from the de novo assembly of six Jamaican genotypes of Eichhornia paniculata using Velvet-Oases .......................................................................................................................... 87

Figure 3S3. Coverage distribution of genomic reads from one selfing Mexican genotype of Eichhornia paniculata mapped to the de novo transcriptome reference .......................................................................................................................... 87

xvii
Figure 3S4. Folded synonymous allele frequency spectra for simulated outcrossing (t=1.00) and selfing (t=0.02) populations under various fixed dominance coefficients (h=0.2, 0.5, or 0.8).....................................................................................................................................88

Figure 3S5. Distribution of fitness effects (DFE) of new nonsynonymous mutations or simulated outcrossing (t=1.00) and selfing (t=0.02) populations under various fixed dominance coefficients (h=0.2, 0.5, or 0.8) when both mutations unique to each population and mutations shared between populations were used to estimate selective and demographic parameters.....................................................................................................................................89

Figure 3S6. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing (t=1.00) and selfing (t=0.02) populations under various fixed dominance coefficients (h=0.2, 0.5, or 0.8) when synonymous mutations occurred only within coding regions were used to estimate demographic parameters.................................................................90

Figure 3S7. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing populations with recombination rate (r) of either 5×10^{-8} or 1.96×10^{-9} per site per generation under various fixed dominance coefficients.................................................................91

Figure 3S8. Distribution of polymorphisms per contig for eight outcrossing Eichhornia paniculata populations from N.E. Brazil and eight selfing populations from the Caribbean.................................................................................................................................................92

Figure 3S9. Cumulative distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of Eichhornia paniculata.................................................................93
Figure 3S10. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* under varying quality cut-off scores (QUAL) for invariant and variant sites.................................................................94

Figure 3S11. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* after 245 loci containing sites heterozygous in two or more selfing populations were excluded from analyses.....95

Figure 3S12. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* after 73 loci containing sites that violated assumptions of Hardy-Weinberg equilibrium in outcrossing populations were excluded from analyses.............................................................................................96

Figure 3S13. Distribution of $\pi_n/\pi_s$ for polymorphic contigs among eight outcrossing *Eichhornia paniculata* populations from N.E. Brazil and eight selfing populations from the Caribbean. Of the 16416 loci, 4485 were polymorphic in outcrossers and 1586 were polymorphic in selfers........................................................................................................97

Figure 3S14. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* after 135 loci with $\pi_n/\pi_s > 1$ in either population were excluded from analyses.............................................................97

Figure 3S15. Folded nonsynonymous and synonymous allele frequency spectra for outcrossing and selfing populations of *Eichhornia paniculata*.................................................................98

Figure 316. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata*.........................................................99
Figure 4.1. Distribution of posterior probabilities for genes showing allele-specific expression (ASE) in *Eichhornia paniculata*..................................................................................................................123

Figure 4.2. Distribution of loge-fold expression difference between outcrossing and selfing alleles of genes with posterior probability for allele-specific expression $> 0.7$ in F$_1$ plants of *Eichhornia paniculata*........................................................................................................125

Figure 4.3. Diversity at nonsynonymous ($\pi_{\text{nonsynonymous}}$) and synonymous sites ($\pi_{\text{synonymous}}$) for genes with posterior probability $> 0.7$ for allele-specific expression (ASE), and for genes without evidence of ASE (posterior probability $< 0.7$) in *Eichhornia paniculata*.........126

Figure 4S1. Number of reads that mapped to reference assembly from (a) floral bud and leaf parental transcriptomes and (b) floral bud and leaf F$_1$ offspring transcriptomes of *Eichhornia paniculata*.........................................................................................................................133

Figure 4S2. Distribution of posterior probabilities that genes show allele-specific expression (ASE) under various $a\ hat$ and $d\ hat$ parameters in *Eichhornia paniculata*.........................134

Figure 4S3. Distribution of loge-fold expression difference between outcrossing and selfing alleles for ASE and non-ASE genes calculated using (a) only sites where the outcrossing allele matched the reference allele, (b) only sites where the selfing allele matched the reference allele or (c) all sites........................................................................................................................................135

Figure 4S4. Distribution of loge-fold gene expression difference between outcrossing and selfing *Eichhornia paniculata* parents for the set of genes with posterior probability for allele-specific expression $> 0.7$ in F$_1$ offspring.........................................................................................................................136

Figure 6.1. Mean anther heights of 227 L’ and 235 M’ F$_2$ *Eichhornia paniculata* progeny......173
Figure 6.2. Genetic map of *Eichhornia paniculata*; the number of markers in each linkage group is indicated on the top of the diagram.................................................................174

Figure 6.3. Composite interval mapping of: A) style length variation and B) variation in the anthers of the upper-level stamens in 227 L’ and 235 M’ F2 backcross progeny of *Eichhornia paniculata*......................................................................................................176

Figure 6.4. Patterns of polymorphism and Tajima's $D$ for *Eichhornia paniculata* genes found within linkage group 5.................................................................177

Figure 6.5. LOD scores on linkage groups 2 and 4 following composite interval mapping of anther heights of L’ and M’ variants in F2 progeny of *Eichhornia paniculata*..............180

Figure 6.6. LOD scores on linkage group 3 of *Eichhornia paniculata* following composite interval mapping of anther heights of the L’ and M’ semi-homostylos morphs in F2 progeny.................................................................183

Figure 6S1. Phenotypic measurements of the flowers of the L’ and M’ semi-homostyles of *Eichhornia paniculata*.........................................................................................................................194

Figure 6S2. Plant height and flower size of semi-homostylos morphs in the backcross F2 of *Eichhornia paniculata* progeny........................................................................................................194

Figure 6S3. Principle component analysis of correlations among plant height, flower size, style length and anther height of 227 and 235 L’ and M’ F2 progeny, respectively, of *Eichhornia paniculata*. .................................................................195

Figure 6S4. Number of genomic reads that mapped to the genomic reference assembly for the L’ (ssmm) and M’ (ssMm) parental genotypes of *Eichhornia paniculata*.........................196
Figure 6S5. Effect plot for the marker with the highest LOD score on linkage group 5 identified during interval mapping of style length in F2 progeny of *Eichhornia paniculata* ........196

Figure 6S6. Comparison of composite interval mapping of the style length of L´ and M´ semi-homostyles in F2 progeny of *Eichhornia paniculata* using the imputation method, expectation-maximization algorithm and extended Haley-Knott regression .................197

Figure 6S7. Distribution of variant and invariant genes along the centre of linkage groups 3 and 7 in *Eichhornia paniculata* ..................................................................................................................197

Figure 6S8. Recombination rates between adjacent sites for genes found along the genomic interval on linkage group 5 of *Eichhornia paniculata* ........................................................................................................198

Figure 6S9. Functional enrichment of *Eichhornia paniculata* genes with candidate SNPs suggesting linkage to the style-morph locus .................................................................................................199

Figure 6S10. Effect plots for the marker with the highest LOD score on linkage group 2 identified during interval mapping of anther height for the upper three anther levels: A) 4, B) 5, and C) 6, in the L´ semi-homostyle of *Eichhornia paniculata* in F2 progeny ....200

Figure 6S11. Effect plots for the marker with the highest LOD score on linkage group 4 identified during interval mapping of the anther height for modified anther levels: A) 2, B) 3, in the M´-morph and the anther height: A) 2, B) 3, in the L´ semi-homostyle of *Eichhornia paniculata* in F2 progeny ........................................................................................................201

Figure 6S12. Comparison of composite interval mapping of anther 6 of the L´ semi-homostyle of *Eichhornia paniculata* in F2 progeny using the imputation method (IMP), expectation-maximization algorithm (EM) and extended Haley-Knott regression (EHK) ...........202
Figure 6S13. Comparison of composite interval mapping of the second lowest anther of the M′ semi-homostyle of *Eichhornia paniculata* in F$_2$ progeny using the imputation method (IMP), expectation-maximization algorithm (EM) and extended Haley-Knott regression (EHK)...............................................................................................................................203

Figure 6S14. Composite interval mapping of anther heights of the L′ semi-homostyle of *Eichhornia paniculata* in F$_2$ progeny. Shown are LOD scores for: A) lowest three anther levels along linkage group 1 and B) linkage group 6 and C) upper three anther levels along linkage group 1 and D) linkage group 6.................................................................204
List of appendices

Appendix I: Supporting information for Chapter 2.................................................................42

Appendix II: Supporting information for Chapter 3..............................................................86

Appendix III: Supporting information for Chapter 4.............................................................133

Appendix IV: Supporting information for Chapter 6.............................................................190
Chapter 1

General introduction

Floral diversification is commonly associated with extensive mating system variation in flowering plants. Angiosperms are therefore excellent model systems for a range of questions concerning the evolutionary genetics of mating system transitions. Hermaphroditism occurs in the majority of angiosperms and provides an opportunity to investigate the ecological and genetic consequences of the relative frequencies of cross- and self-fertilization (Darwin 1876, 1877; Jain 1976; Uyenoyama et al. 1993; Lloyd 1992; Wright et al 2013). Among the numerous hermaphroditic mating strategies is the floral polymorphism heterostyly in which plant populations are characterized by two or three style morphs with a reciprocal arrangement of stigmas and anthers (Darwin 1877; Ganders 1979; Barrett 1992). Compatible mating generally occurs between style morphs with anthers and stigmas of equivalent height and these are termed “legitimate mating” (Figure 1.1). Darwin (1877) provided the first adaptive hypothesis for the function of heterostyly. He proposed that sex-organ reciprocity promotes animal-mediated cross-pollination, and in so doing reduces pollen wastage on incompatible stigmas. Current empirical evidence supports the Darwinian hypothesis (reviewed in Lloyd and Webb 1992a). The heterostylosus syndrome has originated at least 23 times among 19 orders of flowering plants (Lloyd and Webb 1992b) and represents a striking example of pollinator-driven floral convergence. Although much is known about the comparative biology, reproductive ecology and inheritance of the polymorphism (reviewed in Barrett and Shore 2008), much remains to be understood concerning the genetic architecture and molecular basis of heterostyly, and also the breakdown of the polymorphism to self-fertilization.
Heterostylyous populations are usually comprised of two (distyly) or three (tristyly) mating morphs that can be distinguished by sex-organ reciprocity. In distyloous populations, there are long- and short-styled morphs (hereafter L- and S-morphs) whereas tristyloous populations contain a third style morph, the mid-styled morph (M-morph). Negative-frequency dependent selection maintains the style morphs at equal ratios (isoplethy) in populations at equilibrium (Fisher 1944). Disassortative (intermorph) mating is enforced in most heterostylyous populations due to the
operation of a heteromorphic (dimorphic or trimorphic) incompatibility system that prevents selfing and intramorph mating. However, in numerous heterostyous taxa, predominant selfing has evolved owing to the breakdown of the polymorphism. This evolutionary change results from the spread of homostyous (distyly) and semi-homostyous (tristyly) morphs which are usually self-compatible and possess stigmas and anthers at the same height within a flower (Darwin 1877; Ganders 1979; Charlesworth and Charlesworth 1979b; Barrett 1989). Evidence from biogeography and more recently phylogenetic reconstruction and character mapping has documented this transition from outcrossing to selfing in multiple heterostyous taxa (e.g. *Primula* - De Vos et al. 2004; *Eichhornia* - Kohn et al. 1996; *Amsinckia* - Schoen et al. 1997; *Turnera* - Truyens et al. 2005). Heterostyly therefore offers an outstanding model system to investigate the genetic basis and genomic consequences of evolutionary transitions from outcrossing to selfing in flowering plants and this is a topic I investigate in this thesis.

**The transition from outcrossing to predominant selfing**

An attractive aspect of heterostyly from the viewpoint of mating-system evolution is that there is good evidence for multiple independent evolutionary breakdown of the polymorphisms resulting in transitions from outcrossing to self-fertilization both within and between species (reviewed in Barrett 1989; Weller 1992). In distyous *Primula* (Dowrick 1956; Charlesworth and Charlesworth 1979b) and *Turnera* (Barrett and Shore 1987; Shore and Barrett 1985), recombination within the S locus produces homostyles that lack stigma-anther separation resulting in a high degree of autogamy. In tristyous *Eichhornia paniculata*, semi-homostyous variants with one set of anthers close to the stigma result from the spread of modifier genes that modify stamen height (Fenster and Barrett 1993; Barrett et al. 2009). Transitions from
outcrossing to selfing in flowering plants are associated with diverse changes to floral morphology, sex allocation and life history (Sicard and Lenhard 2011), and these can have significant demographic, genomic and evolutionary consequences (Busch and Delph 2012; Wright et al 2013; Igic and Busch 2013; Chapter 2). The multiple breakdown of the tristyloous polymorphism can therefore serve as useful testing ground for investigating the genetic basis of changes in mating system.

The causes and consequences of the evolution of selfing have generated much theoretical and empirical interest. Selfing species are generally short-lived with only 10-15% of angiosperms that are thought to be highly autogamous. The loss of advantageous heterozygous genotypes and the exposure of recessive deleterious mutations generally results in fitness reductions of inbred compared to outbred progeny (Darwin 1876; Lande and Schemske 1985, Charlesworth and Charlesworth 1987; Charlesworth and Willis 2009). However, there have been numerous independent origins of selfing from outcrossing across unrelated herbaceous taxa (Stebbins 1957; Schoen et al. 1997; Takebayashi and Morrell 2001; Moeller and Geber 2005; Igic et al. 2008; Busch et al. 2011; Igic and Busch 2013; Wright et al. 2013; Chapter 2). There are two main hypotheses to explain the evolution of selfing in flowering plants (reviewed in Busch and Delph 2012). First, selfers contribute one gene copy as a paternal parent and two copies as a maternal parent resulting in a 3:2 transition advantage compared to outcrossers, which can only transmit a single copy as the maternal parent (Fisher 1941). This process is known as the ‘automatic selection’ hypothesis for the evolution of selfing. However, selfing can also evolve because individuals can reproduce under conditions where pollinators or mates are limiting, a benefit commonly known as ‘reproductive assurance’ (Darwin 1876; Jain 1976). Schoen et al. (1996) predicted that the population bottlenecks associated with the reproductive
assurance hypothesis should have genome-wide consequences, whereas only a few modifier loci would likely be affected if selfing was favored due to a transmission advantage. It remains an open question whether the two main hypotheses for the evolution of selfing can be distinguished by genomic data. I address this question in my thesis.

The effects of selfing on the genome

Figure 1.2. Theoretical expectations of the genome-wide reduction in selection efficacy associated with the transition for outcrossing to predominant selfing.

Theoretical models predict that selection efficacy should be lower in the genomes of selfing populations, compared to those that are outcrossing, as the effective population size ($N_e$) is reduced in the former (Charlesworth and Wright 2001, Figure 1.2), all else being equal. In the genomes of selfers, restricted recombination and genetic hitchhiking via selective sweeps of beneficial mutations and background selection against deleterious mutations can reduce $N_e$
(Charlesworth et al. 1993a; Charlesworth and Wright 2001; Wright et al. 2008). Reductions in $N_e$ can also occur due to demographic factors such as population bottlenecks that accompany the shift to selfing, as selfing individuals have the ability to initiate colonies. In genomes with reduced $N_e$, stochastic fixation of alleles or the loss of rare alleles becomes more prevalent. These stochastic effects may constrain the effects of selection. Moreover, weak linkage among sites acted upon by opposing selective forces may interfere with the efficiency of selection (McVean and Charlesworth 2000). The joint influence of genetic and demographic factors can result in a genome-wide relaxation of constraints in selfing populations but the change in the distribution of selective effects in the genome remains unclear.

The efficacy of selection in selfing populations may also be influenced by the dominance of deleterious mutations. A negative relation between the dominance of mutations and the strength of their deleterious effects has been suggested for some time (Simmons and Crow 1977; Crow and Simmons 1983). Although deleterious mutations are on average partially recessive ($h \sim 0.2$; Agrawal and Whitlock 2011), the more homozygous background of selfers offers fewer chances for deleterious mutations to be masked, potentially leading to a higher selective pressure on such mutations in selfing populations (Glémin 2007). Indeed, theoretical expectations predict that strongly deleterious, highly recessive mutations are more likely to be purged by selection after recurrent inbreeding (Hedrick 1994; Wang et al. 1999), even though both strongly and mildly deleterious mutations contribute to inbreeding depression (Charlesworth and Charlesworth 1999; Wang et al. 1999). The association between the countervailing forces of a genome-wide reduction in the efficacy of selection and stronger purifying selection against recessive, highly deleterious mutations remains poorly understood.
The transition from outcrossing to predominant selfing in flowering plants may also alter the regulatory landscape of the genome. This is an important feature from the point of view of mating-system evolution, as pioneering work by King and Wilson (1975) suggested that most of the phenotypic differences in recently diverged species can be attributed to regulatory differences. Even recently derived selfing species can exhibit conspicuous differences in floral morphology when compared to their outcrossing progenitors, with smaller, unshowy flowers, the loss of nectar and reduced pollen production (Lloyd 1965; Morgan and Barrett 1989; Ritland and Ritland 1989; Armbruster et al. 2002). It seems very likely that regulatory divergence is a key aspect of the evolution of selfing, but this topic has not received the attention it deserves and I examine this problem in my thesis using genomic approaches.

**Genetic architecture of heterostyly and its breakdown to selfing**

A seemingly simple inheritance of one (distyly) or two (tristyly) Mendelian loci governing the heterostylish syndrome masks a more complex genetic architecture. Bateson and Gregory (1905) established that the S-locus governs distyly in *P. sinensis* with a dominant allele (*Ss*) resulting in the S-morph and recessive alleles (*ss*) giving rise to the L-morph (Figure 1.1). Single-locus inheritance has been observed across ~10 heterostylish families but with dominance reversals in a few taxa (Lewis and Jones 1992). Investigations of several distyloous taxa (e.g. *Primula*, *Fagopyrum*, *Turnera*) indicate that the S-locus is composed of several linked loci that together make up a supergene complex and the advantageous gene combinations are maintained by balancing selection (Charlesworth and Charlesworth 1979a; Lewis and Jones 1992; Barrett and Shore 2008). The order of loci within the complex is thought to be *G*, *P* and *A* in *Primula*: with *G* controlling style length and incompatibility, *P* controlling pollen size and incompatibility, and
A determining anther height and these inferences were based on studies of rare recombinant phenotypes (Ernst 1928, 1955; Dowrick 1956). Unlike distyly, tristyly is governed by two diallelic loci with the S-locus epistatic to the M locus. The S morph is determined by a dominant S allele (genotypes: SsMM, SsMm, SsMm) and the M morph by a dominant M allele (genotypes: ssMM or ssMm) and the L-morph is of genotype ssmm (Barlow 1923; Fisher and Mather 1943; Weller 1976; Eckert and Barrett 1993; Figure 1.1). More complex inheritance is observed in some taxa including the reversal of epistasis between the S and M loci (Von Ubisch 1926; Fyfe 1956; Mulcahy 1964), linkage between these two loci (Fisher and Martin 1948; Fyfe 1950, 1956; Weller 1976) and the possible presence of a third locus (Fyfe 1956; Bennett et al. 1986). By analogy with distylous species, it has often been assumed that supergene complexes also control tristyly (Barrett 1993). Alternatively, either the S or M locus or both might be comprised of just one gene with pleiotropic effects, as suggested by Charlesworth (1979). More extensive genomic tools have now become available to uncover the genetic architecture of the heterostyly genes and one of the goals of this thesis to investigate this problem in a tristylos species.

Current efforts for studying the genetic basis of heterostyly have focused on uncovering the molecular genetic architecture of the Mendelian loci using high-throughput genomics. Candidate genes and S-linked region have been found in distylous species including T. subulata (Labonne and Shore 2011), Fagopyrum esculentum (Yasui et al. 2012) and P. vulgaris (Li et al. 2015) by genetic mapping and differential gene expression approaches, but no causal genes underlying the S locus have been rigorously characterized. The most recent effort by Nowak et al. (2015) identified 13 S-linked variants and 113 genes with morph-biased expression in distylous P. veris following genome assembly and differential gene expression analysis. There is currently no molecular genetic evidence for the existence of supergenes controlling tristyly and
an objective of my thesis is to investigate the genetic architecture of the loci governing tristyly through genetic mapping approaches.

Interestingly, the breakdown of heterostyly can occur in multiple ways and this was discovered by analyzing the genetic basis of selfing variants. In distylous species, recombination within the $S$-supergene complex is associated with the origin of homostyles (e.g. *Primula* Dowrick 1956; *Tunera* - Shore and Barrett 1985; reviewed in Barrett and Shore 2008). Less is known about how semi-homostyles have originated but there is no evidence to indicate that it is due to recombination within the $S$ or $M$ loci (Charlesworth 1979). The origin of semi-homostyloous variants in *E. paniculata* is governed by recessive modifiers non-allelic to the $S$ or $M$ loci (Fenster and Barrett 1994, Vallejo-Marín and Barrett 2009). Preliminary evidence raises the possibility that the multiple transitions to selfing within this species are governed by separate modifier loci (Barrett et al. 2009) and I explore this possibility in my thesis by mapping the genetic regions associated with independent transitions to semi-homostyly.

**The study system *Eichhornia paniculata***

My thesis investigates the molecular genetic basis for tristyly and the genetic, genomic and phenotypic consequences of the breakdown of the polymorphism in the annual aquatic plant *E. paniculata* (Figure 1.3A). This species is diploid ($n=8$) and belongs to the Pontederiaceae (Monocotyledoneae: Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009), a small family of freshwater aquatics primarily native to the Neotropics, but with a limited distribution in parts of the Old World (Africa, Asia and Australia). All known *E. paniculata* populations that are tristylos occur in the arid caatinga (dry thorn scrub) region of N.E. Brazil (Figure 1.3B). The tristylos polymorphism in *E. paniculata* is associated with a cryptic
trimorphic incompatibility system, in which inter-morph cross-pollination is favored over self and intra-morph pollination as a result of differential pollen-tube growth (Cruzan and Barrett 1993, 2015). This preferential fertilization by legitimate pollen results in phenotypic disassortative mating and negative frequency dependent selection (Barrett et al. 1987) and drives populations to an isoplethic (1:1:1 morph frequencies) equilibrium. The stochastic loss of the $S$ allele from populations through founder events and genetic drift initiates the evolutionary breakdown of tristyly, which occurs through the spread of semi-homostyles in dimorphic populations (Barrett et al. 1989, 2009; Figure 1.3C). During this transition, recessive modifier genes result in stamen elongation to same height as the stigma and this loss of herkogamy results in autonomous self-pollination, especially in the L- and M-morphs (Fenster and Barrett 1993; Barrett et al. 2009; Figure 1.4). Mid-styled semi-homostyles (hereafter M´) are distributed in N.E. Brazil, Jamaica and Cuba. Although much less common, long-styled semi-homostyles (hereafter L´) are also found but are restricted to a few isolated populations in Nicaragua and Mexico. The origin of semi-homostylous variants in *E. paniculata* is associated with long-distance dispersal from N.E. Brazil to the Caribbean and Central America (Barrett et al. 1989; Husband and Barrett 1991; Barrett et al. 2009; Vallejo- Marín and Barrett 2009). Using coalescence modeling of DNA sequences, Ness et al. (2010) estimated that the colonization of the Caribbean from N.E. Brazil occurred ~120,000 years ago. Allozyme (Husband and Barrett 1993) and molecular evidence (Barrett et al. 2009) indicate that transitions to selfing in *E. paniculata* have occurred independently, thus allowing comparison of the genetic basis and morphological and genomic consequences of these shifts in mating system.
Figure 1.3. The study system *Eichhornia paniculata*. (A) A Cuban population near Camelote, Camagüey composed of sem-homostylosous mid-styled plants, (B) a tristylosous outcrossing flower from Brazil and (C) a semi-homostylosous mid-styled selfing flower from Jamaica.

Figure 1.4. Evolutionary pathways from cross-fertilization to self-fertilization in tristylos *Eichhornia paniculata* via the origins of semi-homostyly. The primary pathway from
trimorphism to dimorphism culminates in the fixation of selfing semi-homostyloous forms of the M-morph. A less common pathway leads to populations monomorphic for semi-homostyloous 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. After Barrett et al. (2009).

**Research objectives**

Here, I briefly summarize the main goals of my PhD thesis. In chapter 2, I provide a review of the demographic and population genomic consequences of the evolution of self-fertilization in flowering plants. In this chapter I also investigated if the two key hypotheses on why selfing is favored, the transmission advantage and reproductive assurance hypotheses, can be disentangled using molecular genetic data. In addition I also studied how long it may take for the influence of reduced $N_e$ accompanying the evolution of selfing to be detected across the genome. In chapter 3, I investigated the genomic consequences of the evolution of selfing, specifically the reduction in selection efficacy due to reduction of $N_e$ in selfing populations and evidence suggesting the purging of highly recessive deleterious mutations using both simulated and empirical population genetic data from *E. paniculata*. In chapter 4, examined if the transition to selfing in *E. paniculata* has been accompanied by differences in the levels of gene regulation as expected for recently derived lineages with striking phenotypic differentiation. In chapter 5, I examined the genetic basis of the tristyloous polymorphism and specifically investigated whether the two diallelic loci ($S$ and $M$) governing tristyly are physically linked. Finally, in Chapter 6, I investigated the genetic architecture of the $M$ locus by QTL mapping studies and examined whether independent transitions to selfing were governed by different sets of mating system
modifier genes. Below, I summarize the chapters in my thesis, each of which is written as a self-contained research paper for publication and therefore there is some repetition of material among chapters.

**Chapter 2 - The demography and population genomics of evolutionary transitions to self-fertilization in plants.**

In this chapter, I investigated how demographic and genetic processes influence the loss of genetic diversity accompanying the transition from outcrossing to selfing. Whereas the transmission advantage is expected to only reduce diversity in one or a few modifier loci responsible for the evolution of selfing, bottlenecks associated with reproductive assurance might lead to a genome-wide reduction in diversity (Schoen et al. 1996). Multilocus simulations and comparative molecular data from closely related outcrossing and selfing species indicated that the linked selection quickly reduces genome-wide diversity following the shift to predominant selfing. Therefore, I conclude that it may be difficult to infer the role of historical founder events associated with reproductive assurance using molecular data. I discuss the necessary ecological and genomic parameters for quantifying role of selection and demography during transitions to selfing and the loss of genetic diversity.

Chapter 3 - Reduced efficacy of selection and purging of deleterious mutations accompany the evolution of selfing.

The evolution of selfing is expected to lead to a genome-wide reduction in selection efficacy due to the reduced $N_e$ of highly selfing populations. At the same time, strongly recessive deleterious mutations should be under greater purifying selection in homozygous selfing genomes. Here, I examined the distribution of fitness effects in *E. paniculata* coupled with computer simulations to investigate how the magnitude of effective selection coefficients acting on mutations changed during the transition from outcrossing to selfing. I found that the transition to selfing was accompanied by a larger fraction of effectively neutral sites, a result consistent with the effects of reduced $N_e$ in selfers. Also, a larger fraction of sites were under strong purifying selection in selfers, potentially associated with the exposure of strongly deleterious recessive mutations. I concluded that the genomes sampled from selfing *E. paniculata* populations showed evidence of reduced $N_e$, when compared to the genomes of their outcrossing progenitors, and also showed signatures of purifying selection against strongly deleterious mutations.


Chapter 4 - Recent mating system evolution in *Eichhornia paniculata* is accompanied by cis-regulatory divergence.

The extent of the association between regulatory variation and divergence between outcrossing and selfing lineages is currently unclear. In this chapter, I examined how gene expression
changes during the transition to selfing in *E. paniculata*. I crossed plants from outcrossing and selfing populations and tested for the presence of allele-specific expression (ASE) in floral buds and leaf transcriptomes of F₁ progeny, an approach that would estimate the level of *cis*-regulatory variation in the genome. <10% of the genes in the genome exhibited ASE in floral buds and leaf tissues in the F₁ progeny, and these genes preferentially expressed alleles from outcrossing parents. My findings suggest that the transition from outcrossing to high selfing rates may alter the *cis*-regulatory genomic landscape, and that the down-regulation of alleles following the transition to selfing is associated with accumulation of deleterious mutations in the ASE genes in *E. paniculata*.

**Chapter 5 - Genetic linkage between the S and M loci governing tristyly in Eichhornia paniculata (Pontederiaceae).**

The mode of inheritance and nature of linkage in *E. paniculata* between the diallelic *S* and *M* loci governing the tristylos polymorphism is not well established. Using controlled crosses and progeny tests of genotypes heterozygous at the *S* and *M* loci I found that the loci were closely linked and separated by a map distance of 2.7 cM. A literature survey indicated that in two out of the three tristylos families for which data on the inheritance of tristyly is available the *S* and *M* loci were linked in some species but were unlinked in others. I discuss the influence of linkage between the *S* and *M* loci on style morph ratios in populations and on the evolution of tristyly.
Chapter 6 - The genetic architecture of style morph variation in tristylous Eichhornia paniculata and modifiers associated with evolutionary transitions to self-fertilization.

The genetic architecture governing the tristylos syndrome remains unknown. In this chapter, I investigated the genetic architecture of the $M$ locus in tristylos $E. \text{paniculata}$ and the genetic basis of the breakdown to selfing in the L- and M-morphs. I conducted a crossing program generating genotypes segregating for alternate alleles at the $M$ locus ($ssMm$ or $ssmm$). I genotyped and phenotyped these plants and performed composite interval mapping for a number of floral traits including style length and stamen position. In doing so, I identified the interval regions associated with style-morph polymorphism and found that the stamen modifications characterizing the two semi-homostylous selfing variants were governed by separate genetic regions. I also bulked and sequenced the transcriptomes of these plants to identify 334 genes containing SNPs potentially linked to the $M$ locus. In this chapter, I also localized the genetic region containing the $M$ locus and found evidence indicating that separate transitions to selfing in $E. \text{paniculata}$ have originated by different genetic loci involving contrasting sets of mating-system modifier genes unlinked to the $M$ locus.
Chapter 2

The demography and population genomics of evolutionary transitions to self-fertilization in plants

This chapter was published as Barrett, S.C.H., Arunkumar R., & Wright, S.I. (2014).

Summary

The evolution of self-fertilization from outcrossing has occurred on numerous occasions in flowering plants. This shift in mating system profoundly influences the morphology, ecology, genetics, and evolution of selfing lineages. As a result, there has been sustained interest in understanding the mechanisms driving the evolution of selfing and its environmental context. Here, using multilocus simulations and comparative molecular data from related selfers and outcrossers, we investigate the importance of demographic and genetic processes affecting the selection of selfing, and loss of genetic diversity that accompanies this transition. We conclude that there is little evidence for strong bottlenecks associated with initial transitions to selfing, and our simulation results cast doubt on whether it is possible to infer the role of bottlenecks associated with reproductive assurance in the evolution of selfing. They indicate that the effects of genetic hitchhiking on diversity occur rapidly following the shift to high selfing, consistent with growing evidence for a reduced efficacy of purifying selection, even following recent transitions to selfing. Future comparative studies that integrate explicit ecological and genomic details are necessary for quantifying the independent and joint effects of selection and demography on transitions to selfing and the loss of genetic diversity.
Introduction

Hermaphroditism is the primary sexual condition of most seed plants and some animals, particularly invertebrates. The reproductive opportunities provided by cosexuality result in a range of mating systems from obligate outcrossing, through mixtures of outcrossing and self-fertilization (mixed mating), to predominant selfing (autogamy). In angiosperms, this variation is associated with diverse features of the ecology, life history and demography of populations. For example, long-lived woody species of stable habitats are usually highly outcrossing and maintain large effective population sizes, whereas those that are annual colonizers commonly practice high levels of selfing, and have smaller populations (Barrett et al. 1996). Such associations implicate environmental and demographic conditions as playing important roles in the evolution and maintenance of plant mating systems, which are key components of life history and the recombination system. Because a variety of demographic and genetic processes can influence levels of recombination and effective population size, identifying their relative importance will provide a more complete understanding of the evolution of mating systems.

The evolution of predominant self-fertilization from outcrossing represents the most important reproductive transition in flowering plants. Although only ~10-15% of angiosperm species are highly autogamous, there is extensive evidence from numerous herbaceous taxa of multiple independent origins of autogamy from outcrossing (Barrett et al. 1996; Stebbins 1974; Jain 1976). The frequency with which selfing originates provides opportunities for identifying convergent patterns, and the mechanisms and consequences of this transition. Accordingly, efforts to understand how and why selfing evolves has had a long and venerable history, beginning with the early naturalists, including Charles Darwin, George Henslow, and Hermann Müller (reviewed in Lloyd 1980), and continuing today using diverse comparative, experimental
and genomic approaches (Busch and Delph 2012; Igic and Busch 2013; Wright et al. 2013). In addition, a rich body of theory (reviewed in Goodwillie et al. 2005) has developed on the selective mechanisms governing mating system variation and the genetic and evolutionary consequences of transitions to selfing.

There are four principle reasons why the evolution of selfing from outcrossing has attracted so much attention from evolutionary biologists: 1) transitions to selfing are associated with major changes to floral morphology, sex allocation and life history, thus contributing to taxonomic and ecological diversity and serving as valuable systems for the study of plant adaptation (Sicard and Lenhard 2011); 2) because selfing enables individuals to start colonies following dispersal, or to reproduce under conditions of low density, the transition to selfing can have significant demographic and biogeographical consequences (Lloyd 1980); 3) Predominant selfing has a profound effect on the genome and autogamous species can provide important model systems for investigating the evolutionary consequences of restricted recombination. High rates of selfing reduce levels of heterozygosity, restrict recombination and increase genetic hitchhiking of both deleterious and advantageous mutations, thus leading to a genome-wide loss of genetic diversity. These changes reduce the efficacy of natural selection and increase the probability of fixation of deleterious mutations owing to genetic drift (Charlesworth and Wright 2001); 4) predominant selfing has important influences on macroevolutionary processes affecting speciation and extinction rates and patterns of lineage diversification (Igic and Busch 2013; Wright et al. 2013). These diverse facets of the biology of selfing present both challenges and opportunities for evolutionary biologists interested in adaptation and speciation. In particular, opportunities arise to better understand the evolutionary causes and consequences of transitions in life history and genetic system, although because many of the genetic and demographic factors
involved change concurrently, disentangling their individual and joint effects presents an important challenge.

The late plant evolutionary biologist Leslie Gottlieb, to whom this volume is dedicated, made valuable contributions to our understanding of the evolution of selfing. Gottlieb’s classic investigations of sympatric speciation in annual *Stephanomeria* provided evidence that the origin of *S. malheurensis* from *S. exigua* ssp. *coronaria* was associated with a shift from outcrossing to selfing. This transition resulted from the breakdown of self-incompatibility and the development of hybrid sterility, owing to chromosomal structural differences between the progenitor and derivative (Gottlieb 1973; Brauner and Gottlieb 1987). Today, it is widely recognized that the loss of self-incompatibility is commonly associated with shifts from outcrossing to selfing, and in many cases this also appears to promote speciation (Goldberg and Igic 2012). In *Clarkia*, Gottlieb also investigated the evolutionary origins of selfing from outcrossing. Using allozyme variation he demonstrated two independent origins in *Clarkia concinna* (Allen et al. 1991), and a single origin in *C. xantiana* (Gottlieb 1984). These studies were among the first to use molecular data to trace transitions from outcrossing to selfing, and to show that selfing populations contained markedly reduced amounts of genetic variation compared to outcrossing populations, a result that Gottlieb extended by surveying additional taxa (Gottlieb 1977, and now confirmed in numerous species by Hamrick and Godt 1996). His pioneering efforts helped to stimulate subsequent work on the origins of selfing and its influence on patterns of genetic diversity.

Here, we consider recent developments and ongoing challenges in understanding the evolution of selfing and the demographic context in which it occurs. The advent of high throughput DNA sequencing provides opportunities for investigators to address a variety of questions concerned with the selective mechanisms causing transitions, as well as their genomic
and evolutionary consequences. We focus in particular on the problem of distinguishing the relative importance of demographic and genetic processes affecting the selection of selfing, and the reduction in genetic diversity that accompanies this transition. To provide a general background we begin by summarizing the two main hypothesized selective mechanisms for selfing, and the principle reproductive modifications that cause increased selfing rates and what is known about their genetic architecture. We then consider the demographic circumstances in which selfing evolves and how this may influence mechanisms of selection. Using simulation studies and comparative data from related selfers and outcrossers we then address three main questions: 1) Can molecular evidence be used to distinguish the mechanisms of selection for selfing, as previous authors have proposed? 2) What is the relative importance of demographic factors and genetic hitchhiking in causing reductions in genome-wide diversity in recently derived selfing populations, and over what time period do these processes become evident? 3) How rapidly does the transition to selfing lead to reductions in the efficacy of natural selection across the genome? Our principal conclusions are that the effects of genetic hitchhiking on diversity and the efficacy of selection can act quite quickly following the shift to selfing, and that there is little evidence to be gleaned from molecular data for strong bottlenecks during the earliest stages of the shift to selfing.

Selection of self-fertilization

Diverse reproductive, demographic and genetic factors influence the transition from outcrossing to selfing, of which the magnitude of inbreeding depression is generally considered most influential for the maintenance of outcrossing (Lande and Schemske 1985; Lloyd 1992). The assured reproduction that selfing provides when pollen vectors and/or compatible mates limit
cross-pollination, and the transmission advantage of genes which cause selfing, are the two most
general selective mechanisms that have been proposed to explain the shift to selfing. The
‘reproductive assurance hypothesis’ traces back to Darwin (1976) and is the most common
explanation for why selfing evolves. Less often examined is the ‘automatic selection hypothesis’,
which is based on Fisher’s idea that a gene for selfing has a 3:2 transmission advantage when it
arises in an outcrossing population (Fisher 1941). Determining the relative importance of these
two hypotheses is not straightforward, particularly where both processes operate during the
transition to selfing. Distinguishing the two hypotheses requires determining empirically the
modes of self-fertilization in a population (Lloyd 1992), and the extent to which selection of
mating-system modifiers occurs through pollen and/or seed, which necessitates information on
pollen and seed discounting (Busch and Delph 2012). Neither of these parameters has been
simultaneously measured in any plant population, and currently it is unclear how important they
are in counteracting the selective advantage of selfing, particularly in comparison with
inbreeding depression.

An alternative approach for investigating the relative importance of the two main
hypotheses for the evolution of selfing was first suggested in Schoen et al. (1996). They
suggested that the demographic and genetic processes associated with reproductive assurance
and automatic selection might be expected to result in different molecular signatures that could
potentially be distinguished by examining patterns of neutral diversity among selfing
populations. Specifically, they suggested that the evolution of selfing through reproductive
assurance should be accompanied by a genome-wide reduction in genetic diversity due to
population bottlenecks, potentially resulting in founder events. In contrast, the evolution of
selfing through automatic selection involves the spread of mating-system modifiers in
outcrossing populations, which should cause reduced diversity at the modifier itself, but not at unlinked neutral loci. They also suggested that recurrent bottlenecks due to reproductive assurance should lead to elevated among-population variance in diversity, as populations will be at different degrees of recovery from bottlenecks.

Recently, molecular population genetic data from five taxa and a total of six relatively recent transitions to selfing (<10-150ka) was used to evaluate mechanisms causing shifts to selfing (Busch and Delph 2012). In half of these cases, reductions in diversity were severe, consistent with predictions from reproductive assurance, while the remaining cases showed no evidence for a greater reduction in diversity than expected by selfing alone. However, there are several thorny conceptual and empirical issues that need to be addressed in using molecular population genetic data to distinguish the selective mechanisms driving the evolution of selfing.

The above predictions rest on various assumptions, several of which may be violated in selfing populations. First, one key assumption is that selfing evolves through mutations of small effect (Schoen et al. 1996). If major mutations causing large shifts in selfing rate arise and spread rapidly, genome-wide reductions in diversity may occur in the absence of selection for reproductive assurance. Although small-effect modifiers are reported in some systems, the breakdown of self-incompatibility and floral modifications promoting selfing can involve genes of large effect (see below). Second, subsequent gene flow from outcrossing populations following the shift to selfing could erode an initial signal of a bottleneck during the transition to selfing (Busch and Delph 2012). Third, additional factors during the origin of selfing other than bottlenecks are known to reduce diversity in selfing relative to outcrossing populations (Charlesworth and Wright 2001), as we explore in detail later in this article.
Modification to reproductive traits that promote selfing

The shift to high selfing rates is generally associated with a series of modifications to reproductive traits often beginning with the facility for autonomous self-pollination, and culminating in the evolution of the selfing syndrome (Sicard and Lenhard 2011). A challenging problem, also shared with work in speciation genetics, is to identify which traits directly initiate the process of mating system change, and which are consequences of increased selfing rates; for example, subsequent changes in sex allocation to male function (Charnov 1982). In addition, determining the genetic basis of modifiers of the selfing rate can provide insight into the mechanisms responsible for transitions.

In self-incompatible species, the initial step usually involves the breakdown of the physiological self-recognition system and the evolution of self-compatibility. This change commonly involves major mutations producing fully self-compatible variants, although it can be gradual involving different degrees of self-fertility (Good-Ávila et al. 2008). In self-compatible species, three morphological modifications commonly cause increasing selfing rates: a reduction in stigma-anther separation (herkogamy), the loss of dichogamy (timing of anther dehiscence and stigma receptivity), and a decrease in flower size. The extent to which these traits function alone or in concert to initiate increased selfing is poorly understood in most lineages. This is especially the case in highly autogamous species with well-developed selfing syndromes in which numerous substitutions have probably accumulated since their early divergence from outcrossing ancestors.

In some taxa it has been possible to identify the early changes provoking mating-system evolution. For example, in Collinsia spp. a weakening of dichogamy appears to be the critical first step causing increased selfing (Kalisz et al. 2012), whereas in Eichhornia paniculata the
loss of herkogamy, through stamen elongation governed by recessive modifiers of large effect is of primary importance (Vallejo-Marín and Barrett 2009). In N. American populations of *Arabidopsis lyrata* the breakdown of self-incompatibility system is associated with increased selfing rates (Foxe et al. 2010), but this physiological change does not appear to be associated with obvious morphological modifications to flowers causing autonomous self-pollination. In *Capsella rubella* flower size reduction appears to have followed the loss of self-incompatibility (Sicard et al. 2011). Interestingly, the breakdown of self-incompatibility in this system alone confers a moderate facility for autonomous selfing, which is ~50% that of the derived selfing species, indicating that the mutation causing self-compatibility represents a major effect mating-system modifier. In *Leavenworthia alabamica*, two independent losses of self-incompatibility have occurred, one of which is accompanied by floral modifications and high selfing rates, whereas the other shows no obvious changes to floral traits and populations exhibit mixed mating (Busch et al. 2011). Although the loss of self-incompatibility provides opportunities for selfing, the extent to which this occurs will depend on several factors. In taxa with generalized pollinators and relatively small flowers (e.g. many Brassicaceae) the loss of incompatibility alone may be sufficient to permit substantial selfing, as in *C. rubella*, through various modes of selfing (e.g. prior, delayed and pollinator mediated, see Lloyd 1992 and below). In contrast, in taxa with larger more specialized flowers the availability of standing genetic variation for traits promoting autonomous self-pollination is probably necessary for high selfing rates to evolve.

Information on the genetic architecture of mating system modifiers can provide insight into the selective mechanisms driving the transition from outcrossing to selfing, its demographic context and the tempo with which it occurs. Key questions include: what is the relative importance of mutations of large versus small effect governing selfing rate modifiers? How
important are new mutations versus standing genetic variation as a source of modifier alleles? Answers to these questions are important because theoretical work indicates that a mutation of large effect causing predominant selfing can spread to fixation despite high inbreeding depression (Lande and Schemske 1985). If this occurs it has the potential to result in a genome-wide selective sweep since the highly selfing variant may be reproductively isolated from outcrossing individuals immediately, adding to the diverse forces reducing diversity in recently-derived selfing populations. Moreover, if bottlenecks associated with reproductive assurance play an important role in the shift to selfing the importance of standing genetic variation, as a source of selfing modifiers may be quite limited (Glémin and Ronfort 2013a). Unfortunately, few detailed studies have been conducted on the genetics of selfing traits so at this stage there are no clear answers to these questions.

QTL-mapping of crosses between self-compatible, outcrossing *Mimulus guttatus* and selfing *M. nasutus* (Fishman et al. 2002) provided no evidence of large effect QTL’s for flower width, stigma-anther separation and several other traits associated with selfing. Although the two *Mimulus* species have undergone considerable floral divergence, the absence of major QTL’s governing floral traits is consistent with the hypothesis that selfing has evolved gradually by the sequential fixation of modifiers of small effect. In self-compatible groups such as *Mimulus*, fluctuating selection on outcrossing rates, owing to year-to-year variation in population size and pollinator service, may lead to relatively high levels of standing genetic variation for mating system traits and thus more often polygenic control. Indeed, a novel experimental evolution study, conducted with the presence and absence of pollinators in an outcrossing population of *M. guttatus*, revealed remarkably rapid evolution of floral traits promoting autonomous self-pollination during five generations in the ‘pollinator-free’ treatment (Bodbyl Roels and Kelly
2011), implying high levels of standing genetic variation for traits influencing the mating system. Alternatively, new mutations could play a more prominent role during transitions to selfing in self-incompatible species because guaranteed outcrossing may be associated with less temporal variation in selection and thus lower levels of standing genetic variation on mating system traits.

Consistent with this prediction, QTL-mapping of floral traits from crosses between self-incompatible, outcrossing *Capsella grandiflora* and selfing *C. rubella* revealed a different genetic architecture than was evident in *Mimulus*. Flower size traits were governed by changes at a relative small number of loci, some with large effect, in addition to the single dominant allele that causes self-compatibility (Slotte et al. 2002). Thus, it appears that genes of major effect for both self-compatibility and floral traits may have been important in the spread of selfing in *Capsella rubella*, and the spread of these major-effect alleles could have been an important contributor to genome-wide reductions in diversity in the early stages of mating-system evolution.

**Ecological and demographic context**

The observation that autogamous plant species are not randomly distributed with respect to biogeography, colonizing ability, habitat, demography, and life history has often been made (reviewed in 4). Although there are numerous exceptions, selfing is commonly associated with island colonization and geographical marginality; weediness; ephemeral, open disturbed habitats; low density and short-lived life forms (e.g. annuality). Baker’s Law and various aspects of the reproductive assurance hypothesis are most often used to explain these associations. Individuals that are self-compatible, with a capacity for autonomous self-pollination, are likely to be favored over outcrossing individuals, when pollinators and/or mates are in short supply, as might often
occur at low density or following dispersal. Although such explanations seem perfectly plausible in explaining these correlations, it has proven more difficult to obtain convincing experimental evidence that these ecological associations are directly involved in driving transitions from outcrossing to selfing, rather than simply being a consequence of the possession of an autogamous mating system. The direction of causality cannot be determined from ecological and demographic correlates of selfing alone.

Studies of the ecology of wide ranging species with variable mating systems can provide insights into the potential role of demographic factors in causing shifts to selfing. In *E. paniculata* selfing has evolved on multiple occasions associated with the evolutionary breakdown of tristyly (Barrett et al. 2009). Several lines of evidence indicate that repeated bottlenecks and genetic drift in small populations have played an important role in destabilizing tristyly (Barrett et al. 1989). The main concentration of populations in N.E Brazil is largely outcrossing and primarily pollinated by specialist long-tongued bees. In contrast, in Cuba and Jamaica, where specialist bees do not visit populations, highly selfing populations predominate, consistent with Baker’s Law and reproductive assurance. Selfing and mixed mating populations occur in Brazil, but have not evolved the selfing syndrome evident in island populations, probably owing to their more recent origin and occasional gene flow from tristylos populations. In *E. paniculata*, selfing rates are negatively correlated with population size, and both selfing rates and population size are strongly correlated with diversity (Ness et al. 2010). Both reproductive assurance and automatic selection probably play a role in the evolution of selfing in Brazil, with their relative importance depending on the extent to which small populations of *E. paniculata* receive pollinator service, and selfing variants gain fitness as male parents.
Insufficient outcross pollen delivery owing to low density has frequently been invoked to account for the evolution of selfing, but relatively few studies have investigated this problem directly, despite extensive experimental data on pollen limitation in numerous species (Knight et al. 2005). In *Leavenworthia alabamica* two self-compatible races have evolved independently from self-incompatible, outcrossing populations in different parts of the periphery of the species’ range, and are characterized by significantly smaller populations, often of lower density (Busch 2005). Molecular data have been interpreted as suggesting that a bottleneck could have facilitated the evolution of selfing by reproductive assurance in one race, but not in the other (Busch et al. 2011), although a study of pollen limitation found no evidence that it has played a role in the evolution of self-compatibility (Moeller et al. 2005). In contrast, field experiments on pollen limitation in *Clarkia xantiana* provided evidence for the evolution of selfing by reproductive assurance in populations of small size or low density owing to few mates and effective pollinators (Pettengill and Moeller 2011). Molecular data on the evolutionary history of selfing consistent with a strong genetic bottleneck has also been interpreted to support the reproductive assurance hypothesis (Charlesworth 2013). The field study of *Clarkia* is particularly significant because it demonstrated context-depend natural selection on selfing traits; in small populations, isolated from congeners, selection was strongest for reduced herkogamy and protandry.

**The causes for loss of genetic diversity in selfing populations**

As discussed earlier, researchers have used data on molecular variation from recently derived selfing populations and their outcrossing progenitors to test for evidence of bottlenecks and founder events and assess the role of reproductive assurance in the evolution to selfing (Busch
and Delph 2012). However, as illustrated in Figure 2.1, multiple factors can contribute to reduced diversity in selfing populations, and distinguishing their relative contributions to overall reductions in diversity following the transition to selfing represents a major challenge. Under neutrality, selfing lowers the effective population size up to two-fold due to the reduced effective number of alleles in a population. Also, because of reduced effective rates of recombination, neutral diversity will be lowered further due to the action of genetic hitchhiking, including background purifying selection and selective sweeps of advantageous mutations. If strong enough and with sufficient time to reach the new equilibrium, these effects will erode historical signals associated with the mechanisms of selection for selfing. Thus, in making historical inferences on the selective mechanisms for selfing it is first necessary to assess the effect that different strengths of linked selection have, relative to other factors, in reducing variation in selfing populations.

Figure 2.1. Expected reduction in genetic diversity in selfing compared to outcrossing populations with selfing (solid line), selfing and background selection (long dashed line), and with selfing, background selection and population bottlenecks (short dashed line). The expected
reduction in diversity due to selfing alone was $1/(1+F)$ where $F$ is the inbreeding coefficient. The expected reduction in diversity due to background selection was modeled after (Kimura 1983), with selection and mutation parameters taken from estimates in the Brassicaceae, as described in the supplementary methods (Appendix I). The effect of a bottleneck was assumed to be a reduction in diversity to 25% of the original value, regardless of selfing rate.

Previously, the lack of estimates of deleterious mutation rates and their selection coefficients has made it difficult to assess the impact that background selection might have on plant genomes (Charlesworth 2013). This has been particularly problematic since both background selection and population bottlenecks will cause genome-wide reductions in diversity (Charlesworth and Wright 2001). Although our understanding of mutation rates and their selection coefficients is still limited, it is now possible to more explicitly assess the role of background selection in plant populations to predict the extent to which it can explain the loss of variation in selfing populations. In particular, genome-wide estimates are now available in the Brassicaceae for the per-base pair mutation rate due to nucleotide substitutions (Ossowski et al. 2011), the number of base pairs in the genome under selection (Haudry et al. 2013), and the strength of selection on functional mutations (Slotte et al. 2010). Whereas both mutation rates and estimates of the proportion of the genome under selection are likely under-estimates, these values allow us to obtain a lower-bound prediction of the reduction in within-population diversity expected due to background selection in selfing populations.

Using the approximation of Glemin and Ronfort (2013a,b), and available estimates of mutation and recombination parameters, our simulations indicate that background selection is expected to reduce the equilibrium effective population size to about 10% that of an obligate
outcrosser (Figure 2.1), well below the 50% reduction expected in the absence of background selection. The reductions in diversity as a result of background selection are primarily restricted to populations with rates of selfing >90-95%. The effects of background selection illustrated in Figure 2.1 apply to long-term equilibrium, and thus a critical question concerns how rapidly these effects are likely to occur following the transition to selfing. If reductions in diversity due to linked selection occur quickly following the transition, it will be difficult to use molecular data to infer the role of reproductive assurance because most selfing populations will have rapidly reached the new equilibrium of low diversity.

Figure 2.2. The reduction in synonymous diversity ($\theta_w$) as a function of time since the transition to selfing, in units of $N$ generations, where $N$ is the effective population size of the ancestral outcrossing populations. We performed forward population genetic simulations using 300 individuals with 100Mb genomes comprised of alternating noncoding (NC) and coding (C)
regions of various sizes. All NC sites were neutral, whereas for C sites 25% were neutral and 75% were deleterious under variable selection pressures as measured by $s$, the mean selection coefficient acting on deleterious mutations. Details of simulations are in the supplementary methods (Appendix I). Shown are plots for: (a) $s=0$, C=200bp, NC=800bp, (b) $s=-0.1$, C=200bp, NC=800bp, (c) $s=-0.05$, C=200bp, NC=800bp, (d) $s=-0.005$, C=200bp, NC=800bp, (e) $s=-0.05$, C=350bp, NC=650bp, and (f) $s=-0.05$, C=500bp, NC=500bp. The ratio of the means across five runs for each interval (solid line) and the expected two-fold reduction due to selfing alone (dashed line) are plotted. The plots are ordered by parameters that led to increasingly stronger background selection. Without background selection, the loss of diversity occurred within $\sim 2N$ (a). As $s$ was decreased, the loss of diversity occurred within 1-2$N$ generations (b-d). Reduction in the size of NC, even under moderate $s$, led to the strongest effect of background selection where the loss of diversity occurred within 0.5$N$ (e, f).

To determine how rapidly the loss in diversity occurs following a shift to selfing, we simulated mating-system evolution in the presence and absence of deleterious mutations. We kept demographic variables constant and used mutation and recombination parameters resembling the *Arabidopsis* genome. We also used a more realistic genome structure in the simulations than in analytical approximations, incorporating spatial patterning of neutral and selected sites, rather than equal spacing of selected sites across the genome. In our simulations, we varied the spatial patterning and strength of selection against deleterious mutations. This effectively varies the strength of background selection, because greater spacing between neutral and selected sites and stronger purifying selection should reduce the intensity of linked selection. Figure 2.2 illustrates the ratio of diversity in the derived selfing population relative to the
outcrossing population as a function of time since the origin of selfing. When we did not include deleterious mutations, we observed a two-fold reduction of diversity in selfing populations as expected, and the loss of diversity occurred over a timescale of approximately $2N$ generations, where $N$ is the effective size of the ancestral outcrossing population (Figure 2.2a). This result fits with coalescent predictions, since it takes approximately $4N$ generations to reach equilibrium (Kimura 1983), and the 50% reduction in effective population size causes the decline in diversity to be more rapid. With purifying selection acting on deleterious mutations, we observed a greater than two-fold reduction in diversity (Figure 2.2b-f), and the time to reach equilibrium was more rapid with increasing strengths of background selection due to the lowered effective population sizes. Although not considered here, the effect of selective sweeps of beneficial mutations could also further reduce diversity in recently derived selfing populations. Given emerging evidence for high rates of positive selection in some plant populations (Hough et al. 2013), and the possibility of selective sweeps associated with new fitness optima due to the mating system shift, positive selection may also be a significant factor contributing to reductions in genetic diversity in selfing populations. These results imply that reductions in diversity due to genetic hitchhiking in highly selfing populations can occur relatively rapidly, particularly under conditions of strong background selection where the effective population size is small.

If the effects of linked selection are a major cause of diversity loss in recently derived selfing populations we can make several predictions. First, reductions in diversity should be primarily restricted to populations that initially transition to high selfing rates ($s>90\%$), and be less evident where the early stages of the spread of selfing variants are associated with mixed mating at the population level (Figure 2.1). Second, reductions in diversity should be more prominent in relatively older selfing populations, which have had more time for the effects of
linked selection to operate, and are more likely to have reached the new equilibrium diversity due to linked selection. Alternatively, if selection for reproductive assurance causes the initial spread of selfing modifiers, we might expect populations with smaller increases in selfing rate to show reductions in diversity beyond neutral expectation, due to the effects of population bottlenecks, as illustrated in the dotted line in Figure 2.1 (also see Schoen et al. 1996).

Furthermore, bottleneck effects associated with the transition to selfing should be apparent even in the most recent transitions to higher selfing rates.

To assess these predictions, we conducted a literature survey to obtain estimates of synonymous nucleotide diversity for related outcrossing and selfing species pairs. We only considered studies that reported diversity estimates for species pairs in which selfing species or populations were recently derived. We excluded more distantly related pairs, as there was a greater chance that forces unrelated to mating system shifts might have influenced their patterns of diversity. We estimated the ratio of nucleotide diversity in selfing populations relative to their outcrossing progenitor, and excluded studies that estimated diversity using less than five genes, since small numbers of loci will not give reliable estimates of a reduction in diversity. Consistent with the expectations of linked selection, we found that only shifts to high rates of selfing led to substantial reductions in diversity relative to outcrossing progenitors (Figure 2.3a). Mixed mating populations maintained moderate levels of diversity, indicating that significant population bottlenecks have not occurred and/or that gene flow from outcrossing populations after bottlenecks has masked the initial demographic effects on the genome. Our survey results also indicate that, as might be expected, the reduction of diversity in selfing populations in the early stages of the transition was generally less compared to older lineages (Figure 2.3b). Although the age estimates of selfing lineages are highly dependent on assumptions about
mutation rates and generation times, the populations with the strongest reductions in genetic diversity do appear to be among the most highly diverged, and show the most extreme changes in floral morphology (Busch and Delph 2012). These results are consistent with predictions from linked selection, and/or if older selfing lineages with high selfing rates have experienced recurrent bottlenecks and reduced gene flow. However, they are inconsistent with the prediction that the earliest stages of mating system shifts are typically associated with genetic bottlenecks.

Figure 2.3. Ratio of synonymous diversity ($\theta_\pi$) for related selfing and outcrossing populations and species in relation to: (a) the rate of selfing in populations with the higher degree of self-fertilization and (b) the estimated age (in years) of the transition to higher rates of selfing. In (a) the expected reduction in diversity due to selfing alone was $1/(1+F)$ where $F$ is the inbreeding coefficient (dashed line). Specific estimates and their sources are reported in Tables 2.S1-2.S4 (Appendix I).
Given the evidence for longer-term effects of linked selection reducing effective population size, another key question concerns whether an explicit coalescent framework can still be developed to infer the number of founding lineages. In a recent study (Brandvain et al. 2013) examining founding haplotypes in the highly selfing *C. rubella*, the authors concluded that, because of a long-term reduction in effective population size, the data were consistent with the number of founding individuals ranging from three to infinite, despite the very severe reduction in diversity and effective population size. Thus, in this example the authors pointed out that there was little ability to distinguish modes of selection on mating system, despite the relatively recent shift ~50,000 years ago. While similar analyses have not yet been conducted in other systems, the overall patterns to date suggest that selfing populations experience longer-term reductions in effective population size due to linked selection, restrictions of gene flow from outcrossing populations, and/or subsequent bottlenecks, with limited evidence for a direct collapse in diversity during the transition itself.

**Selective consequences of transitions to selfing**

There is growing evidence that the transition from outcrossing to selfing has important evolutionary consequences influencing speciation and extinction rates and the diversification of lineages (Igic and Busch 2013; Wright et al. 2013). In common with the causes of reduced genetic diversity in selfing populations, both genetic and demographic factors probably influence these evolutionary processes. Thus, a major challenge is to tease apart their relative importance. Here, we restrict our discussion to the question of why selfing lineages often appear to be ephemeral compared to outcrossing lineages. Because highly autogamous species are often short-lived and occupy unpredictable environments their populations are more likely than long-lived
outcrossing species to be susceptible to environmental catastrophes leading to local extinctions. Thus, demographic factors associated with small population size may, in part, contribute to higher extinction rates. However, genetic processes may also contribute towards higher extinction rates because the reduced effective population size of selfing populations is expected to reduce not only neutral diversity but also the efficacy of purifying selection (Glémin and Ronfort 2013a). This reduction in the intensity of natural selection can lead to a long-term accumulation of deleterious mutations, potentially contributing to increased extinction risk (Wright et al. 2013). One key question is how rapidly deleterious mutations accumulate following the shift to selfing, as this may, in part, determine the relative importance of demographic versus genetic effects in driving species’ extinction. Our simulation results highlight that the effects of linked selection can happen rapidly following the shift to selfing, implying that the early stages of deleterious mutation accumulation may be apparent soon after the transition.

Early tests of the hypothesis of a reduced efficacy of selection in selfers generally showed limited support for the prediction, but these studies were limited by the extent of the genome that was surveyed (Cutter et al. 2008; Haudry et al. 2008; Escobar et al. 2010). Furthermore, given the recent origin of most selfing populations, using substitution rates to test for relaxed selection may have limited power because of the difficulty in identifying the changes that have occurred since the shift to selfing. The use of polymorphism data from selfing populations and their outcrossing progenitors can provide a more powerful means to detect evidence for the reduced efficacy of selection in the genome of selfing populations (Brandvain et al. 2013; Hough et al. 2013), and to assess how rapidly reductions in the efficacy of selection can occur.
Figure 2.4. Shared and unique sequence polymorphisms between outcrossing and selfing populations of *Eichhornia paniculata* sampled from different parts of the geographical range: (a) proportion of synonymous polymorphisms, (b) ratio of nonsynonymous relative to synonymous polymorphisms. Shown are pairwise comparisons between a single diploid chromosome from an outcrossing individual and haploid chromosomes from two selfing individuals. Two selfers were used as the high homozygosity of selfing genomes reduces the effective number of chromosomes sampled from selfing individuals to approximately half of that of outcrossing individuals. Details of sampling and methods to estimate the proportion of polymorphisms are provided in the supplementary methods (Appendix I).
We used this approach in *E. paniculata* to identify polymorphisms in populations of two independently derived selfing lineages from the Caribbean and Central America, and their outcrossing progenitors from Brazil. The results indicate that many of the polymorphisms in populations of the selfing lineages were shared with outcrossing populations (Figure 2.4a). Moreover, these shared polymorphic sites were under similar strengths of selection as polymorphic sites unique to the outcrosser, as indicated by comparable ratios of nonsynonymous relative to synonymous polymorphisms (Figure 2.4b). In contrast, relaxation of selective constraints in selfing populations was most prominent for polymorphic sites unique to their populations, as indicated by elevated ratios of nonsynonymous relative to synonymous polymorphisms, when compared to the ratios obtained for shared polymorphic sites, and polymorphic sites unique to outcrossing populations. The signal was strongest for selfing populations from Central America, a result consistent with previous work indicating increased nonsynonymous mutations in these populations (Ness et al. 2012). Overall, these results combined with similar studies in *Capsella* (Hough et al. 2013) and *Collinsia* (Hazzouri et al. 2013), suggest that the reduced efficacy of purifying selection in selfing lineages is detectable even following quite recent transitions to selfing. These population genomic consequences appear to occur quickly following a shift to selfing, consistent with our earlier conclusions that the effects of genetic hitchhiking can occur rapidly due to strong reductions in effective population size.

**Unresolved questions and future work**

The simulations and comparative evidence from related selfers and outcrossers we report in this article indicate that current molecular data provide little evidence for bottlenecks during the
transition from outcrossing to selfing, and that the effect of genetic hitchhiking is likely a major contributor to the loss of genetic diversity in recently derived selfing populations. Of course these results do not rule out an important role for reproductive assurance in the transition to selfing, nor do they reject the possibility that the shift to high selfing rates leads to recurrent and severe bottlenecks. However, they do call into question the power to test hypotheses involving selective mechanisms using molecular data. Selection for reproductive assurance may not necessarily cause a signal of strong founder events, and low diversity in recently diverged selfing populations can result from a number of non-mutually exclusive causes. Given that partially selfing populations also show little molecular evidence for bottlenecks, the results to date do not support the hypothesis that early transitions to partial selfing are accompanied by strong bottlenecks due to selection for reproductive assurance.

Future work integrating detailed demographic and genomic information with variation in selfing rates should help to better determine the relative importance of the diverse factors affecting levels of diversity. With growing amounts of genomic data, explicit parameterizing of background selection and positive selection will be important for predicting the effects of genetic hitchhiking alone, and for assessing how important demographic factors may be in also contributing to reductions in diversity. In some systems, such as *E. paniculata*, demographic and mating system parameters are so strongly linked (Ness et al. 2010) that it may be difficult to uncouple their relative importance in natural populations. In such cases experimental approaches will be required. Field studies of experimental populations of different size and density that measure the strength of selection on traits directly promoting selfing would also be valuable in providing insights into the demographic context in which mating system shifts occur and their effects on genetic diversity.
Appendix I

Supporting information for Chapter 2

Supplementary methods

Background selection parameters

We used estimates available in the literature to estimate the strength of background selection for the results plotted in Figures 2.1 and 2.2. In Figure 2.1, the expected reduction in diversity due to background selection was modeled after (Glémin and Ronfort 2013b) as \( \exp\left\{ -\frac{U}{2(h+F-hF)s+C(1+F)} \right\} \), where \( U \) is the genomic deleterious mutation rate, \( C \) is the genomic recombination rate, \( s \) is the mean selection coefficient acting on deleterious mutations, and \( h \) is their dominance coefficients. \( U \) was calculated as genome size (\(~115\text{Mb} - \textit{Arabidopsis} \) Genome Initiative 2000) \times \) proportion of genome under selection (\(~0.17 - \text{Ossowski et al. 2010} \) \times \) per base substitution rate (\(~7\times10^9 - \text{Haudry et al. 2013} \) \times \) number of copies of chromosome (2) giving a value of \(~0.3 \). Our \( s \) was based on average estimates made from Ossowski et al. (2010) giving a value of \(~0.005 \). We chose \( h \) and \( C \) arbitrarily. We assumed all mutations were partially dominant (\( h=0.5 \)) and \( C \) was one magnitude higher than \( U \) (\( C=10 \)). The patterns remained consistent even when these parameters were altered except under very high values of \( C \), where high recombination rates limited the effects of linked selection. Nevertheless, higher homozygosity in selfing populations reduces the effective recombination rate and our values attempt to predict the effect of that reduction on the genome (Charlesworth et al 1993a; Nordborg 2000).
We performed forward population genetic simulations using SLiM (Messer 2013) to infer the strength of background selection in reducing synonymous diversity in selfing compared to outcrossing populations. The program SLiM implements an extended Wright-Fisher model with selection. It involves discrete generations, and parents for the next generation are drawn from the population in the previous generation at probabilities that are proportional to their fitness. Fitness is determined by all the mutations that are present within a given genotype. We started with a completely outcrossing population ($t=1$) comprised of 300 individuals with 100Mb genomes and constant genome-wide mutation ($\mu=7\times10^9$ - Ossowski et al. 2010) and recombination ($r=5\times10^8$) rates. Genomes comprised of alternating noncoding (NC) and coding (C) regions of various sizes; all sites in NC were neutral whereas 25% of sites in C were neutral and 75% of sites were deleterious under varied selection pressures as measured by $s$, the mean selection coefficient acting on deleterious mutations. The sizes of C:NC was 200:800bp, 350:650bp, or 500:500bp. With 200bp, 17.5% of the genome was under selection similar to the estimates made by Ossowski et al. (2010). We chose $s$ based on estimates obtained from natural populations of species in the Brassicaceae ($s=-0.005$ - Slotte et al. 2010) and varied this parameter. We ran simulations with $s=-0.005$, $s=-0.05$, or $s=-0.1$. All mutations had $h=0.5$ and there were no beneficial mutations. Simulations were run for $10N$ generations to reach stationary equilibria and then there was a split leading to a largely selfing population ($t=0.02$) of the same size. After running the simulations for another $4N$ generations, synonymous $\theta_w$ for selfing and outcrossing populations were estimated by randomly sampling 12 individuals from each population at every $0.1N$ interval. We estimated the ratio of mean diversity for selfers relative to outcrossers. We
plotted the mean ratio of diversity for selfers to outcrossers from five runs for each interval plotted in Figure 2.2.

**Analysis of unique and shared polymorphisms in selfers and outcrossers**

This analysis was conducted on the tristylos, diploid *Eichhornia paniculata* (Pontederiaceae). The Brazilian sample was obtained from a trimorphic population near Choro, Ceara; the two Caribbean samples were obtained from monomorphic populations near St. Elizabeth, Slippe, Jamaica and Chorera, Granma, Cuba; the two Central American samples were obtained from monomorphic populations near Oaxaca, San Mateo del Mar, near Tehuantepec, Mexico and Rio Las Lajas, Rivas, Nicaragua. Briefly, we extracted RNA from floral bud tissue for each sample and this was sequenced using the 100bp paired end protocol on Illumina HiSeq 2000. Reads from all samples were mapped using the Burrows-Wheeler Aligner (Li and Durbin 2009) and Stampy (Lunter G, Goodson M. 2011) to a *de novo* transcriptome reference generated from the selfers using Velvet-Oasis (Zerbino and Birney 2008; Schulz et al. 2012). We identified variants using the Unified genotyper program that is part of the Genome Analysis Toolkit (DePristo et al. 2011). We performed a number of steps to filter the variant calls. We removed sites with mapping quality less than 20 and variants with genotype quality less than 60. We removed loci containing sites that were identified as being heterozygous in multiple selfers. Finally, we mapped genomic reads from the Mexican sample to the transcriptome reference and removed loci that had <15x and >60x coverage from the analysis. After the filtering steps, we calculated the number of shared and unique nonsynonymous and synonymous polymorphisms using custom Perl scripts.
**Supplementary tables**

Table 2S1. Survey of diversity measurements for outcrossing-selfing pairs. Only studies that reported $\pi$ or $\theta_w$ for pairs comprising of outcrossing and selfing populations were considered. We did not consider studies that estimated diversity using less than five genes. References are in Table 2S4.

<table>
<thead>
<tr>
<th>Outcrossing species (population)</th>
<th>Selfing species (population)</th>
<th>No. of genes</th>
<th>Diversity measure</th>
<th>Ratio of diversity (selfer/outcrosser)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis lyrata</em> (PCR+PIN)</td>
<td><em>Arabidopsis lyrata</em> (PTP+RON+WAS)</td>
<td>18</td>
<td>$\pi$</td>
<td>0.36</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em> (PIR)</td>
<td><em>Arabidopsis lyrata</em> (KTT)</td>
<td>18</td>
<td>$\pi$</td>
<td>0.67</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em> (TSS)</td>
<td><em>Arabidopsis lyrata</em> (TC+TCA+TSSA)</td>
<td>18</td>
<td>$\pi$</td>
<td>1.00</td>
</tr>
<tr>
<td><em>Capsella grandiflora</em></td>
<td><em>Capsella rubella</em></td>
<td>26,52</td>
<td>$\pi$</td>
<td>0.15</td>
</tr>
<tr>
<td><em>Collinsia linearis</em></td>
<td><em>Collinsia rattanii</em></td>
<td>17</td>
<td>$\pi$</td>
<td>0.29</td>
</tr>
<tr>
<td><em>Eichhornia paniculata</em> (N.E. Brazil)</td>
<td><em>Eichhornia paniculata</em> (Caribbean)</td>
<td>10</td>
<td>$\pi$</td>
<td>0.42</td>
</tr>
<tr>
<td><em>Leavenworthia</em></td>
<td><em>Leavenworthia alabamica</em></td>
<td>8</td>
<td>$\pi$</td>
<td>0.00</td>
</tr>
<tr>
<td>Alabamica (a1 race)</td>
<td>(a4 race)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leavenworthia</td>
<td>Leavenworthia alabamica</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alabamica (a1 race)</td>
<td>(a2 race)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leavenworthia crassa</td>
<td>Leavenworthia crassa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(SI)</td>
<td>(selfing)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leavenworthia crassa</td>
<td>Leavenworthia crassa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(SI)</td>
<td>(mixed)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>Mimulus nasutus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000s</td>
<td>π</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarkia xantiana ssp.</td>
<td>Clarkia xantiana ssp.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xantiana</td>
<td>parviflora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$\theta_w^*$</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Did not report $\pi$; nevertheless, since Tajima's $D$ was close to 0, the ratio of diversity for the selfer relative to the outcrosser should be the same under $\theta_w$ and $\pi$

Table 2S2. Survey of Tajima’s $D$ measurements and bottleneck signals for outcrossing-selfing pairs. References are in Table 2S4.

<table>
<thead>
<tr>
<th>Outcrossing species (population)</th>
<th>Selfing species (population)</th>
<th>Tajima’s $D$ outcrosser</th>
<th>Tajima’s $D$ selfer</th>
<th>Bottleneck signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis lyrata (PCR+PIN)</td>
<td>Arabidopsis lyrata (PTP+RON+WAS)</td>
<td>0.09</td>
<td>0.12</td>
<td>No</td>
</tr>
<tr>
<td>Species 1</td>
<td>Species 2</td>
<td>$D_{TSS}$</td>
<td>$D_{TC+TCA+TSSA}$</td>
<td>Results</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em></td>
<td>(PIR)</td>
<td>0.17</td>
<td>0.63</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td><em>Arabidopsis lyrata</em> (KTT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em></td>
<td>(TSS)</td>
<td>0.58</td>
<td>0.25</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>(TC+TCA+TSSA)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Capsella grandiflora</em></td>
<td><em>Capsella rubella</em></td>
<td>-</td>
<td>-</td>
<td>Yes (Guo et al. 2014)</td>
</tr>
<tr>
<td><em>Collinsia linearis</em></td>
<td><em>Collinsia rattanii</em></td>
<td>-0.06</td>
<td>0.79</td>
<td>Bottlenecks or sampling</td>
</tr>
<tr>
<td><em>Eichhornia paniculata</em></td>
<td>(N.E. Brazil)</td>
<td>-0.13</td>
<td>-0.59</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>(Caribbean)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Leavenworthia alabamica</em></td>
<td>(a1 race)</td>
<td>-</td>
<td>-</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>(a4 race)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Leavenworthia alabamica</em></td>
<td>(a1 race)</td>
<td>-</td>
<td>-</td>
<td>Yes, but no more than outcrosser</td>
</tr>
<tr>
<td></td>
<td>(a2 race)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Clarkia xantiana</em></td>
<td><em>Clarkia xantiana ssp.</em></td>
<td>-1.40</td>
<td>-0.60</td>
<td>Yes because of higher variance in $D$</td>
</tr>
<tr>
<td></td>
<td>ssp. xantiana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>parviflora</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2S3. Survey of estimated age of selfing and selfing rate estimates for outcrossing-selfing pairs. References are in Table 2S4.

<table>
<thead>
<tr>
<th>Outcrossing species (population)</th>
<th>Selfing species (population)</th>
<th>Estimated selfing age</th>
<th>Selfing rate (outcrosser)</th>
<th>Selfing rate (selfer)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis lyrata</em> (PCR+PIN)</td>
<td><em>Arabidopsis lyrata</em> (PTP+RON+WAS)</td>
<td>10,000</td>
<td>0.09</td>
<td>0.79</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em> (PIR)</td>
<td><em>Arabidopsis lyrata</em> (KTT)</td>
<td>10,000</td>
<td>0.12</td>
<td>0.69</td>
</tr>
<tr>
<td><em>Arabidopsis lyrata</em> (TSS)</td>
<td><em>Arabidopsis lyrata</em> (TC+TCA+TSSA)</td>
<td>10,000</td>
<td>0.08</td>
<td>0.64</td>
</tr>
<tr>
<td><em>Capsella grandiflora</em></td>
<td><em>Capsella rubella</em></td>
<td>50,000-</td>
<td>-</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Collinsia linearis</em></td>
<td><em>Collinsia rattanii</em></td>
<td>500,000-</td>
<td>0.43</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Eichhornia paniculata</em> (N.E. Brazil)</td>
<td><em>Eichhornia paniculata</em> (Caribbean)</td>
<td>125,000</td>
<td>0.27</td>
<td>0.79</td>
</tr>
<tr>
<td><em>Leavenworthia alabamica</em> (a1 race)</td>
<td><em>Leavenworthia alabamica</em> (a4 race)</td>
<td>150,000</td>
<td>~0</td>
<td>0.90</td>
</tr>
<tr>
<td>Leavenworthia alabamica (a1 race)</td>
<td>Leavenworthia alabamica (a2 race)</td>
<td>( 48,000 )</td>
<td>~0</td>
<td>0.52</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------------------------</td>
<td>-----------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>Leavenworthia crassa (SI)</td>
<td>Leavenworthia crassa (selfing)</td>
<td>-</td>
<td>~0</td>
<td>0.67</td>
</tr>
<tr>
<td>Leavenworthia crassa (SI)</td>
<td>Leavenworthia crassa (mixed)</td>
<td>-</td>
<td>~0</td>
<td>0.32</td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>Mimulus nasutus</td>
<td>-</td>
<td>0.41</td>
<td>0.99</td>
</tr>
<tr>
<td>Clarkia xantiana ssp. xantiana</td>
<td>Clarkia xantiana ssp. parviflora</td>
<td>10,000-65,000</td>
<td>0.30</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 2S4. Literature sources for estimates reported in Tables S2.1-2.3.

<table>
<thead>
<tr>
<th>Outcrossing species (population)</th>
<th>Selfing species (population)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis lyrata (PCR+PIN)</td>
<td>Arabidopsis lyrata (PTP+RON+WAS)</td>
<td>Foxe et al. (2010)</td>
</tr>
<tr>
<td>Arabidopsis lyrata (PIR)</td>
<td>Arabidopsis lyrata (KTT)</td>
<td>Foxe et al. (2010)</td>
</tr>
<tr>
<td>Arabidopsis lyrata (TSS)</td>
<td>Arabidopsis lyrata (TC+TCA+TSSA)</td>
<td>Foxe et al. (2010)</td>
</tr>
<tr>
<td>Species 1</td>
<td>Species 2</td>
<td>Reference</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Capsella grandiflora</td>
<td>Capsella rubella</td>
<td>Slotte et al. (2010), Brandvain et al. (2010)</td>
</tr>
<tr>
<td>Collinsia linearis</td>
<td>Collinsia rattanii</td>
<td>Hazzouri et al. (2013)</td>
</tr>
<tr>
<td>Eichhornia paniculata</td>
<td>Eichhornia paniculata</td>
<td>Ness et al. (2010)</td>
</tr>
<tr>
<td>(N.E. Brazil)</td>
<td>(Caribbean)</td>
<td></td>
</tr>
<tr>
<td>Leavenworthia</td>
<td>Leavenworthia alabamica</td>
<td>Busch et al. (2011)</td>
</tr>
<tr>
<td>alabamica (a1 race)</td>
<td>(a4 race)</td>
<td></td>
</tr>
<tr>
<td>Leavenworthia</td>
<td>Leavenworthia alabamica</td>
<td>Busch et al. (2011)</td>
</tr>
<tr>
<td>alabamica (a1 race)</td>
<td>(a2 race)</td>
<td></td>
</tr>
<tr>
<td>(SI)</td>
<td>(selfing)</td>
<td></td>
</tr>
<tr>
<td>(SI)</td>
<td>(mixed)</td>
<td></td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>Mimulus nasutus</td>
<td>Y Brandvain, YW Lee, pers. comm., Aug 2013</td>
</tr>
<tr>
<td>Clarkia xantiana ssp. xantiana</td>
<td>Clarkia xantiana ssp. parviflora</td>
<td>Pettengill and Moeller (2011)</td>
</tr>
</tbody>
</table>
Chapter 3

Reduced efficacy of selection and purging of deleterious mutations accompany the evolution of selfing


Summary

The transition from outcrossing to selfing is predicted to reduce the genome-wide efficacy of selection because of the lower effective population size ($N_e$) that accompanies this change in mating system. However, strongly recessive deleterious mutations exposed in the homozygous backgrounds of selfers should be under strong purifying selection. Here, we examine estimates of the distribution of fitness effects (DFE) and changes in the magnitude of effective selection coefficients ($N_e s$) acting on mutations during the transition from outcrossing to selfing. Using forward simulations, we investigated the ability of a DFE inference approach to detect the joint influence of mating system and the dominance of deleterious mutations on selection efficacy. We investigated predictions from our simulations in the annual plant *Eichhornia paniculata*, in which selfing has evolved from outcrossing on multiple occasions. We used range-wide sampling to generate population genomic datasets and identified nonsynonymous and synonymous polymorphisms segregating in outcrossing and selfing populations. We found that the transition to selfing was accompanied by a change in the DFE, with a larger fraction of effectively neutral sites ($N_e s<1$), a result consistent with the effects of reduced $N_e$ in selfers. Moreover, an increased proportion of sites in selfers were under strong purifying selection
(\(N_{es}>100\)), potentially because of the exposure of recessive deleterious mutations. We conclude that the transition to selfing has been accompanied by the genome-wide influences of reduced \(N_e\) and strong purifying selection against deleterious recessive mutations, an example of purging at the molecular level.

**Introduction**

Mating system transitions provide important opportunities to investigate the influence of genetic drift and natural selection on plant genomes. The evolution of predominant self-fertilization (selfing) from cross-fertilization (outcrossing) is generally recognized as the most frequent evolutionary transition involving the reproductive systems of flowering plants (Stebbins 1957). Although selfing may be favored in the short-term, due to the transmission advantage of selfing genes and ability of individuals to set seed in pollen-limited conditions, selfing species represent only 10-15% of angiosperms and predominant selfing is often viewed as an evolutionary dead end (reviewed in Igic and Busch 2013; Wright et al. 2013; Chapter 2). The ephemeral nature of selfing lineages may be primarily due to the accumulation of deleterious mutations in selfing populations (Charlesworth et al. 1993b). Also, a reduced rate of fixation of beneficial mutations can limit the ability of selfing populations to adapt to changing environmental conditions (Glémin and Ronfort 2013a). Nevertheless, empirical support for selfing as an evolutionary dead end is mixed due to methodological and conceptual issues (reviewed in Takebayashi and Morrell 2001; Igic and Busch 2013). Characterizing the frequency of mutations segregating in selfing populations, and the types of selective pressures they are under, could provide insight into why some populations persist while others go extinct.
Theoretical models predict that the efficacy of selection should be lower in the genomes of selfing compared to outcrossing populations, because all else being equal, the effective population size ($N_e$) is reduced in selfing populations (Charlesworth et al. 1993a; Charlesworth and Wright 2001; Wright et al. 2008). In selfers, $N_e$ is reduced as a result of the genetic effects of restricted recombination, leading to hitchhiking of neutral variants and beneficial mutations, and also background selection against deleterious mutations. $N_e$ can be further reduced through demographic factors, such as population bottlenecks, because single individuals have the ability to found colonies (Baker 1955; Lloyd 1980; Charlesworth and Wright 2001; Pannell and Fields 2014). In genomes in which genetic and/or demographic influences reduce $N_e$, stochastic fixation or loss of rare alleles should become more prevalent and these processes may limit the efficacy of selection. Moreover, weak linkage among sites acted upon by opposing selective forces may interfere with the efficiency of selection (McVean and Charlesworth 2000; Comeron et al. 2008). Thus, a variety of demographic and genetic processes in selfing populations can potentially interact to reduce the efficacy of natural selection.

Early investigations of the genomic consequences of transitions from outcrossing to selfing provided limited support for theoretical predictions of a reduced efficacy of selection in selfing populations (e.g. *Arabidopsis* - Wright et al. 2002; *Mimulus* - Sweigart and Willis 2003; *Caenorhabditis* - Cutter et al. 2006; Triticeae - Haudry et al. 2008; Escobar et al. 2010). However, this lack of support may have arisen because only small portions of the genomes of were analyzed and/or the metrics used in these studies, e.g. substitution rates, may have limited power, particularly given the recent origin of most selfing populations. More recent comparisons of polymorphism data and codon usage bias from related outcrossing and selfing populations have provided some evidence for a reduced selection efficacy in selfers. The findings of elevated
levels of deleterious polymorphisms or a greater frequency of unpreferred codons in selfing *Arabidopsis* (Cao et al. 2011; Qiu et al. 2011), *Eichhornia* (Ness et al. 2012), *Capsella* (Qiu et al. 2011; Slotte et al. 2013; Brandvain et al. 2013), *Collinsia* (Hazzouri et al. 2013), and *Neurospora* (Gioti et al. 2013) are consistent with the hypothesis of relaxed selection in selfing populations.

The efficacy of selection in selfing populations will also be affected by how strongly recessive mutations are masked by heterozygosity. A negative relation between the dominance of mutations and the strength of their deleterious effects has been hypothesized (Simmons and Crow 1977; Crow and Simmons 1983). Although deleterious mutations are on average partially recessive (\( h \sim 0.2-0.25 \); Agrawal and Whitlock 2011; Manna et al. 2011), the more homozygous backgrounds of selfing populations offer fewer chances for deleterious mutations to be masked, and this should result in more effective selection against such mutations (Pollak 1987; Caballero and Hill 1992; Charlesworth 1992; Glémin 2007). Indeed, theory predicts that strongly deleterious, highly recessive mutations are more likely to be purged by selection after recurrent inbreeding (Hedrick 1994; Wang et al. 1999), even though both strongly and mildly deleterious mutations contribute to inbreeding depression (Charlesworth and Charlesworth 1999; Wang et al. 1999; Charlesworth and Willis 2009). If a large proportion of deleterious mutations are strongly recessive, selfing populations could persist as a result of the purging of deleterious mutation load (Barrett and Charlesworth 1991; Glémin and Ronfort 2013a), and the eventual fate of populations will be partly determined by levels of standing genetic variation (Orr and Unckless 2008; Glémin and Ronfort 2013a). Purging has been documented across mammals, insects, mollusks and plants, based on assays of fitness traits (Crnokrak and Barrett 2002). More recently, Szövényi et al. (2014) using divergence-based metrics found genomic evidence for stronger selective pressures acting on a moss species with haploid-dominant life cycle
undergoing intragametophytic selfing, when compared to an outcrossing relative. However, they found no evidence for the role of reduced $N_e$ and hypothesized that such an effect is more likely to be observed in plant species with diploid-dominant life cycles.

One reason for the weak empirical support for theoretical expectations on changes in selection efficacy following the transition to selfing could be that the effects of reduced $N_e$ and increased homozygosity counteract one other. A reduction in $N_e$ is expected to lower selection efficacy, whereas increased homozygosity should make selection more efficient by exposing recessive mutations. Glémin (2007) predicted that the effect of relaxed selection should be detected under a range of mutation and population size parameters in spite of the countervailing homozygosity effect. Moreover, he found that divergence-based measures of selection were more likely to reveal differences compared to those based on polymorphism data. However, recently derived selfing lineages, which may be most common, are unlikely to have had sufficient time to accumulate substitutions. In such cases, comparisons using intraspecific polymorphisms are likely to be more useful for detecting differences in selection efficacy between outcrossing and selfing populations.

An approach for investigating the counteracting effects of reduced $N_e$ and increased homozygosity using polymorphism data is to characterize the underlying distribution of fitness effects of mutations. In particular, the method of Keightley and Eyre-Walker (2007), jointly estimates selective and demographic parameters to infer the distribution of fitness effects (hereafter DFE) of new nonsynonymous mutations. This method allows a single population size change making it more robust to violations of assumptions due to static demographic histories, and summarizes the effective strength of selection ($N_e s$) acting on new nonsynonymous mutations using a gamma distribution. In selfing populations with reduced $N_e$ we expect a shift
in the DFE such that there are a higher proportion of effectively neutral mutations \((N_e s < 1)\). However, the DFE approach does not explicitly account for the dominance of mutations and assumes that all mutations are codominant \((h=0.5)\). More efficient purging of recessive deleterious mutations in selfing populations may shift the DFE in the opposite direction, resulting in a higher fraction of strongly deleterious mutations. It is therefore unclear if this approach is capable of uncovering the joint effects of the dominance of mutations and genetic drift on patterns of selection.

We had two main objectives in this study. The first was to determine the extent to which the DFE inference approach of Keightley and Eyre-Walker (2007) is useful for characterizing changes in selection efficacy following the transition from outcrossing to selfing. To do this we used forward population genetic simulations using realistic mutation, recombination, selection, dominance parameters, and genome structure. We explored how the estimated DFE shifts when the rate of selfing in a population increases. Our second objective was to investigate empirical support for theoretical predictions on the genomic consequences of the transition from outcrossing to selfing in the flowering plant *Eichhornia paniculata* using the DFE inference approach. This species is diploid \((n=8)\), annual, and belongs to the Pontederiaceae (Monocotyledoneae: Commelinales) (Angiosperm Phylogeny Group 2009), a small family of freshwater aquatics primarily native to the Neotropics.

*Eichhornia paniculata* populations possess a wide range of mating systems ranging from outcrossing to predominant selfing (Barrett and Husband 1990). This variation is associated with the evolutionary breakdown of tristyly, a floral polymorphism in which outbreeding populations are composed of three style morphs maintained by negative-frequency dependent mating. All *E. paniculata* populations that are tristyloous occur in the arid caatinga region of N.E. Brazil, where
the species commonly occurs in ephemeral ponds and ditches. Stochastic forces associated with small population size have destabilized the tristylosous polymorphism on multiple occasions resulting in the spread and fixation of semi-homostylyous selfing variants (Barrett et al. 1989; Husband and Barrett 1992; Husband and Barrett 1993). This fixation of these variants has accompanied transitions in morph structure from stylar trimorphism through dimorphism to monomorphism. Independent transitions to selfing associated with long-distance dispersal have given rise to selfing populations in the Caribbean and Central America (Barrett et al. 2009). Mid-styled semi-homostyles occur commonly on Jamaica and Cuba, whereas long-styled semi-homostyles are restricted to a few small isolated populations in Nicaragua and Mexico. Roughly 60% of the variation in outcrossing rates in *E. paniculata* can be explained by the morph structure of populations and the frequency of selfing variants (Barrett and Husband 1990). Because trimorphic populations are predominantly outcrossing and monomorphic populations exhibit high rates of selfing, we use morph structure as a proxy for outcrossing rate in our study. Studies of allozyme variation and nucleotide diversity indicate that monomorphic populations of *E. paniculata* have relatively low levels of heterozygosity (Glover and Barrett 1987; Barrett and Husband 1990; Husband and Barrett 1993; Ness et al. 2010) consistent with the effects of recurrent selfing in such populations and/or colonization bottlenecks associated with long-distance dispersal. Purging of genetic load has been documented in *E. paniculata* based on a study on fitness traits (Barrett and Charlesworth 1991), but as yet there is no molecular genetic evidence for this phenomenon.

Our investigations address the following specific questions: 1) Is the DFE inference approach of Keightley and Eyre-Walker (2007) able to separate the contrasting genomic signals of reduced selection efficacy and stronger purifying selection against recessive mutations
accompanying the transition to selfing? 2) Is there empirical support for these contrasting genomic signals based on the transition to selfing in *E. paniculata*? Our findings demonstrate that the DFE approach provides a valuable tool for studies of the genomic consequences of selfing, and that in *E. paniculata* transitions from outcrossing to selfing are accompanied by genome-wide influences of reduced $N_e$ and the purging of recessive deleterious mutations.

**Materials and methods**

*Genome structure and content of simulated datasets*

Using forward simulations, we investigated how patterns of selection change following the shift to selfing. We conducted forward simulations using the software SLiM (Messer 2013), which implements a Wright-Fisher model with selection and non-overlapping generations. We started with a completely outcrossing population ($t=1$) composed of 1000 individuals (census size, $N$) with 100 Mbp genomes and constant genome-wide mutation ($\mu=7\times10^{-9}$ per site per generation) and recombination ($r=5\times10^{-8}$ per site per generation) rates. Genomes were comprised of alternating 800 bp of noncoding (NC) and 200 bp of coding (C) DNA (Appendix II - Figure 3S1). All NC sites and 25% of C sites were neutral. For the remainder of the sites in the coding regions, we assumed a gamma distribution of selection coefficients with shape parameter ($\beta$) 0.3 and $N_s$ ranging from 0.5 to 95. We assumed there were no beneficial mutations. To isolate the effect of the dominance of mutations on selection efficacy, we supplied fixed $h$ of 0.2, 0.5, or 0.8 across the entire genome.
**Varying outcrossing rate (t), N, or r**

We conducted three independent sets of simulations varying the $t$, $N$ or $r$ parameters one at a time. In the first set, we ran simulations for $10N$ generations to reach stationary equilibrium and then introduced a split leading to a second population. The population resulting from the split had the same $N$ as the ancestral outcrossing population ($t=1$) but we decreased $t$ to 0.02, effectively simulating a shift to selfing. From this set of simulations we estimated the realized $N_e$ for the selfing population by calculating the reduction in synonymous diversity following the shift. After this estimation, we ran a second set of simulations in which the population resulting from the split had the same $N$ as the realized $N_e$ of selfers from the first set of simulations, but with the outcrossing rate unchanged ($t=1$). From the first set of simulations, we also calculated the effective recombination rate ($r_e$) of the selfing population using the equation $r_e = r_{outcrosser} \times (1-F_{is})$ where $F_{is}$ is the coefficient of inbreeding (Nordborg 2000). As we were unable to change $r$ midway during the simulations, we ran a third set of simulations where the ancestral outcrossing population had the same $r$ as the $r_e$ of selfers from the first set of simulations. For this third set of simulations we did not create a population split. We ran all simulations for another $6N$ generations to allow populations to reach equilibrium, as expected under coalescent predictions. During this period, we randomly sampled 8 individuals from each population at each $1N$ interval, and calculated the number of nonsynonymous and synonymous invariant sites and the number of nonsynonymous and synonymous mutations that had accumulated independently. We use these values to generate folded allele frequency spectra (hereafter AFS). We estimated the DFEs for all the simulated datasets using the approach of Keightley and Eyre-Walker (2007), as described in the section below. For each $N_e$s category in the DFE, we generated confidence intervals (CIs) based on 120 independent runs of simulations. We performed two-sample $t$-tests to compare the
DFEs generated for simulated outcrossing and selfing populations using R (R Development Core Team 2011). We also reported the expected DFE for the outcrossing population represented as a gamma distribution using the $\beta$ and $Ns$ parameters at the beginning of the simulations. To generate the expected DFE for the selfing population, we first scaled the $Ns$ parameter by the observed reduction in synonymous diversity accompanying the shift to selfing but leaving $\beta$ unchanged. Further, we multiplied $Ns$ by 1.98 in accordance with the expectation that selection should be scaled by the effective dominance levels experienced by mutations in selfing genomes (Caballero and Hill 1992).

*Jointly varying $h$ and $Ns$*

We ran separate simulations to assess the joint impact of the dominance level of mutations and the strength of purifying selection acting on them. Whereas most of the parameters remained unchanged in these additional simulations, we assumed $h=0.2$ and $Ns=95$ for 31.5% of the sites in the C region to simulate strongly recessive deleterious sites. The remaining 43.5% of the sites in the C region had $h=0.5$ and $Ns$ ranging from 0.5 to 85 to simulate partially codominant weakly and strongly deleterious sites. After simulating a split to a second predominant selfing population ($t=0.02$), we calculated the average fitness and number of neutral and deleterious mutations accumulated in individuals randomly sampled from the simulation output. As SLiM outputs haploid chromosomes, we randomly paired two genomes to construct a diploid individual. For a given site in each of these individuals, the fitness effect of nonsynonymous mutations that were present once was $1-hs$. If the same mutation occurred in both chromosome copies, its fitness effect was $1-s$ as they would be homozygous. The fitness of an individual was the multiplicative product of the fitness effect of all the mutations in their genome. Note, because highly selfing populations are mostly homozygous, this procedure does not simulate the true genotypic
composition of the population, but allows for a more direct comparison with outcrossers of fitness and deleterious mutation accumulation. We estimated the DFEs for the simulated datasets using the approach of Keightley and Eyre-Walker (2007), as described in the section below. For each $N_e s$ category in the DFE, we generated CIs based on 120 independent runs of simulations and performed a two-way analysis of variance with mating system and time as factors using R (R Development Core Team 2011).

Estimating selective and demographic parameters

We inferred the DFE for each dataset using the source code version of the DFE-$\alpha$ software available from Peter Keightley's laboratory website (http://lanner.cap.ed.ac.uk/~eang33/est-dfefiles.tar.gz). To ensure that our final parameter estimates from this maximum likelihood method were more likely to reach global optima, we randomized the starting values for mean selection against deleterious mutations ($N^*E(s)$), $\beta$, and $t_2$ parameters and ran the program 10 times. From the output estimates across the runs, we chose the parameters that resulted in the highest maximum likelihood. We report the proportion of mutations falling into a given $N_e s$ range of the DFE (<1, 1-10, 10-100 and >100).

Sampling of Eichhornia paniculata populations

We sampled open-pollinated seeds from 20 populations of $E. paniculata$ across the species range, including 10 outcrossing (tristyloous) and 10 selfing (predominantly semi-homostyloous) populations (Appendix II - Table 3S1). All outcrossing populations were from N.E. Brazil. Ness et al. (2010) found that the outcrossing populations from this region clustered in the same genetic structure group based on analyses done using the program InStruct. The genotypes used in this study were a subsample of those used by Ness et al. (2010). The selfing populations originated from Cuba ($n=3$ populations), Jamaica ($n=5$ populations), Mexico ($n=1$ population), and
Nicaragua ($n=1$ population). We germinated and grew plants under uniform glasshouse conditions at the University of Toronto. To maximize our sampling effort, we selected one individual per population following the scattered sampling approach (Wakeley and Lessard 2003; Städler et al. 2009). This sampling assumes that alleles coalesce faster within demes compared to between demes and, therefore, this sampling should maximize the number of unique alleles represented.

**Nucleic acid extraction and sequencing**

We extracted both RNA and DNA from floral buds of *E. paniculata* for sequencing, where RNA sequencing of multiple samples was used for de novo assembly and polymorphism analysis, whereas DNA was extracted from a single individual to aid in filtering out duplicate genes based on coverage. We chose floral buds as they contain both gametophytic and sporophytic genes thus maximizing our sampling breath. Buds can be easily and quickly removed and flash-frozen in liquid N$_2$, and contain low levels of secondary compounds. We extracted RNA using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich). The extracted RNA samples were used to make Illumina TruSeq RNA libraries that were sequenced using the 100 bp paired end protocol on Illumina HiSeq 2000 at the McGill University and Génome Québec Innovation Centre. The samples were sequenced across two lanes with 10 samples multiplexed in each lane. The outcrossing and selfing samples were evenly distributed across the two lanes to avoid lane effects confounding comparisons between mating systems. We further extracted genomic DNA from the floral buds of the Mexican sample using a modified variant of the CTAB extraction protocol (Doyle and Doyle 1987, 1990; Edwards et al. 1991) treating the lysed cells with ribonuclease A to remove contaminant RNA. An Illumina TruSeq DNA library was prepared from this sample and was sequenced on a separate lane on the Illumina HiSeq 2000. After sequencing, we
removed reads < 50 bp and reads that had more than 10% of "N" bases using a custom Perl script, retaining about 92% of the original data. The raw sequence data are available under accession number SRP049636 at the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and the associated BioProject accession number is PRJNA266681 (http://www.ncbi.nlm.nih.gov/bioproject/).

Assembly and identification of coding regions

As *E. paniculata* is a non-model system, we generated a new *de novo* transcriptome assembly. Although Ness et al. (2011) earlier generated a transcriptome reference for *E. paniculata*, we conducted an independent assembly as the greater number of genotypes, longer reads, and new assembly software facilitated more accurate construction of contigs. Nevertheless, we followed a similar approach as the previous study. We chose selfing genotypes to generate the consensus assembly under the expectation that their more homozygous genomes would limit the problem caused by alleles that appear heterozygous but actually belong to paralogous loci. We generated a *de novo* assembly from the six Jamaican transcriptome samples using the programs Velvet 1.2.08 (Zerbino and Birney 2008) and Oases 0.2.08 (Schulz et al. 2012). We found the best parameters for the assembly using VelvetOptimizer 2.2.4 (Zerbino et al. 2010) that indicated that k-mers of length 75-85 were optimal. The assembled reference transcriptome had a total size of 65.53 Mbp with an $N_{50}$ of 2.2kb (Appendix II - Figure 3S2). For the contigs in the assembled reference, we predicted the location of the coding regions through BLAST searches to known proteins in plant databases and removed contigs without any matches. We did this using a combination of tBLASTx (Altschul et al. 1990) to Viridiplantae database and GeneWise 2.4.1 (Birney et al. 2004). After identifying coding regions in assembled contigs, we trimmed bases.
before start and after stop codons using custom Perl scripts retaining 29336 loci at the end of this step.

Read mapping

After generating the reference assembly, we mapped all reads from the *E. paniculata* transcriptome samples and the genomic DNA sample to the assembly. First, we mapped short reads with Burrows-Wheeler Aligner (BWA, v0.6.2-r126) using default parameters (Li and Durbin 2009). Further, we used the BWA Sampe command to combine the paired end read mapping results together and Samtools view command (Samtools v0.1.18 r982:295; Li et al. 2009) to convert the mapping results into a binary alignment format (BAM format). After the first round of read mapping, we used the Stampy 1.0.20 software with default parameters (Lunter and Goodson 2011) to map more divergent reads, as well as to identify insertions and deletions (indels). At the end of this phase, ~92% of the filtered reads for each sample were successfully mapped. Next, we processed the read mapping output into a format required for variant calling software using four programs that were part of the Picard tools package 1.100 (1571) using default settings (http://picard.sourceforge.net): 1) SamFormatConverter that converted SAM output from Stampy into a BAM file, 2) ReorderSam that reordered mapped reads so that they conformed to the order of contigs in the assembly file, 3) AddOrReplaceReadGroups that grouped all reads sequenced from a given lane before variant calling, and 4) BuildBamIndex that indexed the BAM file. The processed BAM files are available under accession no. SRP049636 at the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra).

Variant calling and filtering

After read mapping, we identified single nucleotide polymorphisms (SNPs) segregating among the *E. paniculata* transcriptome samples using a set of programs from the Genome Analysis
Toolkit (GATK) v2.7-4-g6f46d11 (DePristo et al. 2011). We first used RealignerTargetCreator and IndelRealigner with default parameters and identified and realigned sequences within ~3 kb of an indel where erroneous mismatches were more likely to have occurred. Further, we ran UnifiedGenotyper with the BadCigar read filter, which uses a Bayesian model to call invariants sites, SNPs, and indels from all samples. For a given variant, we allowed for a maximum of six alternate alleles to be genotyped. We performed a number of filtration steps to minimize spurious SNP calls. From the UnifiedGenotyper output, we only retained sites for which the Phred scaled quality score (QUAL) was >60 and depth in each individual sequenced was >20. We only retained a SNP if the Phred scaled genotype quality (GQ) for all samples was >60. We also excluded 5 bp on either side of an indel where spurious SNP calls are likely to be made. For the SNPs identified in the outcrossing genotypes, we performed a test for Hardy-Weinberg equilibrium (Wigginton et al. 2005) as implemented in VCFtools (v0.1.11) (Danecek et al. 2011). We validated our SNP calls by comparisons to a range-wide polymorphism dataset for E. paniculata generated by Barrett et al. (2009) and Ness et al. (2010). These studies extensively sampled 225 E. paniculata individuals from 25 populations across Brazil, the Caribbean, and Central America for variants at 10 EST-derived nuclear loci. We performed a custom BLAST search comparing the 10 loci to our reference using the makeblastdb and blastn programs part of the NCBI BLAST+ toolkit, version 2.2.26 (Camacho et al. 2009). Of the 10 EST-derived nuclear loci generated by these authors, seven matched loci in our dataset that we had removed during the filtering stages. We aligned the nuclear loci and sequences for the best matching loci in our study using MUSCLE (Edgar 2004) as implemented in MEGA6 (Tamura et al. 2013) excluding sites with gaps. Only the alignment between EP0314 and the best matching locus from our study (Locus_13533) identified polymorphic sites segregating in both set of sequences and the region-
specific polymorphisms segregating in our samples matched those identified in the previous studies (Appendix II - Table 3S2).

*Filtering paralogous sequences*

If duplicated regions were assembled into a single contig, differences between paralogs might erroneously be called as a polymorphic difference at a single site. To address this we first removed loci containing sites that were heterozygous in all 10 selfing genotypes identified with the aid of VCFtools (v0.1.11) (Danecek et al. 2011). Such variants are unlikely to be real given the largely homozygous backgrounds of selfing populations and the occurrence of at least two independent transitions to selfing (Barrett et al. 2009). Secondly, we mapped genomic reads from a selfing genotype from Mexico to the transcriptome reference. We used the Samtools *depth command* (Samtools v0.1.18 r982:295; Li et al. 2009) to find depth per site and used a custom Perl script to calculate the mean across a given contig. We removed loci with less than 15x or greater than 60x genomic coverage (Appendix II - Figure 3S3) based on the assumption that coverage should be even across the entire genome.

*Removing shared variant sites*

To focus on contemporary selection pressures, we used only sites that were unique to each lineage in comparisons of the efficacy of selection between outcrossing and selfing populations of *E. paniculata*. To do so, we first used a custom Perl script to convert the UnifiedGenotyper VCF output into a FASTA format. Next, we identified polymorphic sites that were shared between outcrossing and selfing populations using a custom Perl script. Rather than removing those sites and breaking the reading frame of the sequence, we substituted bases at those sites in our FASTA files to "N" using a script from http://raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/perl-scripts.html
(selectSites.pl). We retained a total of 16416 transcripts at the end of the filtering stages. Note that our simulations used a similar approach, as only lineage-specific mutations were analyzed while inferring the DFE.

**Comparing outcrossing and selfing populations of Eichhornia paniculata**

We first processed the output from the SNP filtering steps and generated the necessary variables so we could compare the efficacy of selection in outcrossing and selfing populations. Because selfers are more highly homozygous than outcrossers, they have effectively half the number of chromosomes. To keep the number of chromosomes analyzed the same for comparisons, we generated a haploid copy of the diploid chromosome for each individual and randomly chose one of the bases at a given heterozygous site using a custom Perl script. As we only had two samples from Central America, we were unable to characterize the AFS and therefore we did not include the Central American populations in most of our comparisons. To keep the number of chromosomes sampled the same while performing comparisons, eight of the ten outcrossing genotypes from Brazil were randomly selected.

**Estimating the strength of selection acting on mutations in Eichhornia paniculata populations**

We used the Polymorphorama script (Andolfatto 2007; Haddrill et al. 2008) to generate locus-specific nonsynonymous and synonymous folded AFS and the number of invariant sites for the outcrossing and selfing populations. This program includes two-fold degenerate sites while calculating the AFS. Two-fold degenerate sites might inflate the number of nonsynonymous compared to synonymous polymorphisms resulting in biased conclusions on estimates of selection. However, Williamson et al. (2014) reported that inferences about the DFE were similar with or without the use of such two-fold sites. We generated 200 bootstrap replicate *E. paniculata* datasets, after resampling randomly across the loci using R (R Development Core
Team 2011), and generated the sums and means for each dataset. We inferred the DFE for each replicate dataset as described earlier in this section. For the resulting $N_{es}$ categories in the DFE, we generated mean values and 95% CIs across the 200 bootstrap replicates. To generate the CIs, we first excluded the top and bottom 2.5% of the bootstrap replicates. From the remainder of the replicates, we used the smallest and largest values to represent the lower and upper limits of the CI, respectively. We determined if the $N_{es}$ categories from the DFE differed between outcrossing and selfing populations using the randomization test following Keightley and Eyre-Walker (2007). It is important to note that a difference in one $N_{es}$ category of the DFE will not be independent of differences in the other categories. Finally, we used the $\beta$ and $-N^*E(s)$ parameters estimated by the DFE-α software to plot the cumulative proportion of nonsynonymous mutations. This allowed us to have more confidence in our inferences from the DFE, even if mutations of extremely large deleterious effect ($N_{es}>>100$) were absent in the samples analyzed.

**Results**

**Inferring the DFE of simulated outcrossing and selfing populations**

*Fixed dominance coefficients across the genome*

Using forward simulations, we investigated changes in selection efficacy in outcrossing and selfing populations after a split from a common outcrossing ancestor. After $6N$ generations, there was ~75% reduction in synonymous diversity following the shift to selfing. As the level of dominance across the entire genome was increased from 0.2 to 0.8, there was a directional shift in the estimated DFE of outcrossing populations, such that a larger fraction of sites were inferred to be under stronger levels of selection (Figure 3.1). Consistent with expectations, there was a
deficit of sites in the $N_{es} > 100$ category of the estimated DFE of the outcrossing population under $h=0.2$, when compared to the level expected from the DFE under an additive model. In contrast, the estimated DFE for selfing populations remained largely the same as the level of dominance increased. Under all three dominance levels, there was an excess of sites under stronger levels of selection than expected due to the $N_e$ reduction and increased homozygosity accompanying the shift to selfing. Under $h=0.2$, a significantly larger fraction of sites were in the $N_{es} > 100$ category, and a significantly smaller fraction of sites were in the $N_{es} 10$-100 categories of the estimated DFE of selfers when compared to outcrossers (Figure 3.1, Appendix II - Table 3S3), consistent with the effects of purging. In contrast, a significantly larger and smaller fraction of sites were in the $N_{es} < 1$ and $N_{es} 10$-100 categories, respectively, under $h=0.8$. These results indicate that with increasingly recessive mutations there is power to detect a purging effect in selfing populations, whereas with increasing dominance there is evidence for a greater proportion of effectively neutral mutations.

Figure 3.1. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing ($t=1.00$) and selfing ($t=0.02$) populations under various fixed dominance coefficients
(h=0.2, 0.5, or 0.8). N_e\sigma is the product of N_e and the selection coefficient (s). Shown are A) the expected DFEs, B) h=0.2, C) h=0.5, and D) h=0.8. Simulations illustrate the change in DFE for both populations after 6N generations after a split from the common outcrossing ancestor. The coding regions were under various selection coefficients (N_s=0.5-95) all sampled from a gamma distribution with shape parameter (\beta) 0.3. We generated the expected DFE for the outcrossing population represented as a gamma distribution using supplied \beta and N_s parameters. We generated the expected DFE for the selfing population by scaling supplied N_s parameter by the observed 75% reduction in synonymous diversity in selfers and multiplying it by 1.98 to account for effective dominance levels of mutations in selfing genomes while leaving \beta unchanged. We generated the observed DFEs by randomly sampling and generating allele frequency spectra using eight genomes from populations of size 1000. Shown are the mean proportions of sites for each N_e\sigma category and their respective confidence intervals based on 120 simulations.

Nonequilibrium demographic factors

The difference between the expected and estimated proportion of strongly selected sites in the selfing populations may reflect a failure to fully control for nonequilibrium demographic factors caused by the transition to selfing and the effects of Hill-Robertson interference. Because the DFE inference approach uses the synonymous AFS to estimate demographic parameters, we examined the neutral AFS produced in the simulated dataset. Under all dominance levels, the synonymous AFS for outcrossing and selfing populations were skewed towards rare variants, when compared to expectations under neutral equilibrium (Appendix II - Figure 3S4A, B). Such a skew was no longer observed in the outcrossing population, when we included mutations that were shared between both populations in our analyses (Appendix II - Figure 3S4C). This highlights that the skew was due to the use of unique mutations to estimate contemporary
selection pressures, which are likely rare given their recent origin. In contrast, selfers also showed a skewed AFS even when such mutations were included (Appendix II - Figure 3S4D). The inclusion of shared mutations did not affect our inferences about the effect of the dominance of mutations on selection efficacy (Appendix II - Figure 3S5). However, there was a larger difference between the observed and expected number of mutations in the \( N_s > 100 \) category under \( h=0.5 \) and \( h=0.8 \). Additionally, as the simulated genomes had a much larger fraction of noncoding compared to coding regions (Appendix II - Figure 3S1), our analysis may be biased by the fact that a large proportion of neutral sites were a considerable distance from the selected sites. Using synonymous mutations occurring only within the coding regions did not affect the observed differences between the DFEs of outcrossing and selfing populations (Appendix II - Figure 3S6), although the estimated CIs were larger.

Reduced \( N_e \) and stronger background selection

![Figure 3.2](image)

Figure 3.2. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing populations of two census sizes (\( N \)) under various fixed dominance coefficients. Shown are A) the expected DFEs, B) \( h=0.2 \), C) \( h=0.5 \), and D) \( h=0.8 \). Simulations illustrate the
change in DFE for both populations after 6N generations after the split of a population of N=250 from the common ancestral population of N=1000. The coding regions were under various selection coefficients (Ns=0.5-95) all sampled from a gamma distribution with shape parameter (β) 0.3. We generated the expected DFEs for the populations represented as a gamma distribution using supplied β and Ns parameters. We generated the observed DFEs by randomly sampling and generating allele frequency spectra using eight genomes from populations of size 1000. Shown are the mean proportions of sites for each Ns category and their respective confidence intervals based on 120 simulations.

To disentangle how reduced Ne and re accompanying the shift to selfing interact with the dominance of mutations, we varied these parameters one at a time, simulating outcrossing populations with equivalent reductions in effective size and recombination to our selfing populations. There was an increase in the fraction of sites in the lower Ns categories of the observed DFE under N=250, equivalent to the realized neutral Ne of the previously simulated selfing population, when compared to N=1000 (Figure 3.2). The observed DFEs for N=250 under all dominance levels harbored a greater fraction of sites in the larger Ns categories when compared to the expected levels. However, simulating this reduction in Ne did not show the purging signal we observed in selfing populations; there was no evidence for a significant increase in the proportion of strongly selected mutations in simulations with recessive mutations (compare Figure 3.2B with Figure 3.1). The DFE of the outcrossing population, with r=1.96×10^{-9} per site per generation, equivalent to the realized re of previously simulated selfing populations, had a larger fraction of mutations in smaller Ns categories when compared to the DFE of the outcrossing population, with r=5×10^{-8} per site per generation under h=0.5 and under h=0.8
(Figure 3.1 and S7). However, this change was not as large as the level expected due to the \( N_e \) reduction accompanying the shift to selfing. As with the simulations of outcrossing populations subjected to reduced \( N_e \), we found no evidence of the purging signal of a larger fraction of strongly deleterious mutations compared with the ancestral outcrosser, again showing that this signal is specific to the transition to selfing. Thus, these simulations indicate that the signal of increased proportions of strongly selected sites is a unique effect of the transition to selfing, and driven by the purging effect of increased homozygosity.

**Mixed dominance coefficients across genome**

![Graph showing the type and effects of mutations accumulated by four outcrossing (black) and four selfing (grey) individuals over 6\( N \) generations after a split from a common outcrossing ancestor.]

**Figure 3.3.** The type and effects of mutations accumulated by four outcrossing (black) and four selfing (grey) individuals over 6\( N \) generations after a split from a common outcrossing ancestor. A) Using forward population genetic simulations, we calculated the number of nonsynonymous
and synonymous mutations accumulated in eight haploid genomes randomly sampled from populations of size 1000. The eight haploid genomes were randomly paired to create four diploid individuals. B) Shown are 95% confidence intervals for the fitness of four individuals.

Figure 3.4. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing ($t=1.00$) and selfing ($t=0.02$) populations when both dominance and selection coefficients were jointly varied. Within the coding region of the genome supplied in the simulations, 31.5% of sites had $h=0.2$ and $Ns=95$, 43.5% of the sites had $h=0.5$ and $Ns$ ranging from 0.5 to 85, and the remaining had $h=0.5$ and $Ns=0$. Simulations illustrate the change in DFE for both populations over $6N$ generations after a split from the common outcrossing ancestor. We
generated DFEs by randomly sampling and generating allele frequency spectra using eight genomes from populations of size 1000. $N_e s$ is the product of $N_e$ and the selection coefficient ($s$). Shown are the mean proportions of sites for each $N_e s$ category and their respective confidence intervals based on 120 simulations.

When both dominance and selection coefficients were jointly varied, selfers accumulated more nonsynonymous mutations compared to outcrossers even though both accumulated the same number of synonymous mutations (Figure 3.3A). This increased accumulation of nonsynonymous mutations resulted in a greater fitness decline of selfers (Figure 3.3B). We further investigated how such mutation accumulation influenced the DFE for outcrossing and selfing populations. After the split from the common outcrossing ancestor, the estimated DFE of the selfing population shifted such that a significantly larger fraction of sites were in the $N_e s < 1$ category at $1N$ and $3N$ generations (Figure 3.4, Appendix II - Table 3S4). Also, there was a significantly larger fraction of sites in the $N_e s > 100$ category during the first $4N$ generations. The interaction between mating system and time was a significant factor influencing the proportion of mutations in the $N_e s < 1$ and $N_e s > 100$ categories. Moreover, the number of sites in each $N_e s$ category of the DFE fluctuated over generations. When there were both weakly selected additive sites and strongly selected recessive sites in the simulated genome, the DFE approach inferred the occurrence of more very weak and very strong deleterious mutations in selfers, although such effects varied with the time since the transition to selfing. Thus, with a model with varying dominance coefficients, it is possible to detect signals of both relaxed purifying selection and purging, although the relaxed selection signal is less prevalent, despite an observed decline in fitness in selfing populations.
Inferring the DFE of outcrossing and selfing populations of *Eichhornia paniculata*

*Selection and demographic parameters*

![Graph showing folded nonsynonymous and synonymous allele frequency spectra for outcrossing and selfing populations of *Eichhornia paniculata*. Haploid chromosomes from one individual from each of eight outcrossing populations from N.E. Brazil and eight selfing populations from the Caribbean were used to generate the frequency spectra. Eight outcrossing individuals from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparison.]

Figure 3.5. Folded nonsynonymous and synonymous allele frequency spectra for outcrossing and selfing populations of *Eichhornia paniculata*. Haploid chromosomes from one individual from each of eight outcrossing populations from N.E. Brazil and eight selfing populations from the Caribbean were used to generate the frequency spectra. Eight outcrossing individuals from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparison.
Figure 3.6. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing *Eichhornia paniculata*. $N_e s$ is the product of $N_e$ and the selection coefficient ($s$). Eight Caribbean selfing and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. Error bars on top of each $N_e s$ category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci. We used a randomization test (see Keightley and Eyre-Walker 2007) to compare outcrossing and selfing populations and to assess significance at 0.5% level (indicated by *).

Of the 16416 loci, 4485 were polymorphic in outcrossers and 1586 were polymorphic in selfers; polymorphic loci in outcrossers and selfers had between 1-40 and 1-20 segregating sites, respectively, after excluding variants shared between outcrossing and selfing populations, with the distribution of polymorphisms per locus right-skewed (Appendix II - Figure 3S8). Mean $\pi_n$. 

77
for outcrossers and selfers was $4.93 \times 10^{-4}$ and $8.99 \times 10^{-5}$, respectively; mean $\pi$ for outcrossers and selfers was $3.09 \times 10^{-3}$ and $5.62 \times 10^{-4}$, respectively. We identified around 20388 nonsynonymous and 32384 synonymous polymorphic sites in outcrossers and around 3613 nonsynonymous and 5730 synonymous polymorphic sites in selfers for estimating selective and demographic parameters (Appendix II - Table 3S5). There was an excess of rare nonsynonymous polymorphisms compared to synonymous polymorphisms in both outcrossing and selfing populations (Figure 3.5). Selfing populations had a slight deficit of rare synonymous polymorphisms compared to outcrossing populations. Randomization tests, as implemented in Keightley and Eyre-Walker (2007), indicated there were significantly fewer sites in the $N_\epsilon s$: 1-10, 10-100 categories and a significantly greater proportion of sites in the $N_\epsilon s>100$ in selfers compared to outcrossers (Figure 3.6). Although we did not detect a significant difference for the $N_\epsilon s<1$ category at the 0.5% level, there was a general trend for a larger number of mutations in this category for selfing populations. Indeed, our plots of the cumulative proportion of nonsynonymous mutations using the $\beta$ and -$N^*E(s)$ parameters indicated that selfers had a marginally larger fraction of mutations with $N_\epsilon s<3$ when compared to outcrossers (Appendix II - Figure 3S9). The plots also showed that ~63.7% of the mutations had $N_\epsilon s>150$ in selfing populations, whereas only ~46.2% of the mutations had $N_\epsilon s>150$ in outcrossing populations.

Data filtering

As inferring the DFE is dependent on the AFS being compared, we investigated how different quality cut-offs and filtering affected the patterns observed. A single site quality cut-off for both invariant and variant sites could disproportionately exclude invariant sites. However, in our case reducing the quality cut-off for both sites did not affect the DFE inference for outcrossing and selfing populations (Appendix II - Figure 3S10). We also performed more stringent filtering to
remove potential paralogs by excluding an additional 245 loci that contained at least one site heterozygous in two or more selfers (Appendix II - Table 3S6). Removing such loci from the comparisons did not alter our inferences (Appendix II - Figure 3S11). Less than 3% of polymorphic sites across outcrossing genotypes departed from Hardy-Weinberg equilibrium at the 5% level and these sites were localized to 73 loci. Again, removal of these loci did not affect our inferences on the differences between outcrossing and selfing genotypes (Appendix II - Figure 3S12). Finally, we investigated how the relative proportions of nonsynonymous and synonymous polymorphisms per loci influenced our results. As the distribution of $\pi_n/\pi_s$ was right skewed (Appendix II - Figure 3S13), we excluded 135 loci that had $\pi_n/\pi_s > 1$ in either outcrossers or selfers and repeated the DFE analyses. Although we still observed an excess of sites in the $N_e s > 100$ category in selfers, there was substantial overlap in the proportion of sites in the $N_e s < 1$ category between outcrosser and selfers (Appendix II - Figure 3S14, Table 3S7).

**Pooling independent shifts to selfing**

We investigated whether pooling additional samples from an independent transition to selfing influenced the AFS and the inferred DFE. As the inclusion of the two additional samples from Central America resulted in a total of 10 selfing samples, we repeated the analyses using all 10 outcrossing genotypes to keep the number of chromosomes compared the same. With the inclusion of the additional selfing genotypes, we still observed an excess of rare nonsynonymous polymorphisms compared to synonymous polymorphisms in both outcrossing and selfing populations (Appendix II - Figure 3S15). In the synonymous AFS, selfing populations had a deficit of singletons and an excess of doubletons compared to outcrossing populations. Moreover, we found that using the randomization tests, as implemented in Keightley and Eyre-
Walker (2007), the underlying DFE and all discrete $N_e s$ categories (Appendix II - Figure 3S16) were significantly different between outcrossing and selfing populations at the 0.5% level.

**Discussion**

Our forward population-genetic simulations demonstrate that reduced $N_e$ and increased purging accompanying the transition to selfing can be detected by estimating the underlying distribution of deleterious mutational effects segregating in populations. Our simulations show that this is not expected under a single dominance coefficient for amino acid mutations, but can be explained if there is a mixture of strongly selected recessive mutations that experience effective purging in selfing populations, and more weakly selected additive mutations subject to relaxed selection. Our empirical study of *E. paniculata* revealed a small increase in the proportion of effectively neutral amino acid mutations and a significant increase in the proportion of strongly selected sites in selfing compared to outcrossing populations. The spread in the distribution of selective coefficients observed in both the simulated and empirical datasets is consistent with effects of both purging and a reduced efficacy of selection in selfing populations. We now consider the effects of reduced $N_e$ and increased homozygosity on the genome-wide distribution of polymorphisms.

*Consequences of reduced $N_e$ in selfing populations*

Our simulations revealed that there was a change in the number of sites under weaker purifying selection following the transition from outcrossing to selfing as a result of reduced $N_e$ and $r_e$. Further, relaxed selection was driven largely by partially dominant mutations rather than by mutations that are recessive. The observed pattern is consistent with theoretical predictions that
selection should be less efficient in selfers (Charlesworth and Wright 2001). In both the simulated and empirical results the estimated magnitude of relaxed selection appears to be small (Figures 4 and 6), and there may be several reasons for this. First, it is possible that even a large reduction in \( N_e \) does not substantially alter the DFE in selfing populations if a significant proportion of mutations are strongly selected \((N_{es} >>> 1)\). Second, the relatively recent evolutionary transition to selfing in \( E. paniculata \) (see Ness et al. 2010) might not have allowed for sufficient time for the effects of reduced \( N_e \) to be detected at genome-wide scales. However, our simulations examining the shift to selfing over \( 6N \) generations did not find a very large reduction in the \( N_{es} <1 \) category of the DFE and patterns of relaxed selection disappeared after \( 3N \) generations. The reduced \( N_e \) accompanying the shift to selfing may have only moderately detectable effects on the genome-wide reduction in selection efficacy. Even so, our simulation results imply that even subtle shifts detected using the DFE approach may be associated with early and cumulative declines in fitness in selfing populations (Figure 3.3).

The small magnitude of the increase in relaxed selection accompanying the shift to selfing in the forward simulations compared to expected levels might suggest that the DFE inference approach underestimates the effects of reduced \( N_e \) on selection efficacy. If the method underestimated the magnitude of relaxed selection, it could explain why the observed shift in the empirical data towards effectively neutral sites in the DFE of selfers was relatively weak. An alternate possibility is that even datasets of genome-wide polymorphism might have limited power to detect patterns of relaxed selection (and see Glémin 2007). The pattern of relaxed selection in the empirical dataset became less apparent when we excluded loci with \( \pi_{w}/\pi_s > 1 \) indicating that they might be contributing to genome-wide patterns. Furthermore, even low levels
of migration among populations could have resulted in a greater chance that fewer distinct genotypes were sampled.

Our empirical results demonstrate demographic effects associated with the colonization history of selfing populations of *E. paniculata*. Pooling genotypes from independent shifts to selfing had a significant influence on the AFS as indicated by the excess of doubletons in selfers. The Central American genotypes were fixed for alternate alleles compared to samples from the Caribbean and this probably explains this result. *Eichhornia paniculata* occurs in ephemeral aquatic habitats and populations experience dramatic annual variation in size, including frequent local extinctions (Barrett and Husband 1997; Husband and Barrett 1998). Recent studies suggest that nonsynonymous mutations may reach equilibrium faster than synonymous mutations after population bottlenecks; therefore comparisons of the two could generate spurious signals of relaxed selection (Pennings et al. 2014; Simons et al. 2014). Because *E. paniculata* only colonized the Caribbean ~125,000 years ago (Ness et al. 2010), populations from that region might not have had sufficient time to recover from founding events. If the fit of demographic models for selfing populations was less accurate compared to outcrossing populations, our signal of relaxed selection could, in part, reflect this effect. However, we found that even when the synonymous spectra did not meet neutral expectations, the inference of the DFE was similar to the case when the spectra did match such expectations. Although the demographic correction implemented in the DFE inference approach should account for distortions of the AFS away from equilibrium expectations, it is unclear if the marginal trend of relaxed selection in the empirical data, which was significant when samples from separate origins were pooled, was influenced by violations of neutral expectations. Our simulations gave us qualitatively similar differences in the DFE to the empirical data and provided evidence for a fitness decline and the
accumulation of deleterious mutations in selfing populations (Figure 3.3). These results suggest that the weak trend observed in our empirical results reflects a true signal of the reduced efficacy of selection in selfing populations.

*Exposure of recessive deleterious mutations*

We observed a large proportion of strongly deleterious sites in the DFEs obtained from both simulated and empirical data. By using forward simulations, we found that the DFE approach was able to infer the effect of varying dominance levels of mutations on patterns of selection. As deleterious mutations became more dominant, the effective strength of selection acting against them in outcrossing populations increased. In contrast, the dominance level of mutations had little effect in selfing populations, where the largely homozygous background of genotypes exposes the deleterious effects of recessive and dominant mutations. Further, our simulations also indicated that the DFE approach was able to identify how both partially dominant and strongly recessive deleterious mutations result in differences in the selection efficacy experienced by selfers and outcrossers. At various intervals of time, linkage between neutral, weakly and strongly deleterious mutations, under variable dominance coefficients, may have complicated attempts to disentangle their individual effects on the DFE. Whereas linkage effects may have contributed to the observed variation in the proportion of sites under a given strength of selection category over time, there was no consistent pattern in the way fluctuations occurred. Overall, our simulation results indicated that the empirical patterns of increased accumulation of effectively neutral and strongly deleterious mutations occur across a wide range of selection and dominance coefficients, after transitions to selfing. Previous experimental studies in *E. paniculata*, involving multigenerational fitness comparisons of phenotypic traits following selfing and outcrossing, were unable to detect significant inbreeding depression in selfing
populations from the Caribbean and the authors proposed that this was largely a result of purging of genetic load (Barrett and Charlesworth 1991). Our findings based on data from genome-wide polymorphisms are consistent with this earlier study in providing molecular evidence to support the purging hypothesis.

A key question is whether the observed excess of strongly deleterious mutations in selfing populations was, in part, due to the methods used to infer selection efficacy. The approach of Keightley and Eyre-Walker (2007) is known to imprecisely infer the DFE when the distribution is multimodal (Kousathanas and Keightley 2013), and also to overestimate the strength of purifying selection in the face of linked selection (Messer and Petrov 2013). It is worth noting that the DFE estimation approach infers the proportion of mutations that are strongly deleterious \( (N_s > 100) \) from the \( \beta \) parameter, even if there are no mutations of this strength segregating in samples. However, plotting the DFE as a continuous distribution indicated there was a greater spread of selection coefficients in selfers. Furthermore, we observed that the magnitude of the shift towards effectively neutral sites in simulated selfing populations was lower than expected as a result of reduced \( N_e \), or the level experienced by simulated outcrossing population with a census size equivalent to the realized \( N_e \) of the selfers. Both of these results suggest that countervailing homozygosity effects stifle a unidirectional change in the DFE. Therefore, the observed patterns are consistent with an effect of both purging and reduced efficacy of selection in selfing populations, even if estimates of the magnitude of the effect may be imprecise.

At this stage we are unable to assess the extent to which a particular assumption of the DFE inference approach, namely that selection at one site is independent of selection acting on other sites, was violated in our study. It seems probable that in our selfing populations this
assumption would be violated, with many neutral sites linked to deleterious ones. In this case, the DFE approach might falsely indicate that there were a greater proportion of deleterious sites. However, over short distances in the genome our bootstrapping by locus approach addresses some of the uncertainty due to linkage blocks, especially in selfing populations, as some bootstrap replicates may not contain all of the genes within a given block. Furthermore, our simulations suggest that the patterns we observed are expected specifically under a model that incorporates both strongly recessive deleterious mutations and slightly deleterious, more additive mutations, consistent with a role for purging of harmful recessive alleles and an accumulation of weakly deleterious mutations.

Finally, we have not considered the effects of beneficial mutations in interpreting the results of our study. Recessive beneficial mutations may also be exposed in selfing populations leading to selective sweeps. As beneficial mutations are likely to be fixed rapidly after they appear in selfing populations it seems unlikely that they would have been segregating in the small number of *E. paniculata* samples used in our study. The interplay between recessive beneficial and deleterious mutations is important to consider as linkage between such sites should increase the time taken for sweeps to occur (Hartfield and Glémin 2014). Future studies that investigate the signature of sweeps and estimates of the rate of adaptation in selfing populations should consider the influence of strongly deleterious mutations as our results indicate that they can influence genome-wide patterns of selection.
Appendix II

Supporting information for Chapter 3

Supplementary figures

Figure 3S1. Genome structure implemented in the simulations. Each genome was 100 Mbp comprised of alternating 800 bp noncoding (NC) and 200 bp coding (C) regions. All sites in NC and 25% of sites in C were neutral with $h=0.5$ and the remaining 75% of C were under varying selection coefficients.
Figure 3S2. Length distribution of all contigs from the *de novo* assembly of six Jamaican genotypes of *Eichhornia paniculata* using Velvet-Oases. The summary statistics for the assembly are shown in the table within the figure. The $N_{50}$ value shows that 50% of all bases in the assembled reference are in contigs of size corresponding to this value or larger.

<table>
<thead>
<tr>
<th>Summary statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{50}$</td>
<td>2188 bp</td>
</tr>
<tr>
<td>Total size of assembled reference</td>
<td>65.33 Mbp</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>55655</td>
</tr>
<tr>
<td>Number of contigs &gt;1kb</td>
<td>21255</td>
</tr>
</tbody>
</table>

Figure 3S3. Coverage distribution of genomic reads from one selfing Mexican genotype of *Eichhornia paniculata* mapped to the *de novo* transcriptome reference.
Figure 3S4. Folded synonymous allele frequency spectra for simulated outcrossing ($t=1.00$) and selfing ($t=0.02$) populations under various fixed dominance coefficients ($h=0.2, 0.5, \text{ or } 0.8$). We generated the allele frequency spectra by randomly sampling eight genomes from populations of size 1000. To generate the frequency spectra we either used A) mutations unique to outcrossing population, B) mutations unique to selfing population, C) mutations unique to outcrossing populations and shared between outcrossing and selfing populations, or D) mutations unique to selfing populations and shared between outcrossing and selfing populations. The white open bars depict the neutral equilibrium expectation.
Figure 3S5. Distribution of fitness effects (DFE) of new nonsynonymous mutations or simulated outcrossing ($t=1.00$) and selfing ($t=0.02$) populations under various fixed dominance coefficients ($h=0.2$, 0.5, or 0.8) when both mutations unique to each population and mutations shared between populations were used to estimate selective and demographic parameters. $N_{e}s$ is the product of $N_e$ and the selection coefficient ($s$). Shown are A) the expected DFEs, B) $h=0.2$, C) $h=0.5$, and D) $h=0.8$. Simulations illustrate the change in DFE for both populations after 6$N$ generations after a split from the common outcrossing ancestor. We generated the expected DFE for the outcrossing population represented as a gamma distribution using supplied $\beta$ and $N_s$ parameters. We generated the expected DFE for the selfing population by scaling supplied $N_s$ parameter by the observed 75% reduction in synonymous diversity in selfers and multiplying it by 1.96 to account for effective dominance levels of mutations in selfing genomes while leaving $\beta$ unchanged. We generated observed DFEs by randomly sampling and generating allele
frequency spectra using eight genomes from populations of size 1000. The coding regions were under various selection coefficients ($N_s$=0.5-95) all sampled from a gamma distribution with shape parameter ($\beta$) 0.3. Shown are the mean proportions of sites for each $N_s$ category and their respective confidence intervals based on 120 simulations.

Figure 3S6. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing ($t$=1.00) and selfing ($t$=0.02) populations under various fixed dominance coefficients ($h$=0.2, 0.5, or 0.8) when synonymous mutations occurred only within coding regions were used to estimate demographic parameters. $N_{es}$ is the product of $N_e$ and the selection coefficient ($s$). Shown are A) the expected DFEs, B) $h$=0.2, C) $h$=0.5, and D) $h$=0.8. Simulations illustrate the change in DFE for both populations after $6N$ generations after a split from the common outcrossing ancestor. We generated the expected DFE for the outcrossing population represented.
as a gamma distribution using supplied $\beta$ and $N_s$ parameters. We generated the expected DFE for the selfing population by scaling supplied $N_s$ parameter by the observed 75% reduction in synonymous diversity in selfers and multiplying it by 1.96 to account for effective dominance levels of mutations in selfing genomes while leaving $\beta$ unchanged. We generated observed DFEs by randomly sampling and generating allele frequency spectra using eight genomes from populations of size 1000. The coding regions were under various selection coefficients ($N_s$=0.5-95) all sampled from a gamma distribution with shape parameter ($\beta$) 0.3. Shown are the mean proportions of sites for each $N_{es}$ category and their respective confidence intervals based on 120 simulations.

Figure 3S7. Distribution of fitness effects (DFE) of new nonsynonymous mutations for simulated outcrossing populations with recombination rate ($r$) of either $5 \times 10^{-8}$ or $1.96 \times 10^{-9}$ per site per
generation under various fixed dominance coefficients. Shown are A) the expected DFEs, B) $h=0.2$, C) $h=0.5$, and D) $h=0.8$. Simulations illustrate the DFE for both populations after 16N generations. The coding regions were under various selection coefficients ($Ns=0.5-95$) all sampled from a gamma distribution with shape parameter ($\beta$) 0.3. The population had a constant mutation rate ($\mu=7 \times 10^{-9}$ per site per generation) and population size ($N=1000$) during the simulation runs. We generated the expected DFE under $N=1000$ for an outcrossing population represented as a gamma distribution using supplied $\beta$ and $Ns$ parameters. We expected DFE under $N=250$ by scaling supplied $Ns$ parameters while leaving $\beta$ unchanged. We generated the observed DFEs by randomly sampling and generating allele frequency spectra using eight genomes. Shown are the mean proportions of sites for each $N_e,s$ category and their respective confidence intervals based on 120 simulations.

Figure 3S8. Distribution of polymorphisms per contig for eight outcrossing *Eichhornia paniculata* populations from N.E. Brazil and eight selfing populations from the Caribbean. Of the 16416 loci, 4485 were polymorphic in outcrossers and 1586 were polymorphic in selfers.
Figure 3S9. Cumulative distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata*. $N_e s$ is the product of $N_e$ and the selection coefficient ($s$). Eight Caribbean selfing samples and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. The shape ($\beta$) and mean selection against deleterious mutations ($N^*E(s)$) parameter estimates from the approach of Keightley and Eyre-Walker (2007) were used to plot the gamma distributions. For outcrossers, $\beta=0.3691$, $-N^*E(s)=365.16$; for selfers $\beta=0.2517$, $-N^*E(s)=3112.84$. 
Figure 3S10. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* under varying quality cut-off scores (QUAL) for invariant and variant sites. $N_e s$ is the product of $N_e$ and the selection coefficient ($s$). Eight Caribbean selfing samples and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. Error bars for each $N_e s$ category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci.
Figure 3S11. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of Eichhornia paniculata after 245 loci containing sites heterozygous in two or more selfing populations were excluded from analyses. $N_{es}$ is the product of $N_e$ and the selection coefficient ($s$). Eight Caribbean selfing samples and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. Error bars for each $N_{es}$ category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci.
Figure 3S12. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* after 73 loci containing sites that violated assumptions of Hardy-Weinberg equilibrium in outcrossing populations were excluded from analyses. $N_e s$ is the product of $N_e$ and the selection coefficient ($s$). Eight Caribbean selfing samples and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. Error bars for each $N_e s$ category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci.
Figure 3S13. Distribution of $\pi_n/\pi_s$ for polymorphic contigs among eight outcrossing *Eichhornia paniculata* populations from N.E. Brazil and eight selfing populations from the Caribbean. Of the 16416 loci, 4485 were polymorphic in outcrossers and 1586 were polymorphic in selfers.

Figure 3S14. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata* after 135 loci with $\pi_n/\pi_s>1$ in either
population were excluded from analyses. \( N_s s \) is the product of \( N_e \) and the selection coefficient (s). Eight Caribbean selfing samples and eight outcrossing individuals were used to generate the DFEs. Eight outcrossing samples from the 10 that were sequenced were randomly selected to keep the number of chromosomes sampled the same while performing the comparisons. Error bars for each \( N_s s \) category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci.

![Graph](image)

Figure 3S15. Folded nonsynonymous and synonymous allele frequency spectra for outcrossing and selfing populations of *Eichhornia paniculata*. We used haploid chromosomes from one individual each from 10 outcrossing and 10 selfing populations to generate the frequency spectra. Populations from the entire range of *E. paniculata* were sampled, with outcrossing populations from N.E. Brazil and selfing populations from the Caribbean and Central America.
Figure 3.16. Distribution of fitness effects (DFE) of new nonsynonymous mutations for outcrossing and selfing populations of *Eichhornia paniculata*. \( N_{es} \) is the product of \( N_e \) and the selection coefficient (\( s \)). Ten outcrossing samples from N.E. Brazil and 10 selfing samples from the Caribbean and Central America were used to generate the DFE. Error bars for each \( N_{es} \) category are 95% confidence intervals from 200 bootstrap replicates generated by resampling over loci. We used a randomization test (see Keightley and Eyre-Walker 2007) to compare outcrossing and selfing populations and to assess significance at 0.5% level (indicated by *).
**Supplementary tables**

Table 3S1. The sample of *Eichhornia paniculata* individuals used for estimating the efficacy of selection in outcrossing and selfing populations

<table>
<thead>
<tr>
<th>Code</th>
<th>Country</th>
<th>Province/Department</th>
<th>Nearest City</th>
<th>Morph structure</th>
<th>Morph&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>B179</td>
<td>Brazil</td>
<td>Alagoas</td>
<td>Igaci</td>
<td>trimorphic</td>
<td>S</td>
</tr>
<tr>
<td>B191</td>
<td>Brazil</td>
<td>Pernambuco</td>
<td>Lajedo</td>
<td>trimorphic</td>
<td>M</td>
</tr>
<tr>
<td>B192</td>
<td>Brazil</td>
<td>Pernambuco</td>
<td>Cupira</td>
<td>trimorphic</td>
<td>L</td>
</tr>
<tr>
<td>B194</td>
<td>Brazil</td>
<td>Pernambuco</td>
<td>São Caetano</td>
<td>trimorphic</td>
<td>S</td>
</tr>
<tr>
<td>B197</td>
<td>Brazil</td>
<td>Pernambuco</td>
<td>Camocim de Sao Felix</td>
<td>trimorphic</td>
<td>S</td>
</tr>
<tr>
<td>B199</td>
<td>Brazil</td>
<td>Pernambuco</td>
<td>Caruaru</td>
<td>trimorphic</td>
<td>S</td>
</tr>
<tr>
<td>B202</td>
<td>Brazil</td>
<td>Ceará</td>
<td>Quixada</td>
<td>trimorphic</td>
<td>M</td>
</tr>
<tr>
<td>B207</td>
<td>Brazil</td>
<td>Ceará</td>
<td>Choro</td>
<td>trimorphic</td>
<td>L</td>
</tr>
<tr>
<td>B210</td>
<td>Brazil</td>
<td>Ceará</td>
<td>Canindé</td>
<td>trimorphic</td>
<td>L</td>
</tr>
<tr>
<td>Code</td>
<td>Country</td>
<td>Province/Region</td>
<td>Town/Village</td>
<td>Morph Type</td>
<td>Morph</td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>------------</td>
<td>--------</td>
</tr>
<tr>
<td>B211</td>
<td>Brazil</td>
<td>Ceará</td>
<td>Fortaleza</td>
<td>trimorphic</td>
<td>M</td>
</tr>
<tr>
<td>C1</td>
<td>Cuba</td>
<td>Granma</td>
<td>Yara</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>C3</td>
<td>Cuba</td>
<td>Granma</td>
<td>Cholera</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>C5</td>
<td>Cuba</td>
<td>Las Tunas/ Camagüey</td>
<td>Camalote</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>J29</td>
<td>Jamaica</td>
<td>St. Elizabeth</td>
<td>Fullerswood</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>J30</td>
<td>Jamaica</td>
<td>St. Elizabeth</td>
<td>Cataboo</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>J31</td>
<td>Jamaica</td>
<td>St. Elizabeth</td>
<td>Slipe</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td>J32</td>
<td>Jamaica</td>
<td>Westmoreland</td>
<td>Little</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>London</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J33</td>
<td>Jamaica</td>
<td>Westmoreland</td>
<td>Georges</td>
<td>monomorphic</td>
<td>M’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Plain</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mexico</td>
<td>Mexico</td>
<td>San Mateo del Mar,</td>
<td>monomorphic</td>
<td>L’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nr. Tehuantepec</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nicaragua</td>
<td>Nicaragua</td>
<td>Rivas</td>
<td>monomorphic</td>
<td>L’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rio Las</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lajas</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3S2. Polymorphisms segregating in *Eichhornia paniculata* populations identified in aligned portions of EST-derived nuclear locus EP0314 from Barrett et al. (2009) and Ness et al. (2010) and the best matching contig from this study

<table>
<thead>
<tr>
<th>Site</th>
<th>12</th>
<th>14</th>
<th>177</th>
<th>21</th>
<th>22</th>
<th>262</th>
<th>266</th>
<th>29</th>
<th>634</th>
<th>652</th>
<th>655</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil South</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selfer (<em>t</em>=0)</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>A/C</td>
<td>C/T</td>
<td>C/</td>
</tr>
<tr>
<td>Brazil South</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>outcrosser (<em>t</em>&gt;0.8)</td>
<td>A/</td>
<td>C/</td>
<td>G</td>
<td>C/</td>
<td>C/</td>
<td>C/G/</td>
<td>A/</td>
<td>C/</td>
<td>C/T/</td>
<td>A/C/T/</td>
<td>C/T/</td>
<td>C/</td>
</tr>
<tr>
<td>Brazil North</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>outcrosser (<em>t</em>&gt;0.8)</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C/T</td>
<td>C</td>
<td>C</td>
<td>C/</td>
</tr>
<tr>
<td>Caribbean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(<em>t</em>&lt;0.15)</td>
<td>A</td>
<td>C</td>
<td>A/</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T/C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Central America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(<em>t</em>=0)</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C/?</td>
<td>C/?</td>
<td>C/?</td>
<td>C/</td>
</tr>
<tr>
<td>Brazil sample1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil sample2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

102
<table>
<thead>
<tr>
<th>Brazil sample3&lt;sup&gt;b&lt;/sup&gt;</th>
<th>C  C  G  C  C  G  A  C  T  C  C  C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil sample4&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  C  G  C  C  G  A  C  C  C  C  C</td>
</tr>
<tr>
<td>Brazil sample5&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  C  G  C  C  G  G  G  C  T  C  C</td>
</tr>
<tr>
<td>Brazil sample6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  C  G  C  C  G  A  C  C  C  C  C</td>
</tr>
<tr>
<td>Brazil sample7&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  C  G  T  C  C  G  G  C  C  C  C</td>
</tr>
<tr>
<td>Brazil sample8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  G  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Brazil sample9&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  C  G  C  C  G  A  C  C  C  C  C</td>
</tr>
<tr>
<td>Brazil sample10&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C  T  G  T  C  C  A  C  C  T  C  C</td>
</tr>
<tr>
<td>Caribbean sample1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  T  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Caribbean sample2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  T  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Caribbean sample3&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  T  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Caribbean sample4&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  T  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Caribbean sample5&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A  C  A  C  C  T  G  C  T  C  C  C</td>
</tr>
<tr>
<td>Sample</td>
<td>Sequence</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Caribbean sample5&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A C A C C T G C T C C C</td>
</tr>
<tr>
<td>Caribbean sample6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A C A C C T G C T C C C</td>
</tr>
<tr>
<td>Caribbean sample7&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A C A C C T G C T C C C</td>
</tr>
<tr>
<td>Caribbean sample8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>A C A C C T G C T C C C</td>
</tr>
<tr>
<td>Central America sample1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C C G C C G A C C C C C</td>
</tr>
<tr>
<td>Central America sample2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>C C G C C G A C C C C G</td>
</tr>
</tbody>
</table>

<sup>a</sup>Polymorphisms identified in populations sampled in Barrett et al. (2009) and Ness et al. (2010)

<sup>b</sup>Polymorphisms identified in the 10 outcrossing and 10 selfing populations from this study
Table 3S3. Nonsynonymous and synonymous diversity of simulated outcrossing and selfing populations under fixed dominance levels for all mutations

<table>
<thead>
<tr>
<th>Dominance of mutations (h)</th>
<th>Mating system</th>
<th>$\pi_n$</th>
<th>$\pi_s$</th>
<th>$\pi_n/\pi_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>Outcrosser</td>
<td>1.63E-05</td>
<td>3.51E-05</td>
<td>0.46</td>
</tr>
<tr>
<td>0.2</td>
<td>Selfer</td>
<td>3.82E-06</td>
<td>8.75E-06</td>
<td>0.44</td>
</tr>
<tr>
<td>0.5</td>
<td>Outcrosser</td>
<td>1.36E-05</td>
<td>3.41E-05</td>
<td>0.40</td>
</tr>
<tr>
<td>0.5</td>
<td>Selfer</td>
<td>3.67E-06</td>
<td>8.88E-06</td>
<td>0.41</td>
</tr>
<tr>
<td>0.8</td>
<td>Outcrosser</td>
<td>1.24E-05</td>
<td>3.34E-05</td>
<td>0.37</td>
</tr>
<tr>
<td>0.8</td>
<td>Selfer</td>
<td>3.73E-06</td>
<td>8.73E-06</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 3S4. P-values from a two-sample $t$-test comparing $N_{es}^a$ categories of simulated outcrossing and selfing populations under fixed dominance levels for all mutations

<table>
<thead>
<tr>
<th>Dominance of mutations (h)</th>
<th>&lt;1</th>
<th>1-10</th>
<th>10-100</th>
<th>&gt;100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.257</td>
<td>0.479</td>
<td>0.002**</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>0.5</td>
<td>0.163</td>
<td>0.415</td>
<td>0.609</td>
<td>0.651</td>
</tr>
<tr>
<td>0.8</td>
<td>&lt;0.001***</td>
<td>0.181</td>
<td>0.030*</td>
<td>0.275</td>
</tr>
</tbody>
</table>
\( aN_e^s \) is the product of \( N_e \) and the selection coefficient \( (s) \)

*Significant at the 5% level

**Significant at the 1% level

***Significant at the 0.1% level

Table 3S5. Two-way analysis of variance comparing the \( N_e^s^a \) categories of simulated outcrossing and selfing populations under varying dominance for mutations

<table>
<thead>
<tr>
<th>Factor</th>
<th>Degrees of freedom</th>
<th>(&lt;1)</th>
<th>1-10</th>
<th>10-100</th>
<th>&gt;100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mating system</td>
<td>1</td>
<td>0.049*</td>
<td>0.012*</td>
<td>&lt;0.001***</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Time ((N))</td>
<td>5</td>
<td>0.011*</td>
<td>0.089</td>
<td>0.243</td>
<td>0.121</td>
</tr>
<tr>
<td>Mating system * Time ((N))</td>
<td>5</td>
<td>0.011*</td>
<td>0.056</td>
<td>0.092</td>
<td>0.028*</td>
</tr>
<tr>
<td>Error</td>
<td>1428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( aN_e^s \) is the product of \( N_e \) and the selection coefficient \( (s) \)

*Significant at the 5% level

***Significant at the 0.1% level
Table 3S6. Counts of nonsynonymous and synonymous invariant sites and minor allele frequency spectra bins summed across loci for eight outcrossing and eight selfing populations of *Eichhornia paniculata*

<table>
<thead>
<tr>
<th></th>
<th>Invariant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcrosser nonsynonymous</td>
<td>5205480</td>
<td>14392</td>
<td>3552</td>
<td>1702</td>
<td>742</td>
</tr>
<tr>
<td>Outcrosser synonymous</td>
<td>1438285</td>
<td>19738</td>
<td>6799</td>
<td>3933</td>
<td>1914</td>
</tr>
<tr>
<td>Selfer nonsynonymous</td>
<td>5177565</td>
<td>2478</td>
<td>561</td>
<td>402</td>
<td>172</td>
</tr>
<tr>
<td>Selfer synonymous</td>
<td>1454240</td>
<td>3562</td>
<td>918</td>
<td>910</td>
<td>340</td>
</tr>
</tbody>
</table>

Table 3S7. Number of loci containing sites heterozygous across multiple selfing populations from the Caribbean

<table>
<thead>
<tr>
<th></th>
<th>Number of loci containing such sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterozygous in two or more populations</td>
<td>245</td>
</tr>
<tr>
<td>Heterozygous in three or more populations</td>
<td>140</td>
</tr>
<tr>
<td>Heterozygous in four or more</td>
<td>113</td>
</tr>
<tr>
<td>Comparisons</td>
<td>$P$-values</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Original, Figure 6</td>
<td>0.190</td>
</tr>
<tr>
<td>Including sites with QUAL&gt;30, Figure 3S10</td>
<td>0.190</td>
</tr>
<tr>
<td>Including sites with QUAL&gt;0, Figure 3S10</td>
<td>0.267</td>
</tr>
<tr>
<td>Excluding 245 loci with sites heterozygous in &gt;1 selfers, Figure 3S11</td>
<td>0.370</td>
</tr>
<tr>
<td>Excluding 73 loci with sites violating Hardy-Weinberg in outcrossers, Figure 3S12</td>
<td>0.250</td>
</tr>
<tr>
<td>Excluding 135 loci with $\pi_w/\pi_s$&gt;1 in outcrossers or selfers, Figure 3S14</td>
<td>0.533</td>
</tr>
</tbody>
</table>

$^a$ $N_e s$ is the product of $N_e$ and the selection coefficient ($s$)
Chapter 4

Recent mating-system evolution in *Eichhornia* is accompanied by *cis*-regulatory divergence

Summary

The evolution of predominant self-fertilization from cross-fertilization in plants is accompanied by diverse changes to morphology, ecology and genetics, some of which likely result from regulatory changes in gene expression. We examined changes in gene expression during early stages in the transition to selfing in populations of animal-pollinated *Eichhornia paniculata* with contrasting mating patterns. We crossed plants from outcrossing and selfing populations and tested for the presence of allele-specific expression (ASE) in floral buds and leaf tissue of F\textsubscript{1} offspring, indicative of *cis*-regulatory changes. We identified 1,365 genes exhibiting ASE in floral buds and leaf tissue. These genes preferentially expressed alleles from outcrossing parents. Moreover, we found evidence that genes exhibiting ASE had a greater nonsynonymous diversity compared to synonymous diversity in the selfing parents. Our results suggest that the transition from outcrossing to high rates of self-fertilization may have the potential to shape the *cis*-regulatory genomic landscape of angiosperm species, but that the changes in ASE may be moderate, particularly during the early stages of this transition.

Introduction

The evolution of self-fertilization (selfing) from cross-fertilization (outcrossing) is the most frequent reproductive transition in flowering plants. This change in mating system has a major influence on the morphology, ecology, and genetics of selfing lineages. Populations with high
selfing rates generally have smaller and less showy flowers, often lacking physical separation between stigmas and anthers, and usually exhibit reduced pollen to ovule ratios and nectar production (Darwin, 1876; Lloyd 1965; Cruden 1977). Due to the independent origin among angiosperm families of these convergent floral changes they are jointly referred to as the ‘selfing syndrome’ (Ornduff 1969; Stebbins 1974; Sicard and Lenhard 2011). Predominant selfing can also have significant demographic and biogeographical consequences as this mating strategy allows individuals to reproduce under pollinator- and mate-limited conditions, and to found colonies following dispersal events (Lloyd 1980). Theoretical studies also indicate that changes in mating system from outcrossing to selfing can have important genomic consequences (reviewed in Charlesworth and Wright 2001; Wright et al. 2008; Barrett et al. 2014). The reduction in effective population size following the transition to selfing is associated with the accumulation of weakly deleterious mutations, higher mutation and recombination rates, and altered representation of selfish genetic elements in selfing genomes, when compared to outcrossing genomes. Genetic analyses of populations have detected both minor and major effect mutations governing morphological traits promoting self-pollination (e.g. Turnera - Shore and Barrett 1985, 1990; Mimulus - Fishman et al. 2002, 2015; Capsella - Sicard et al. 2011; Slotte et al. 2012). However, the underlying molecular genetic changes involved in transitions from outcrossing to selfing remain largely uncharacterized.

The evolution of the selfing may often be associated with changes in gene expression, but little is known about this potential association. Gene expression is controlled by the interaction between linked elements proximal to the gene (cis-regulatory elements) and unlinked elements that occur elsewhere in the genome (trans-acting elements) (Carroll 2005). Cis-regulatory elements are of particular interest as mutations in these regions have a lower chance of resulting
in deleterious pleiotropic effects compared to those in coding regions, or at trans-acting elements (Emerson and Li 2010; Wittkopp and Kalay 2012). A common approach for isolating the effects of cis-regulatory divergence is to generate hybrids between related species and test for allele-specific expression (ASE) (Pastinen 2010). For example, using RNA-seq data, Bell et al. (2013) detected ASE in 51% of genes (~35,000) in hybrids generated from invasive and noninvasive populations of Cirsium arvense. In a study of hybrids between wild teosinte and domesticated maize lines, Lemmon et al. (2014) reported that 70% of genes (~17,000) exhibited ASE; however, only 1,079 showed consistent differences among all lines indicating standing genetic variation in both the wild and domesticated samples. Recently, significant genetic variation associated with ASE was detected at ~6,000 genes in a single population of the outcrossing Capsella grandiflora (Josephs et al. 2015). These studies detected widespread ASE in populations of the same species, or between closely related taxa.

Shifts in gene expression may accompany mating-system evolution for several reasons. First, adaptive phenotypic changes accompanying the transition in mating system may be mediated by shifts in expression of genes involved in modifications to floral morphology and/or life history (Sicard and Lenhard 2011). Second, species and population differences in the genomic distribution and/or regulation of transposable elements (TEs) can drive shifts in gene expression of adjacent genes through the effects of TE silencing on nearby genes (Lister et al. 2008; Hollister et al. 2011). Third, genome-wide reductions in the efficacy of selection on regulatory sequences in selfing populations (Charlesworth and Wright 2001), and/or relaxed selective pressures due to loss of pollinators, and life history evolution (Lloyd 1980) could each lead to changes in gene expression. Finally, recent theory also predicts that a proliferation of DNA regions that enhance gene expression can accumulate in outcrossing genomes, but
inbreeding and selfing have the potential to reduce the rate of accumulation of such elements (Fyon et al. 2015).

The influence of mating system evolution on gene expression is largely unexplored in plants. Using SNP arrays, He et al. (2012) found that approximately 2,205 out of the 14,462 genes that were assayed showed evidence for ASE in hybrids generated from outcrossing *Arabidopsis lyrata* and selfing *A. thaliana*. Strikingly, the outcrossing *A. lyrata* allele was preferentially expressed over the selfing allele for 90% of the ASE genes, and these genes accumulated more nonsynonymous mutations than synonymous mutations when compared to non-ASE genes. He et al. (2012) attributed the observed outcrossing bias in allelic expression to transposable element-induced gene silencing in *A. thaliana*. Generally, increased expression intensity is associated with slower protein evolution due to stronger selection against protein misfolding (Drummond and Wilke 2008), protein-protein misinteractions (Yang et al. 2012), and mRNA misfolding (Park et al. 2013). Therefore, gene silencing could be associated with the accumulation of nonsynonymous mutations in the down-regulated ASE genes in *A. thaliana* (He et al. 2012). As the proportion of genes displaying ASE in floral bud and leaf tissues in *A. thaliana* were similar, there was no direct evidence indicating that cis-regulatory divergence is specifically associated with adaptive differences in floral morphology or function. Steige et al. (2015) reported that 44% of 18,452 genes showed evidence for ASE in hybrids generated from more closely related outcrossing and selfing *Capsella* species. In contrast to the *Arabidopsis* results, they observed preferential expression of selfing alleles for ASE genes. Analysis of small RNA expression indicated that the observed bias may have resulted from transposable element-induced silencing of outcrossing alleles due to their greater representation near ASE genes. Although they did not observe an enrichment of ASE genes in floral tissues of *Capsella*, a
number of genes showing ASE patterns were found in previously identified regions containing quantitative trait loci (QTL) for floral and reproductive traits, consistent with the hypothesis that some of the genes exhibiting ASE may have been subject to adaptive regulatory evolution in association with the evolution of the selfing syndrome. Currently, the strength and extent of the association of cis-regulation with mating-system variation has been assessed in a single plant family only, and the nature of the association during the early stages of mating-system divergence remains unknown.

Here, we investigate if divergence in cis-regulatory elements has accompanied the recent evolution of selfing from outcrossing in the diploid, annual Eichhornia paniculata (Pontederiaceae). Outcrossing in E. paniculata is promoted by the floral polymorphism tristyly, in which populations are composed of three floral morphs with a reciprocal arrangement of stigma and anther heights (Darwin, 1877; Barrett 1992). In contrast, selfing populations exhibit stigmas and anthers at the same position within a flower and self-pollinate autonomously. The change in spatial separation of stigmas and anthers in selfing variants is governed by recessive modifiers that cause elongation of stamens to the same height as the stigma (Fenster and Barrett 1994). Whereas tristylos populations possess large, showy, blue-purple flowers, those in selfing populations are much smaller and possess traits characteristic of the selfing syndrome (Morgan and Barrett 1989). In E. paniculata the transition to selfing has occurred on multiple occasions and is associated with long-distance dispersal from Brazil to the Caribbean and Central America (Husband and Barrett 1993; Barrett et al. 2009). Ness et al. (2010) estimated that the colonization of the Caribbean occurred ~120,000 years ago, suggesting the transition to selfing occurred relatively recently. Bottlenecks and the shift to selfing are associated with a reduction in genetic diversity in populations (Barrett and Husband 1990; Ness et al. 2010). Recent studies report
reduced codon usage bias and a small increase in the proportion of potentially deleterious and effectively neutral mutations in selfing populations, findings consistent with a genome-wide reduction in selection efficacy (Ness et al. 2012; Arunkumar et al. 2015). *Eichhornia paniculata* therefore provides a valuable system to explore the associations among selection efficacy, *cis*-regulatory divergence and mating-system transitions.

We tested for the occurrence of ASE in F_1 offspring from four crosses between different outcrossing genotypes from Brazil and independently derived selfing genotypes from the Caribbean and Central America. We investigated ASE in two contrasting plant tissues: floral bud tissue with gametophytically- and sporophytically-expressed genes, and leaf tissue that only expresses sporophytic genes, and analyzed the biological function and patterns of selection acting on ASE genes. We addressed the following specific questions: (1) What proportion of genes sampled from F_1 plants exhibit ASE, and is the observed ASE specifically associated with modifications to floral architecture associated with the evolution of selfing? If the evolution of the selfing syndrome in flowers was the major factor influencing *cis*-regulatory variation, we would expect to see a strong signal of ASE in floral tissue but only minimal levels in leaf tissue, and genes showing ASE to be involved in outcrossing floral function. However, a variety of other factors unrelated to floral evolution could also influence patterns of *cis*-regulation, including changes in vegetative and life-history traits associated with the demographic history and ecology of selfing populations, genome-wide changes in TE distributions, or alterations in the efficacy of selection. If these factors play an important role in influencing ASE expression we might expect to see minimal differences in levels of ASE between floral and leaf tissue. (2) Do genes displaying ASE preferentially express outcrossing or selfing alleles, and are there any differences in the proportion of nonsynonymous and synonymous mutations between ASE and
non-ASE genes in outcrossing and selfing populations? If transitions to selfing are accompanied
by widespread gene silencing, we might expect to see higher expression of outcrossing alleles
and an accumulation of nonsynonymous mutations in selfing compared to outcrossing
populations. Conversely, outcrossing alleles might be under-expressed if rare mutations resulting
in gene silencing in outcrossers were absent in selfers, due to stochastic loss or the reduced
activity of transposable elements during the transition to selfing (see Arvid et al. 2014; Steige et
al. 2015).

**Materials and methods**

*Sampling parents and generating F₁ crosses*

We obtained open-pollinated seeds from four outcrossing populations from N.E. Brazil and four
selfing populations from Jamaica (n=2), Cuba and Mexico. We used the morph structure of
populations as a determinant of its mating system, as trimorphic and monomorphic populations
are predominantly outcrossing and selfing, respectively (Barrett and Husband 1990). We grew
plants under uniform glasshouse conditions at the University of Toronto and chose one
individual from each population and crossed each outcrosser to a unique selfer (Appendix III -
Table 4S1). For each cross, the outcrosser and selfer were the paternal and maternal parents,
respectively. We emasculated maternal parents four hours prior to anther dehiscence and applied
pollen from the outcrossing paternal parent using forceps. Approximately two weeks after each
cross, we collected mature seeds and stored them in dry conditions. Following germination of F₁
crosses eight months later, we chose one plant from each of the four crosses for further study.
RNA extraction and sequencing

We extracted RNA from floral buds and leaf tissue. We note that what we refer to as ‘leaf tissue’ is technically tissue obtained from elongated internodes or petioles subtending ‘leaves’ [see Richard and Barrett (1984) for details of the organography of *E. paniculata*]. The reproductive and vegetative tissue was obtained from each of the eight parents and four F₁ offspring for sequencing. We extracted RNA using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich). We used the extracted RNA samples to make Illumina TruSeq RNA libraries that were sequenced using the 100 bp paired end protocol on three lanes of the Illumina HiSeq 2000 at the McGill University and Génome Québec Innovation Centre. The parental bud transcriptomes were sequenced as part of Arunkumar et al. (2015) on lanes where 12 samples were multiplexed in each lane. The raw sequence data for the parental bud transcriptomes are available under accession number SRP049636 at the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra, accessed August 24 2015) and the associated BioProject alias is PRJNA266681 (http://www.ncbi.nlm.nih.gov/bioproject/, accessed August 24 2015). For this study, eight libraries were prepared from the leaf tissue of parents, four were leaf tissue of F₁ progeny and four were floral buds of F₁ progeny. The 12 libraries from leaf tissue were multiplexed on each lane. We sequenced each of these libraries twice, with a total of two lanes dedicated to these samples. However, we only used one of the two sequenced libraries for parental leaf tissue in our analyses to keep the sequencing depth among the floral bud and leaf parental transcriptomes the same. In contrast, we used both of the sequenced libraries prepared from the leaf tissue of the F₁ progeny in our analyses. The four libraries prepared from floral buds of F₁ progeny were sequenced on an individual lane. The raw sequence data for the parental leaf transcriptomes and the F₁ floral bud and leaf transcriptomes are available under accession
number SRP060405 at the Sequence Read Archive and the associated BioProject alias is PRJNA288861.

Read mapping

We previously generated a 65.53 Mbp de novo transcriptome assembly \((N_{50} = 2.2\ \text{Kbp})\) from selfing transcriptomes from six populations in the Caribbean using the programs Velvet 1.2.08 (Zerbino and Birney 2008) and Oases 0.2.08 (Schulz et al. 2012) and predicted coding regions by BLAST searches to plant databases (Arunkumar et al. 2015). Further, we filtered paralogous loci by removing contigs containing sites heterozygous across multiple selfing populations, and those for which coverage of mapped genomic reads were greater than the genomic sequencing depth. We mapped the RNA-seq short reads and genomic reads with Burrows-Wheeler Aligner v0.7.8-r455 (Li and Durbin 2009) using default parameters and used the Sampe command to combine the paired end read mapping results. Then, we used the Stampy 1.0.23 software (Lunter and Goodson 2011) with default parameters to map more divergent reads and identify insertions and deletions (indels). We processed the read mapping output using the SamFormatConverter, ReorderSam, AddOrReplaceReadGroups and BuildBamIndex programs from of the Picard tools package 1.124 using default settings (http://picard.sourceforge.net, accessed August 24 2015). We assigned the libraries prepared from leaf tissue of the F\(_1\) progeny that were sequenced on separate lanes to the same read group. We also used the MarkDuplicates program, part of the Picard tools package 1.124, to identify and tag duplicate reads. As erroneous mismatches might have occurred near indels, we used the RealignerTargetCreator and IndelRealigner programs, part of the Genome Analysis Toolkit (GATK) 3.3-0-g37228af (DePristo et al. 2011), with default parameters to realign sequences within \(~3\ \text{kb}\) of an indel. We subsampled the reads from the resulting BAM files of F\(_1\) floral bud transcriptomes to 67% of their original value using Samtools.
view command (Samtools v1.2; Li et al. 2009) to control for the higher sequencing depth of these libraries compared to those prepared from the leaf tissue of the F1 progeny.

**Identifying genes showing ASE**

We ran the UnifiedGenotyper program, also part of GATK 3.3-0-g37228af (DePristo et al. 2011), with the BadCigar read filter to identify invariant sites, SNPs and indels from all samples and we retained sites with Phred scaled quality scores > 60 and sequencing depth > 20. We retained SNPs if their Phred scaled genotype quality for all parental and progeny samples from the floral bud and leaf tissue was > 60 and if they were at least 5 bp away from either side of an indel. We used custom Perl scripts to identify sites that were homozygous for alternative SNPs in both parents and heterozygous in F1 plants. For these sites, we extracted the depth of the outcrossing and selfing alleles in the F1 transcriptomes from the UnifiedGenotyper output. We used the depth information to estimate the posterior probability that a gene shows ASE using hierarchical Bayesian inference developed by Skelly et al. (2011). We compared the level of ASE detected using this Bayesian inference to that detected using a binomial exact test, a widely applied method for estimating the level of differential expression (Robinson et al. 2010; Chen et al. 2011; Skelly et al. 2011). While the binomial test has a higher false discovery rate (FDR) when compared to Bayesian methods (Skelly et al. 2011; León-Novelo et al. 2014), it is a powerful approach when low quality sites are excluded (Castel et al. 2015). As we lacked genomic reads from the parents, we adopted the genomic model inferred for humans and yeast by Skelly et al. (2011). In their paper, they summarized the genomic model using the $a$ hat and $d$ hat parameters. For both species, the $a$ hat ranged from 3,000 to 6,500 indicating that both parental copies were expressed at approximately equal levels in the genomic data and $d$ hat of ~550 indicating that noise around this 50:50 expectation for each gene was small. Following
their model, we used large values for both parameters \((a \text{ hat} = 5,000, \ d \text{ hat} = 500)\). Even so, we investigated the effect of increasing or decreasing noise in the genomic model by repeating the analysis using 10 fold larger and smaller values of \(a \text{ hat} \) and \(d \text{ hat} \) parameters. We ran 500,000 MCMC iterations, sampling every 100\(^{th}\) iteration to check for convergence. We ran 2,000 scaling iterations allowing for a maximum of eight rounds of scaling. After convergence, we calculated the posterior probability that each gene showed ASE. We also investigated the presence of ASE by performing a two-sided negative binomial exact test on each gene. To do this, we averaged the reads counts across all SNPs for each gene using edgeR v3.1 (Robinson et al. 2010). We only used genes that had at least one site heterozygous in all progeny. We used a FDR of 5\% with a null hypothesis that paternal and maternal alleles were expressed equally to call ASE genes using this approach.

*Filtering ASE calls*

While the lack of availability of parental genomic sequence required us to explore a range of parameters in the Bayesian analysis, availability of transcriptome data from one outcrossing genotype and genomic data from a cross between selfing populations allowed us to identify and filter some potential SNPs showing spurious ASE, or ASE reflecting standing variation within outcrossing populations. We inferred ASE in the outcrossing parents by estimating allelic expression for sites heterozygous in each outcrosser. We excluded from our analyses genes that had posterior probabilities of ASE > 0.7 in transcriptomes generated for outcrossing parents, as such genes likely reflect ASE present in natural populations unrelated to differences between the outcrossing and selfing genotypes. To exclude spurious ASE sites caused by possible mapping biases, we also inferred the extent of ASE when genomic reads from two *E. paniculata* individuals were mapped to the transcriptome reference. One individual was a selfer from
Oaxaca, Mexico and the other was a genotype derived from a cross between a selfing individual from Slipe, Jamaica and the aforementioned selfer from Mexico. We calculated allelic depth for sites heterozygous in each individual for identifying ASE and used the DNA model scripts generated by Skelly et al. (2011) to estimate the $a$ hat and $b$ hat parameters. We excluded from our analyses of the genomic read data genes that had posterior probabilities of ASE > 0.7, as genomes would be uninformative for assessing differential expression, and any observed ASE may be spurious and associated with technical variability in read mapping.

*Molecular functional analyses*

For genes displaying evidence for ASE, we performed custom translated nucleotide BLAST searches (Altschul et al. 1990), with *E. paniculata* genes as the query against the *Arabidopsis* protein database obtained from The Arabidopsis Information Resource (TAIR) (Lamesch et al. 2012). We performed a singular enrichment analysis, using the genes expressed in each tissue as the query and the TAIR10 *Arabidopsis* gene model as the reference, using the agriGO toolkit and database (Du et al. 2010). We found significant gene ontology terms were found after performing hypergeometric tests with the Yekutieli multiple test correction at a significance level of 0.05.

*Calculating expression intensity*

We calculated the number of reads that mapped to the transcriptome reference assembly from the floral bud and leaf transcriptomes of the parents and F$_1$ offspring using HTSeq-count 0.6.1p1 (Anders et al. 2014) under the union mode. For genes with significant evidence for ASE, we averaged the allelic depth across all sites to calculate the mean expression intensity for the outcrossing and selfing alleles in the F$_1$ progeny. Further, we estimated the expression levels for these ASE genes in the parental floral bud and leaf transcriptomes by using HTSeq-count 0.6.1p1. We assessed if comparisons of expression levels between outcrossing and selfing alleles
in the F1 progeny, and expression levels of ASE genes in the outcrossing and selfing parental transcriptomes, were statistically significant using R (R Development Core Team 2011) by first generating bootstrap replicates, after resampling randomly across the genes. We performed a two-tailed permutation test comparing the overall sums for bootstrap datasets to generate a P-value for each comparison.

**Comparing nonsynonymous and synonymous diversity**

We performed a 2x2 contingency test to investigate if there was interaction between nonsynonymous and synonymous polymorphisms and floral bud and leaf ASE genes. For this test, we summed the number of polymorphisms observed across all ASE genes in each tissue. Further, we compared genetic diversity ($\pi$) for nonsynonymous ($\pi_{\text{nonsynonymous}}$) and synonymous sites ($\pi_{\text{synonymous}}$) for genes with ASE based on posterior probabilities > 0.7 against genes without evidence for ASE. All genes with ASE would have had heterozygous site(s). To reduce potential directional biases due to invariant genes from the set that did not show ASE, we only analyzed genes with one or more heterozygous sites. We excluded the genotype from Mexico when calculating $\pi$ for selfing populations. This genotype was a result of an independent transition to selfing (see Arunkumar et al. 2015), and therefore substitutions between genotypes from separate transitions may have been erroneously called as SNPs and thus distorted estimates of $\pi$. Our ASE gene set only included those with posterior probability for ASE > 0.7 across all of the remaining three progeny. All other genes were defined as non-ASE genes. We randomly sampled three outcrossing genotypes from the four sequenced parents to keep the number of chromosomes being compared the same. We estimated $\pi$ using the Polymorphorama script (Andolfatto 2007; Haddrill et al. 2008) and generated mean values and 95% confidence intervals (CI) across 1000 bootstrap replicates, after resampling randomly across the genes using R (R Development Core
To obtain the 95% CI, we excluded bootstrap replicates with mean values below the 2.5 percentile and above the 97.5 percentile. Then, we used the smallest and largest values from the rest to represent the lower and upper limits of the CIs, respectively. We performed permutation tests, as in the previous section, to assess if $\pi_{\text{nonsynonymous}}/\pi_{\text{synonymous}}$ comparisons between ASE and non-ASE genes were statistically significant. This resampling approach would make it less likely that a few loci with extremely high or low levels of genetic diversity would lead to the erroneous detection of significant differences in the comparisons of ASE and non-ASE genes.

**Results**

*Allele-specific expression (ASE)*

The distribution of coverage of genes from floral bud and leaf transcriptomes from parents and offspring were similar, but there were many genes that were not expressed in leaf transcriptomes (Appendix III - Figure 4S1). We identified 60,000-70,000 sites homozygous for alternative SNPs in outcrossing and selfing parents and heterozygous in F$_1$ progeny for testing for the presence of ASE in floral bud and leaf tissues (Appendix III - Table 4S2). Of these sites, only 3,118 and 2,836 sites were heterozygous in floral bud and leaf tissues, respectively, across all four F$_1$ progeny, and 1,798 sites were heterozygous in both tissue types in all four progeny. 450-800 genes sampled from the floral bud tissue had probability values > 0.7 (Figure 4.1a), but only 53 genes showed ASE across all four F$_1$ genotypes. Similarly, 350-650 genes sampled from leaf tissue had probability values > 0.7 (Figure 4.1b), but only 46 genes showed ASE across all four F$_1$ genotypes. Only five genes showed ASE across both floral bud and leaf tissues across all four progeny at this posterior probability threshold. Note that a 2x2 contingency test did not reveal an
interaction between tissue type, i.e. floral bud or leaf, and the number of nonsynonymous and synonymous polymorphisms summed across all ASE genes for each tissue at the 5% significance threshold. Genes displaying ASE in floral buds and leaves were involved in a range of functions from growth, binding, regulation, maintaining cell structure and defense response (Appendix III - Table 4S3), but there was no clear association to floral function for any of the genes.

Figure 4.1. Distribution of posterior probabilities for genes showing allele-specific expression (ASE) in *Eichhornia paniculata*. Probability distributions are illustrated for genes sampled from (a) floral buds and (b) leaf tissue of F₁ plants generated from crosses between selfing (maternal)
and outcrossing (paternal) parents. We used a Bayesian binomial test as implemented in Skelly et al. (2011) to determine the probability of ASE against the null expectation of equal expression intensity of both alleles.

We tested the effects of varying the parameters and approaches for detecting ASE. First, we varied the $a_{\hat{}}$ and $d_{\hat{}}$ parameters, which represent parameters quantifying technical variability, using the Skelly et al. (2011) approach. When we used larger $a_{\hat{}}$ (50000) and $d_{\hat{}}$ (5000) parameters to identify genes showing ASE in floral buds of one of the F1s, the mode of the posterior probability distribution shifted to the right (Appendix III - Figure 4S2). In this case, many more genes had larger posterior probabilities for ASE. In contrast, the use of smaller $a_{\hat{}}$ (500) and $d_{\hat{}}$ (50) parameters (allowing for greater overdispersion due to technical variability) resulted in a distribution similar to the original, using an $a_{\hat{}}$ of 5,000 and $d_{\hat{}}$ of 500 (Figure 4.1a), with similar number of genes with probabilities $> 0.7$. Further, a negative binomial exact test indicated that of the 6360 genes that had at least one site heterozygous across the floral buds of all four F1 progeny, 764 showed significance evidence for ASE at a 5% FDR threshold.

Expression fold difference

We calculated differences in expression intensity of outcrossing and selfing alleles in F1 plants for genes with posterior probability of ASE $> 0.7$. We chose this probability threshold, as there were few ASE genes with larger posterior probability values (Figure 4.1). Strikingly, almost all genes with ASE in floral bud (Figure 4.2a) and leaf tissues (Figure 4.2b) had 1.5-2 fold increase in expression of outcrossing compared to selfing alleles, across all comparisons. Permutation tests indicated the differences in expression intensity of outcrossing and selfing alleles were significant at the 0.1% threshold for all comparisons. This result does not appear to be due to
mapping biases; although we observed a general outcrossing bias in allelic expression, it was often the selfing allele that matched the reference allele in the three progeny with Caribbean selfing maternal parents, because of the source of our reference transcriptome (Appendix III - Table 4S2). For the progeny from the selfing parent from Mexico, the outcrossing allele matched the reference allele for ~40% of sites and the selfing allele matched the reference allele for the remaining 60%. For this progeny, we consistently observed an outcrossing bias in allelic expression intensities for ASE genes, when repeating the calculations using only sites where the outcrossing allele (Appendix III - Figure 4S3a), or selfing allele (Appendix III - Figure 4S3b), matched the reference allele. In contrast, for genes without significant evidence for ASE, both alleles were expressed at approximately equal levels regardless of which site types were analyzed (Appendix III - Figure 4S3a-c). Genes showing ASE in floral bud and leaf tissues of the F₁ plants also had a general trend of increased expression levels in outcrossing compared to selfing parents (Appendix III - Figure 4S4 a,b), although this difference was not significant at the 5% level across all comparisons (Appendix III - Table 4S4).

Figure 4.2. Distribution of logₑ-fold expression difference between outcrossing and selfing alleles of genes with posterior probability for allele-specific expression > 0.7 in F₁ plants of
*Eichhornia paniculata*. Illustrated are the distributions for genes sampled from (a) floral buds and (b) leaf tissue of F1 plants. Large (> 0) and small (< 0) values indicate higher expression of outcrossing and selfing alleles, respectively.

**Relative proportions of nonsynonymous and synonymous mutations**

Figure 4.3. Diversity at nonsynonymous ($\pi_{\text{nonsynonymous}}$) and synonymous sites ($\pi_{\text{synonymous}}$) for genes with posterior probability > 0.7 for allele-specific expression (ASE), and for genes without evidence of ASE (posterior probability < 0.7) in *Eichhornia paniculata*. Illustrated are the mean $\pi_{\text{nonsynonymous}}/\pi_{\text{synonymous}}$ estimates for ASE and non-ASE genes in: (a) floral buds of outcrossing genotypes (b) leaf tissue of outcrossing genotypes (c) floral buds of selfing genotypes and (d) leaf tissue of selfing genotypes. Three outcrossing genotypes from N.E. Brazil and three selfing genotypes from the Caribbean were used to estimate $\pi$. All analyzed genes had one or more heterozygous sites. There were 106 and 86 genes showing ASE in floral buds and leaf tissues, respectively. We randomly subsampled from the set of non-ASE genes to keep the number of
genes compared the same. 95% confidence intervals were generated from 1000 bootstrap replicates, resampling across genes. *** indicates comparison significant at the 0.1% level based on permutation tests.

We compared \( \pi_{nonsynonymous}/\pi_{synonymous} \) for 106 floral bud and 86 leaf tissue genes that had posterior probability > 0.7 for ASE across the three progeny with Caribbean selfing maternal parents against all other genes (posterior probability < 0.7). We randomly subsampled 106 floral bud and 86 leaf tissue genes from the set without significant evidence for ASE to keep the number of genes being compared the same. As expected, there was a ~80% and ~50% reduction in diversity for ASE and non-ASE genes, respectively, in floral bud and leaf tissues of selfing parents when compared to outcrossing parents (Appendix III - Table 4S5). The percent diversity reductions at nonsynonymous and synonymous sites were similar. ASE genes for floral bud and leaf tissue of outcrossing parental genotypes, \( \pi_{nonsynonymous}/\pi_{synonymous} \) was not significantly different from the estimate based on all other genes (Figure 4.3a, b). In contrast, \( \pi_{nonsynonymous}/\pi_{synonymous} \) was significantly higher for ASE genes compared to all other genes in floral buds of parental selfing genotypes (Figure 4.3c). While a similar trend of higher \( \pi_{nonsynonymous}/\pi_{synonymous} \) was also observed for ASE in leaf tissue compared to non-ASE genes, this difference was not significant at the 5% threshold (Figure 4.3d). Note that a non-normal distribution of \( \pi_{nonsynonymous} \) and \( \pi_{synonymous} \) values resulted in lower \( \pi_{nonsynonymous}/\pi_{synonymous} \) when we resampled individual estimates for each gene to generate bootstrap replicate datasets (Figure 4.3) compared to a single global estimate (Appendix III - Table 4S5). The former approach is less susceptible to deviations from assumptions of normality. The \( \pi_{nonsynonymous}/\pi_{synonymous} \) estimates remained significantly different between ASE and non-ASE genes when we repeated the analyses, after restricting the ASE gene set to those...
with posterior probabilities for ASE > 0.75 across the three progeny. In contrast, when the posterior probability threshold for ASE genes was reduced beyond 0.65, there were no longer significant differences in $\pi_{\text{nonsynonymous}}/\pi_{\text{synonymous}}$ between ASE and non-ASE genes.

**Discussion**

We detected allele-specific expression for genes sampled from floral bud and leaf tissue of F$_1$ plants from crosses between outcrossing and selfing genotypes of the annual plant *E. paniculata*. However, only moderate levels of *cis*-regulatory differentiation accompanied the transition to selfing, probably because our study involved an intraspecific comparison of mating system differentiation in which selfing is of relatively recent origin, and we factored out genes with evidence of ASE segregating within outcrossing populations. The majority of genes we investigated showed biased expression of the outcrossing allele in both floral bud and leaf tissue. When compared against genes without evidence for ASE in F$_1$ plants, genes with ASE had more nonsynonymous mutations than synonymous mutations in selfing parents but not outcrossing parents. Below, we discuss the possible reasons for the associations between *cis*-regulatory variation, nonsynonymous polymorphism, and mating system divergence in *E. paniculata*.

*The evolution of selfing is accompanied by differences in cis-regulation*

Our results are consistent with the hypothesis that there is an association between mating system divergence and patterns of ASE. We observed preferential expression of the outcrossing allele in genes showing ASE in floral and leaf tissues, similar to patterns reported from *Arabidopsis* (Chang et al. 2010; He et al. 2012), and in contrast to those from *Capsella* (Steige et al. 2015). We would have expected preferential mapping of selfing alleles if read mapping bias
significantly influenced the detection of ASE (Degner et al. 2009; Stevenson et al. 2013) because the transcriptome reference we used was from Caribbean selfing populations (Arunkumar et al. 2015). However, this was not the case as we observed a strong bias in the expression of outcrossing alleles. Genes showing ASE in progeny also showed generally greater expression intensities in outcrossing compared to selfing parents. The difference in level of gene expression between the two parents was marginal in some of the parental pairs across both tissue types. This difference may be associated with the reduced power, owing to the lower sequencing depth of the parents when compared to the F1 offspring. Alternatively, it is possible in some cases that compensatory trans effects are present, whereby cis-regulatory changes revealed in F1 hybrids are compensated for by trans effects in the parental populations (Bell et al. 2013).

Given that we observed as much ASE in leaves as flower buds, there is no clear evidence for specific regulatory changes associated with the evolution of the selfing syndrome in E. paniculata. Patterns of ASE in this species may also be associated with adaptation to new biogeographical zones with potentially novel environments, as a result of the colonization of the Caribbean and Central America following long-distance dispersal from Brazil (Ness et al. 2010). However, it is not clear why such changes would consistently be associated with reduced expression in selfing populations. Instead, the overall patterns we observed may be more likely driven by general shifts in gene regulation caused by enhancer evolution (Fyon et al. 2015), TE-mediated gene silencing (Lister et al. 2008; Hollister et al. 2011), and/or changes in the strength or efficacy of selection on ASE genes following the transition to selfing (Charlesworth and Wright 2001). Elevated nonsynonymous polymorphism at down-regulated genes is consistent with a reduced strength and/or efficacy of selection, since it implies that this set of genes in particular has experienced less constraint. The finding of ASE in both floral and leaf tissue
suggests that a variety of evolutionary forces have probably shaped cis-regulatory variation accompanying the change in mating system in *E. paniculata*.

Another factor that can contribute to allele-specific expression is genomic imprinting. In flowering plants imprinting has been found to primarily affect endosperm (Messing and Grossniklaus 1999; Köhler and Weinhofer-Molisch 2010; Nodine and Bartel 2012), although more recent evidence from *A. thaliana* suggests that imprinting may also affect the contribution of alleles from the maternal and paternal parents to the embryo (Raissig et al. 2013). Brandvain and Haig (2005) predicted that in crosses between an outcrossing paternal parent and a selfing maternal parent, maternally inherited genes might exhibit imprinting. Under this scenario, the outcrossing allele might be preferentially expressed in progeny, similar to the allelic bias we observed. However, we collected floral buds from F1 progeny prior to any possibility of autonomous selfing, and thus it seems unlikely that the ASE we observed resulted from early endosperm development. Although our experimental design minimized the influence of genetic imprinting, it cannot be completely ruled out as a possible contributor to the observed patterns of ASE. A recent study suggested that parent-of-origin effects on gene expression may exist in later stages of development (Videvall et al. 2015). It is also worth noting that in studies of allele-specific expression involving mating system differences (*Capsella* - Steige et al. 2015, *Arabidopsis* - He et al. 2015, this study) the paternal allele showed ASE bias in all cases. Reciprocal crosses between outcrossing and selfing genotypes might provide further insights on the potential contribution of genetic imprinting and/or inheritance patterns of small RNAs to ASE.
Patterns of ASE and association to the evolution of the selfing syndrome

The observed *cis*-regulatory differences in selfers and outcrossers could be due to a variety of causes including changes to floral function and life history traits associated with the evolution of selfing, genomic changes such as enhancer evolution (Fyon et al. 2015) or genome-wide reductions in the efficacy of selection (Charlesworth and Wright 2001). As there was no significant interaction between tissue type and ASE shaping nonsynonymous and synonymous diversity, or evidence for an enrichment of genes involved in floral or reproductive function, selection on floral function is clearly not the only factor associated with the down-regulation of selfing alleles in F1 plants of *E. paniculata*.

The evolution of the selfing syndrome is unlikely to explain the overall patterns of regulatory evolution that we observed. However, this does not rule out a role for regulatory evolution in morphological changes associated with the shift to selfing. Some of the genes showing ASE in floral tissue may have important roles in traits associated with the selfing syndrome. We detected 100 genes showing ASE in floral buds and not leaves; similarly, 85 genes of the 1,305 genes displaying ASE in floral buds of hybrid *Capsella* were found within genetic regions associated with floral trait differentiation between selfers and outcrossers (Steige et al. 2015). In comparison to the number of ASE genes identified, genetic mapping studies have found only 1-6 QTLs governing changes in 6-10 traits associated with floral morphological differences in *Mimulus* (Fishman et al. 2002, 2015) and *Capsella* (Sicard et al. 2011; Slotte et al. 2012). Although the number of genes showing ASE is higher than the number of QTLs associated floral phenotypic differentiation in these earlier studies, there may be multiple genes occurring within the QTL regions identified. Also, some QTLs underlying floral traits may comprise “master regulators” that co-ordinate the expression of a suite of genes. Our
comparisons of patterns of ASE in floral bud and leaf tissues suggest that overall cis-regulatory
differentiation is more likely to be a general consequence of mating system divergence.
Nevertheless, the floral-specific ASE genes that we have identified comprise an important set of
candidate genes for investigating the role of morphological evolution in mating system
transitions.
Appendix III

Supporting information for Chapter 4

Supplementary figures

Figure 4S1. Number of reads that mapped to reference assembly from (a) floral bud and leaf parental transcriptomes and (b) floral bud and leaf F₁ offspring transcriptomes of *Eichhornia paniculata*. 
Figure 4S2. Distribution of posterior probabilities that genes show allele-specific expression (ASE) under various $\hat{a}$ and $\hat{d}$ parameters in *Eichhornia paniculata*. Illustrated are probability distributions for genes sampled from floral buds of an F1 plant from a cross between selfing (maternal, MEX) and outcrossing (paternal, B192) parent under various $\hat{a}$ and $\hat{d}$ parameters. A Bayesian binomial test as implemented in Skelly et al. (2011) was used to determine the probability of ASE against the null expectation of equal expression intensity of both alleles. Smaller $\hat{a}$ and $\hat{d}$ values increase the chances of departure from the 50:50 expectation for expression intensity of null genes and vice-versa.
Figure 4S3. Distribution of log$_e$-fold expression difference between outcrossing and selfing alleles for ASE and non-ASE genes calculated using (a) only sites where the outcrossing allele matched the reference allele (b) only sites where the selfing allele matched the reference allele or (c) all sites. Illustrated are distributions for genes sampled from floral buds of an F$_1$ plant from a cross between selfing (maternal, MEX) and outcrossing (paternal, B192) parent. ASE genes and non-ASE genes were defined as those with posterior probability values $> 0.7$ and $< 0.7$ respectively. Large ($> 0$) and small ($< 0$) values indicate higher expression of outcrossing and selfing alleles, respectively.
Figure 4S4. Distribution of log$_e$-fold gene expression difference between outcrossing and selfing *Eichhornia paniculata* parents for the set of genes with posterior probability for allele-specific expression $> 0.7$ in F$_1$ offspring. Illustrated are the distributions for genes sampled from (a) floral buds and (b) leaf tissue of the parental transcriptomes. Large ($> 0$) and small ($< 0$) values indicate higher expression of genes in the outcrossing and selfing transcriptomes, respectively.
### Supplementary tables

Table 4S1. Population codes and localities for parental genotypes of *Eichhornia paniculata* used in the study.

<table>
<thead>
<tr>
<th>Cross</th>
<th>Maternal parent population (selfer)</th>
<th>Paternal parent population (outcrosser)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yara, Granma, Cuba (C1)</td>
<td>Igaci, Alagoas, Brazil (B179)</td>
</tr>
<tr>
<td></td>
<td>Slipe, St. Elizabeth, Jamaica (J31)</td>
<td>Caruaru, Pernambuco, Brazil (B199)</td>
</tr>
<tr>
<td>2</td>
<td>Little London, Westmoreland, Jamaica (J32)</td>
<td>Lajedo, Pernambuco, Brazil (B191)</td>
</tr>
<tr>
<td>3</td>
<td>Oaxaca, San Mateo del Mar, Mexico (Mex)</td>
<td>Cupira, Pernambuco, Brazil (B192)</td>
</tr>
<tr>
<td>4</td>
<td>Nr. Tehuantepec, Mexico (Mex)</td>
<td></td>
</tr>
</tbody>
</table>
Table 4S2. Number of sites homozygous for alternative SNPs in outcrossing and selfing parents and heterozygous in F1 progeny.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Status of F1 progeny alleles</th>
<th>C1 x</th>
<th>J31 x</th>
<th>J32 x</th>
<th>MEX x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B179</td>
<td>B199</td>
<td>B191</td>
<td>B192</td>
</tr>
<tr>
<td>Floral bud</td>
<td>Selfing allele matched reference</td>
<td>72996</td>
<td>66686</td>
<td>60294</td>
<td>26821</td>
</tr>
<tr>
<td></td>
<td>Outcrossing allele matched reference</td>
<td>7461</td>
<td>3787</td>
<td>7467</td>
<td>37483</td>
</tr>
<tr>
<td></td>
<td>Neither allele matched reference</td>
<td>190</td>
<td>96</td>
<td>165</td>
<td>597</td>
</tr>
<tr>
<td></td>
<td>Selfing allele matched reference</td>
<td>59730</td>
<td>65056</td>
<td>58316</td>
<td>24812</td>
</tr>
<tr>
<td>Leaf</td>
<td>Outcrossing allele matched reference</td>
<td>6167</td>
<td>4130</td>
<td>7641</td>
<td>34071</td>
</tr>
<tr>
<td></td>
<td>Neither allele matched reference</td>
<td>186</td>
<td>122</td>
<td>147</td>
<td>487</td>
</tr>
</tbody>
</table>
Table 4S3. Functional classification for genes with posterior probability for allele-specific expression > 0.7 sampled from floral buds and leaf tissue of F1 plants from crosses between selfing (maternal) and outcrossing (paternal) parents of *Eichhornia paniculata*.

<table>
<thead>
<tr>
<th>Tissue(s)</th>
<th>ID</th>
<th>Ontology</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>floral bud and leaf</td>
<td>GO:0001558</td>
<td>biological process</td>
<td>regulation of cell growth</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0003700</td>
<td>molecular function</td>
<td>sequence-specific DNA binding transcription factor activity</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0005102</td>
<td>molecular function</td>
<td>receptor binding</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0005520</td>
<td>molecular function</td>
<td>insulin-like growth factor binding</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0005576</td>
<td>cellular component</td>
<td>extracellular region</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0006355</td>
<td>biological process</td>
<td>regulation of transcription, DNA-templated</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0006952</td>
<td>biological process</td>
<td>defense response</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0008270</td>
<td>molecular function</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td>floral bud and leaf</td>
<td>GO:0009405</td>
<td>biological process</td>
<td>pathogenesis</td>
</tr>
<tr>
<td>GO:0016068</td>
<td>type I hypersensitivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0043565</td>
<td>sequence-specific DNA binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0046872</td>
<td>metal ion binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0051537</td>
<td>2 iron, 2 sulfur cluster binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005199</td>
<td>structural constituent of cell wall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005515</td>
<td>protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008061</td>
<td>chitin binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009277</td>
<td>fungal-type cell wall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4S4. *P*-values from comparisons of gene expression level differences between outcrossing and selfing *Eichhornia paniculata* parents for the set of genes showing allele-specific expression in the F1 offspring.

<table>
<thead>
<tr>
<th>Parental comparisons</th>
<th>Floral bud</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 v. B179</td>
<td>0.33</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>J31 v. B199</td>
<td>0.004</td>
<td>0.408</td>
</tr>
<tr>
<td>J32 v. B191</td>
<td>&lt;0.001</td>
<td>0.214</td>
</tr>
<tr>
<td>MEX v. B192</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 4S5 Diversity at nonsynonymous and synonymous sites for *Eichhornia paniculata* floral bud and leaf genes in outcrossing parents from Brazil and selfing parents from the Caribbean.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Floral bud</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>ASE</td>
<td>non-ASE</td>
</tr>
<tr>
<td>Outcrosser nonsynonymous</td>
<td>2.46E-03</td>
<td>1.18E-03</td>
</tr>
<tr>
<td>Outcrosser synonymous</td>
<td>1.42E-02</td>
<td>7.68E-03</td>
</tr>
<tr>
<td>Selfer nonsynonymous</td>
<td>5.75E-04</td>
<td>6.85E-04</td>
</tr>
<tr>
<td>Selfer synonymous</td>
<td>2.63E-03</td>
<td>4.42E-03</td>
</tr>
</tbody>
</table>
Chapter 5

Genetic linkage between the $S$ and $M$ loci governing tristyly in

*Eichhornia paniculata* (Pontederiaceae)

Summary

A two-locus model ($S$ and $M$) is most commonly invoked to explain the inheritance of tristyly, a floral polymorphism that has evolved independently in several angiosperm families. Several deviations from this general model are reported in the literature, including whether the $S$ and $M$ loci are linked or not. Earlier controlled crosses and progeny tests in the neotropical herb *Eichhornia paniculata* (Pontederiaceae) established that two diallelic loci, with $S$ epistatic to $M$, control tristyly. Here, I extend this work by synthesizing double heterozygous short-styled plants to determine whether the two loci are linked, and if so to estimate the map distance between them. Crosses of four $S$- and $L$-morph parents indicated that the $S$ and $M$ loci were linked in one $S$-morph parent and were separated by a map distance of 2.7 cM. Progenies from the remaining $S$-morph parents were uninformative with respect to linkage. The deficit of $S$-morphs in the selfed progeny of the parent showing linkage was consistent with expectations of selection against genotypes homozygous at the $S$ locus, perhaps owing to the sheltering of deleterious alleles. A literature survey indicated that tristyloous species with and without linkage of the $S$ and $M$ loci occur in two of the three families in which the inheritance of the polymorphism has been investigated. The consequences of linkage between the $S$ and $M$ loci on style-morph ratios in finite populations, and for the evolution of the polymorphism are discussed.
Introduction

Tristylos flowering plants are generally characterized by populations comprised of three style morphs possessing a reciprocal arrangement of stigmas and anthers. The only compatible mating in most heterostylous populations is between anthers and stigmas of equivalent height and this results in phenotypic disassortative mating (Darwin 1877; Ganders 1979; Barrett 1992). Fisher (1944) predicted that populations at equilibrium should contain equal ratios of the long-, mid- and short-styled morphs (hereafter L-, M and S- morphs), as long as there are no fitness differences between the morphs. Six angiosperm families are known which contain tristylos species but genetic investigations of the polymorphism, using controlled crosses and progeny ratios, have only been conducted in three families (reviewed in Barrett 1993).

Despite the polyphyletic origins of tristyly, a similar two-locus model (S and M loci) with dominance and epitasis between the two diallelic loci govern the inheritance of the floral polymorphism in Lythraceae (Barlow 1923; Fisher and Mather 1943; Eckert and Barrett 1993), Oxalidaceae (Von Ubisch 1926; Fisher and Martin 1948; Fyfe 1950, 1956; Weller 1976; Leach 1983; Bennett et al. 1986; Weller 2015) and Pontederiaceae (Gettys and Wofford, 2008; S.C.H. Barrett unpubl. data). Genotypes governing the three style morphs under this model of inheritance for the diploid case are: L-morph – sm/sm; Mid-morph – sM/sM, sm/sm; S-morph – SM/sM, SM/sm, Sm/sM, Sm/sm). Why the same inheritance pattern has evolved independently in the unrelated flowering plant families remains unknown.

Homozygosity for the dominant allele at the S locus is usually absent in tristylos species because disassortative mating enforced by trimorphic incompatibility prevents selfing and assortative mating, but in self-compatible tristylos species genotypes homozygous at the S locus are possible. Barrett et al. (1989) reported 837 shorts and 333 non-shorts (n=1,170 plants) pooled
from selfing 12 S-/s- genotypes and the authors noted a deficit of S-morph plants in these
progenies. They suggested that the deficiency could be due to viability selection against
homozygous dominant short-styled individuals that are produced during selfing. Heterozygosity
at the S locus could result in the accumulation of deleterious mutations (‘sheltered load’ -
Uyenoyama 1997; van Oosterhout 2009) leading to selection acting against genotypes
homozygous dominant at the S locus when they are produced (Mather and De Winton 1941;
Strobeck 1980). The shared genetic basis of this polymorphism across multiple plant families has
stimulated much theoretical and empirical attention.

Variation from the general diploid two-locus model has been reported in several
tristyloous species. For example, polyploidy with double reduction occur in autotetraploid
Lythrum salicaria (Fisher 1941, 1944; Fisher and Martin 1947). In Oxalis, a large genus with
numerous tristyloous species, dominance reversal with the S-morph recessive to the L-morph
(Von Ubisch 1926; Fyfe 1956; Mulcahy 1964), a putative third locus (Fyfe 1956; Bennett et al.
1986; Trognitz and Hermann 2001), and different degrees of genetic linkage between the S and
M loci (Fisher and Martin 1948; Fyfe 1950, 1956; Leach 1983; Weller 1976, 2015) are each
reported in various species. In contrast, there is no evidence of linkage in Lythraceae (Lythrum
salicaria - Fisher and Mather 1943; Decodon verticillatus - Eckert and Barrett 1993) and
Pontederia cordata of the Pontederiaceae (Gettys and Wofford 2008). However, a preliminary
study of E. paniculata suggested the possibility of linkage between the S and M loci (S.C.H.
Barrett unpublished data), and this finding motivated the present study on this species.

Here, I investigate further if linkage occurs between the S and M loci in E. paniculata.
This species is diploid (n=8), annual, and belongs to the Pontederiaceae (Monocotyledoneae:
Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009), a small family of
freshwater aquatics primarily native to the Neotropics. All known *E. paniculata* populations that are tristylos occur in the arid caatinga region of N.E. Brazil. Earlier controlled crosses and progeny tests established that two diallelic loci, with *S* epistatic to *M* control tristyly in *E. paniculata* (Barrett et al. 1989; S.C.H. Barrett unpubl. data). To investigate whether linkage occurs in *E. paniculata* I conducted controlled selfs of the S-morphs, putatively heterozygous at both the *S* and *M* loci, and also crossed them to plants of the L-morph and tested the segregation ratios of style morphs in the progeny against models with complete linkage and independent assortment between the *S* and *M* loci.

**Materials and methods**

*Test crosses*

Table 5.1. Locality and morph structure of the four trimorphic *Eichhornia paniculata* populations in N.E. Brazil from which the plants of the S-morph used in this study originated via subsequent crossing. Data from Barrett et al. (1989) and S.C.H. Barrett unpubl data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B187</td>
<td>Pernambuco</td>
<td>Garanhuns</td>
<td>1750</td>
<td>0.44</td>
<td>0.31</td>
<td>0.25</td>
<td>0.97</td>
</tr>
<tr>
<td>B192</td>
<td>Pernambuco</td>
<td>Cupira*</td>
<td>8000</td>
<td>0.35</td>
<td>0.45</td>
<td>0.09</td>
<td>0.81</td>
</tr>
<tr>
<td>B197</td>
<td>Pernambuco</td>
<td>Camocim de SãoFélix*</td>
<td>1500</td>
<td>0.37</td>
<td>0.55</td>
<td>0.06</td>
<td>0.74</td>
</tr>
</tbody>
</table>
*These populations also contain modified selfing variants and therefore the ratios of the three style morphs do not sum to 1.0.

Controlled selfs and crosses involving plants of the M and S-morph sampled from outcrossing populations in N.E. Brazil and assessments of progeny ratios in the F1 progeny revealed four S-morph parents that were suspected of being heterozygous at both the S and M loci (SM/sm or Sm/sM). Note that I am using this convention to write putative genotypes, in order to highlight allelic associations between the S and M loci; these genotypes are also often designated as SsMm.

I crossed the four putative double heterozygotes to plants of the L-morph. Three of the four S- and L-morph pairs were crossed reciprocally with both plants acting as ovule and pollen parents. I also self-fertilized two of the S-morph plants. The genotypes sampled for selfing and performing test crosses originated from four trimorphic populations of contrasting population sizes and morph diversity (Table 5.1). I used the following designation to uniquely label the plants in this study: population-individual-morph (e.g. B197-2-S). In September 2012, I germinated multiple seeds from each self and each reciprocal cross in separate 12.7 cm plastic pots filled with potting soil in individual trays filled with water in a glasshouse at the University of Toronto. During the germination period, temperatures in the glasshouse were maintained at around 35°C to stimulate germination with a 16-hour day length. I misted the seedlings with water regularly to maintain humid conditions and promote growth. When the seedlings were at the three-leaf stage, I transplanted them into individual 7 cm plastic pots filled with potting soil using fine forceps and transferred them onto benches flooded with water with a separate bench for each cross. At this time, the temperature in the glasshouse was reduced and maintained between 26-31°C. I grew plants in these conditions till they flowered and scored the style morph.
of plants resulting from the crosses. The vast majority of plants (93-98%) flowered from the selfs and crosses.

Inferring genotypes and estimating linkage

I compared the observed progeny style morphs ratios to expected ratios assuming various possible genotypes for the S-morph parents with complete linkage or independent assortment of the S and M loci (Table 5.2). I assessed if the progeny segregation ratios for the two reciprocal crosses generated from the same S-morph plant differed using $\chi^2$ tests. If the $P$-values of the $\chi^2$-square tests were greater than 0.05, I pooled the progeny from the reciprocal crosses. I calculated the recombination fraction ($r$) as the number of recombinant non-short-styled offspring ($R$) divided by the total number of non-short-styled offspring ($NS=M+L$). I also calculated the odds ratio (LOD) score, which is a likelihood ratio test that can be used to detect linkage (Morton 1955). It evaluates the likelihood of linkage ($r=\hat{r}$) against the likelihood of the null hypothesis of independent assortment ($r=0.5$). The LOD score was calculated as:

$$LOD = \log_{10} \frac{\left(\frac{NS}{R}\right)^{1-rNSR}}{\left(\frac{NS}{0.5R}\right)^{0.5NS0.5R}}$$

I also performed individual $G$-tests to test if segregation ratios in the progeny differed from expectations of complete linkage in the selfs and crosses.
Table 5.2. Expected segregation of style morphs in progeny from selfs and crosses involving short-styled (S) and long-styled (L) homozygous recessive individuals (sm/sm) under various genotypes for the S-morph parent. $r$ is the recombination fraction between the $S$ and $M$ loci. L, M, and S refer to the long-, mid- and short-styled morphs, respectively.

<table>
<thead>
<tr>
<th>Self/cross</th>
<th>$r$</th>
<th>L</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>sm/sm x Sm/sm</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sm/sm x Sm/sm</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sm/sm x SM/sm</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>sm/sm x SM/sm</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sm/sm x Sm/sM</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>sm/sm x Sm/sM</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sm/sm x SM/sM</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sm/sm x SM/sM</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sm/sm self</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Sm/sm self</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SM/sm self</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>SM/sm self</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
Results

Of the four short styled plants used in test crosses, only one cross and its reciprocal provided information on the presence and strength of linkage between the $S$ and $M$ loci. B197-1-S x B187-1-L resulted in 10 L-morph, 483 M morph and 534 S-morph plants in the progeny and B187-1-L x B197-1-S resulted in 13 L-morph, 550 M morph and 717 S-morph plants in the progeny (Table 5.3). Comparisons of the observed and expected style morphs ratios (Table 5.2) in the progeny indicated that the only possible genotype of B197-1-S was that it is heterozygous at both the $S$ and $M$ loci, as all three morphs were evident in both reciprocals crosses. $\chi^2$ test for heterogeneity between these reciprocal crosses resulted in a $P$-value greater than 0.05 ($\chi^2=3.091$, degrees of freedom=2). As there was no significant evidence of heterogeneity between the reciprocal crosses, we pooled the progeny from the reciprocal crosses and LOD score testing for linkage between the $S$ and $M$ loci in B197-1-S indicated an $r$ of 127.7 and the estimate of the map distance between the $S$ and $M$ loci was 2.7 cM. The two loci in B197-1-S were linked in repulsion ($Sm/sM$). As I detected recombinants in only one genotype that was used in the crossing program, I was unable to calculate variances for the degree of genetic linkage between the $S$ and $M$ loci.
Table 5.3. Observed segregation of style morphs in *Eichhornia paniculata* progeny from selfs and crosses involving short-styled (S) individuals suspected of being double heterozygotes *(SM/sm or Sm/sM)* and long-styled (L) individuals *(sm/sm)*.

<table>
<thead>
<tr>
<th>Ovule parent x pollen parent</th>
<th>Frequency</th>
<th>Total progeny</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>B192-1-S self</td>
<td>0.31</td>
<td>0</td>
</tr>
<tr>
<td>B192-1-S x B192-1-L</td>
<td>0.49</td>
<td>0</td>
</tr>
<tr>
<td>B192-1-L x B192-1-S</td>
<td>0.47</td>
<td>0</td>
</tr>
<tr>
<td>B192-2-S x B192-2-L</td>
<td>0.43</td>
<td>0</td>
</tr>
<tr>
<td>B192-2-L x B192-2-S</td>
<td>0.44</td>
<td>0</td>
</tr>
<tr>
<td>B197-1-S self</td>
<td>0</td>
<td>0.35</td>
</tr>
<tr>
<td>B197-1-S x B187-1-L</td>
<td>0.01</td>
<td>0.47</td>
</tr>
<tr>
<td>B187-1-L x B197-1-S</td>
<td>0.01</td>
<td>0.43</td>
</tr>
<tr>
<td>B211-1-L x B211-1-S</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The remaining three test crosses were not informative for assessing the presence or strength of genetic linkage between the *S* and *M* loci as no recombinant phenotypes were evident in their progenies. The B192-1-S x B192-1-L cross and its reciprocal both resulted in ratios close to
1S:1L and the B192-1-S self resulted in a morph ratio close to 2S:1L in the progeny (Table 5.3). The B192-2-S x B192-2-L cross and its reciprocal also resulted in a morph ratio close to 1S:1L. Comparisons of the observed to expected ratios under various possible genotypes for the S-morph parent (Table 5.2) indicated B192-1-S and B192-2-S could either be of genotypes $SM/sm$ or $Sm/sm$. The B211-1-L x B211-1-S cross resulted in a morph ratio close to 1S:1L (Table 2.3) and comparisons to expectations indicated that B211-1-S could be of genotype $Sm/sM$ or $SML/sM$ (Table 2.2).

Of the two selfed progenies of short-styled plants that I produced, there was a deficit of S-morph plants in one progeny. The 2S:1M ratio resulting from the B197-1-S self (95 M morphs, 176 S-morphs) was significantly different from three short to one non-short expectation with or without linkage between the $S$ and $M$ loci (Table 5.3, deviation from 3S:1M – $G = 5.95$, $P = 0.015$). In contrast, the ratio of progeny resulting from the B192-1-S self (73 L-morphs, 163 S-morphs) was not significantly different from the three short to one non-short expectation (Table 5.3, deviation from 3S:1M – $G = 1.78$, $P > 0.1$).

**Discussion**

*Linkage between the $S$ and $M$ locus and sheltered load at the $S$ locus*

Comparisons of observed and expected ratios of style morphs in test crosses of short-styled plants of *E. paniculata* indicated that one (B197-1-S) of the four plants was unequivocally heterozygous at both the $S$ and $M$ loci yielding a total of 23 recombinant plants among the 1986 offspring that were grown to flowering. Moreover, my calculation of the degree of genetic linkage indicated that the two loci were strongly linked in repulsion ($Sm/sM$) with an estimated
map distance of 2.7 cM. Among the remaining three short-styled plants the progeny ratios could have resulted from several alternative genotypes and therefore these plants provided no information on the presence and degree of linkage between the $S$ and $M$ loci. Assessing the presence and strength of genetic linkage requires the detection one or more recombinant phenotypes in progeny resulting from controlled crosses and, if linkage is very strong, requires a very large number of progeny to assess the strength of linkage. It is still possible that the plants of the S-morph that produced no recombinants were double heterozygotes and the lack of detection of recombinant phenotypes was due to an inadequate number of progeny analyzed. Indeed, I only detected linkage in the two reciprocal test crosses in which close to 1000 progeny were grown to flowering.

There was a significant deficit of plants of the S-morph in the selfed progeny of a single short-styled plant (B197-1-S) compared to the expectation of 3 short : 1 non-short. The lower frequency of S-morphs in the progeny from this self could be due to selection acting against $S$-/$S$-genotypes similar to the results and conclusions reached by Barrett et al. (1989). While the B192-1-S self also exhibited a slightly lower frequency of S-morph plants in the progeny, the observed ratio was not significantly different from the 3:1 expectation, possibly owing to a smaller number of progeny analyzed. My sample of plants of the S-morph was limited and therefore to test this hypothesis rigorously a larger samples of maternal parents and progenies would be required.

**Consequences of linkage for the dynamics of style morph ratios in populations**

The theoretical results of Heuch and Lie (1985) demonstrate that the equilibrium ratio of 1:1:1 (isoplethy) of the three style morphs is not affected by the presence or absence of linkage between the $S$ and $M$ loci. However, many tristylos populations are not at equilibrium when
sampled (Morgan and Barrett 1988; Eckert and Barrett 1995; Eckert et al. 1996b). Founder effects in tristylos species with linkage between the S and M loci could affect their evolutionary dynamics and approach to equilibrium in ways different from species without linkage. For example, if populations were initiated by founders involving a single double heterozygous genotype (SM/sm or Sm/sM) without linkage between the two loci, selfing would rapidly generate all three morphs, assuming that the species was self-compatible (Heuch and Lie 1985; Morgan and Barrett 1988; Eckert and Barrett 1992). However, with linkage this process would be considerably slower and depend on the recombination fraction. Another consequence of genetic linkage between the S and M loci is that there would be reduced chances for sib-mating when selfing a linked double heterozygous genotype because the resulting progeny would be comprised predominantly of two rather than three morphs (Fisher 1949; Fyfe 1950). More explicit modeling work is necessary to determine how linkage might influence the dynamics of style morph ratios in non-equilibrium populations of tristylos plants and how this might influence patterns of disassortative mating.

Variation in linkage relations among tristylos species

Tristylos species with and without linkage occur in two (Oxalidaceae and Pontederiaceae) of the three families in which the inheritance of the polymorphism has been investigated through controlled crosses and progeny tests (summarized in Table 5.4). My estimates of genetic linkage between the S and M loci among Oxalis species indicate that the two loci are strongly linked (r = 0.05-0.07) in O. valdiviensis (Fisher and Martin 1948; Fyfe 1950), O. lasiandra (Weller et al. 1976) and O. alpina (Weller et al. 2015) but are only very weakly linked in O. articulata (Fyfe 1956, r = 0.26). Note, I estimated r in Oxalis species by dividing the number of recombinant genotypes by the total number of progeny in the informative crosses. These estimates were not
explicitly reported in the published studies. Test crosses involving L- and S-morph parents produced style morphs in the progeny that deviated significantly from ratios expected under independent assortment of the $S$ and $M$ loci in *O. compressa* (Leach 1983), *O. macrocarpa* and *O. magnifica* (Weller 1976). As no recombinants were detected in the respective crosses for these three species, accurate estimates of the degree of linkage were not possible to make. Even so, my estimate of the upper bound of $r$, assuming that at least one recombinant progeny would eventually be observed, ranged from 0.02-0.07. Similar assessments of genetic linkage between the $S$ and $M$ loci for *O. physocalyx*, *O. rosea* and *O. tuberosa* did not favour the hypothesis of genetic linkage (Weller 1976). In Lythraceae there is also no evidence of linkage between the $S$ and $M$ loci in *L. salicaria* (Fisher and Mather 1943) and *D. verticillatus* (Eckert and Barrett 1993) even though these studies analyzed style-morph ratios from crosses involving a large number of progeny. In *P. cordata*, a tristylos species from the sister genus to *Eichhornia* (see Ness et al. 2011), linkage was not detected in crosses between plants of the S-morph that were double heterozygotes, synthesized from controlled crosses as in this study, and the L-morph (Gettys and Wofford 2008), although sample sizes were smaller than in the present study (~650 plants). Also, preliminary evidence suggests that a congener of *E. paniculata*, the clonal tetraploid *E. crassipes*, shows no evidence of linkage between the $S$ and $M$ loci, based on controlled crosses of the S-morph (S.C.H Barrett unpubl. data). It is unclear what role taxon age might have in the evolution of linkage between the $S$ and $M$ loci, as species with and without linkage occur within the same plant family and genus (e.g. *Oxalis*, *Eichhornia*). The lack of consistent patterns involving the presence and degree of genetic linkage between the $S$ and $M$ loci among multiple tristylos species raises several evolutionary questions concerning the origin and function of linkage for the genes controlling tristyly.
Table 5.4. Studies on the inheritance of tristyly and degree of linkage between the $S$ and $M$ loci in three unrelated plant families.

<table>
<thead>
<tr>
<th>Family</th>
<th>Species</th>
<th>Presence of linkage</th>
<th>Progeny size</th>
<th>$r^A$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lythraceae</td>
<td>Lythrum salicaria</td>
<td>No</td>
<td>~673</td>
<td>-</td>
<td>Fisher and Mather (1943)</td>
</tr>
<tr>
<td>Lythraceae</td>
<td>Decodon verticillatus</td>
<td>No</td>
<td>828</td>
<td>-</td>
<td>Eckert and Barrett (1993)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>Oxalis physocalyx</td>
<td>No</td>
<td>81</td>
<td>-</td>
<td>de Barros Pacheco and Coleman (1989)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. rosea</td>
<td>No</td>
<td>105</td>
<td>-</td>
<td>Bennett et al. (1986)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. tuberosa</td>
<td>No</td>
<td>543</td>
<td>-</td>
<td>Trognitz and Hermann (2001)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. valdiviensis</td>
<td>Yes</td>
<td>746, 1397</td>
<td>0.05, 0.07</td>
<td>Fisher and Martin (1948), Fyfe (1950)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. articulata</td>
<td>Yes</td>
<td>1289</td>
<td>0.26</td>
<td>Fyfe (1956)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. compressa</td>
<td>Yes</td>
<td>208</td>
<td>&lt;0.02</td>
<td>Leach (1983)</td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td>O. lasiandra</td>
<td>Yes</td>
<td>97</td>
<td>0.02</td>
<td>Weller (1976)</td>
</tr>
<tr>
<td>Family</td>
<td>Species</td>
<td>Parity</td>
<td>Recombination Fraction</td>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------</td>
<td>--------</td>
<td>------------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td><em>O. macrocarpa</em></td>
<td>Yes</td>
<td>45</td>
<td>&lt;0.1 Weller (1976)</td>
<td></td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td><em>O. magnifica</em></td>
<td>Yes</td>
<td>136</td>
<td>&lt;0.04 Weller (1976)</td>
<td></td>
</tr>
<tr>
<td>Oxalidaceae</td>
<td><em>O. alpina</em></td>
<td>Yes</td>
<td>~400</td>
<td>0.05-0.07 Weller et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>Pontederiaceae</td>
<td><em>Pontederia cordata</em></td>
<td>No</td>
<td>650</td>
<td>- Gettys and Wofford (2008)</td>
<td></td>
</tr>
<tr>
<td>Pontederiaceae</td>
<td><em>Eichhornia paniculata</em></td>
<td>Yes</td>
<td>1986</td>
<td>0.03 This study</td>
<td></td>
</tr>
</tbody>
</table>

^ The recombination fraction ($r$) is the number of recombinant offspring divided by the total number of offspring and is calculated for species with evidence of linkage between the $S$ and $M$ loci.

Increased linkage between genes is a key step in the evolution of supergenes and linkage between the $S$ and $M$ loci in tristylos species could have been important in the origin of one of the two loci. Darlington and Mather (1949) emphasized the role of inversions in the origin of the linkage group governing distyly since this mechanism could bring about favorable gene combinations and suppress the role of recombination in producing maladapted trait combinations. Similarly, Mather (1950) suggested that selection might act on initially unlinked loci, such as self-incompatibility loci, favoring increased linkage and leading to the evolution of a composite functional unit. Several authors have also invoked linkage between the two loci governing tristyly as being potentially important in the evolution of a second tristyly locus (Crowe 1964; Sved 1965; Richards 1986; Olmstead 1990). For example, one possibility for the
evolution of a second locus is via gene duplication of the first locus with the two loci subsequently diverging in function, but this hypothesis remains to be rigorously investigated. It is also possible that linkage may have been gained and then lost in some tristylos lineages. More detailed studies of other tristylos species may provide insight into how linkage between the $S$ and $M$ loci originated and why variation in the presence and absence of linkage occurs across different tristylos lineages. In particular, it will be important to determine if linkage is a necessary condition for the evolution of the putative supergenes that have been postulated for tristylos species.
Chapter 6

The genetic architecture of style morph variation in tristylos Eichhornia paniculata and modifiers associated with evolutionary transitions to self-fertilization.

Summary

Although the morphology and functional significance of tristyly is well-established, the genetic architecture governing this complex floral syndrome remains unknown. The polymorphism is characterized by three discrete style morphs that possess a reciprocal arrangement of stigma and anther heights governed by two diallelic loci ($S$ and $M$), with the $S$ locus epistatic to the $M$ locus. Tristyly generally functions to promote animal-mediated cross-pollination, but is susceptible to evolutionary breakdown resulting in transitions to predominant self-fertilization. The morphological basis to this change in mating system involves modifications to stamen position causing autonomous self-pollination in semi-homostylos variants. Here, we investigate the genetic architecture of the $M$ locus governing the control of the long- and mid-styled morphs in tristylos Eichhornia paniculata and the genetic basis of the breakdown to selfing in each style morph. We crossed independently derived semi-homostylos long- and mid-styled individuals fixed for alternate alleles at the $M$ locus ($ssmm$ and $ssMM$, respectively) and backcrossed an $F_1$ plant to the parental $ssmm$ genotype. We phenotyped and genotyped 462 backcross progeny using 1,450 markers to identify genetic regions governing style length and anther height variation. We performed composite interval mapping (CIM) to identify the interval regions containing quantitative trait loci (QTL). The QTL associated with style-morph variation mapped to the center of one linkage group, and a narrow section of this interval contained polymorphic genes with elevated Tajima's $D$. Further, we bulked and sequenced progeny segregating for
alternate alleles at the $M$ locus to identify 334 genes containing SNPs potentially linked to the $M$ locus. The stamen modifications characterizing the two semi-homostylyous variants were governed by loci on separate linkage groups. Our results indicate that the $M$ locus localizes to a single genetic region that shows signatures consistent with balancing selection resulting from the frequency-dependent selection that maintains tristyly within populations. Transitions to selfing in the two semi-homostylyous morphs appear to have originated by different loci involving contrasting sets of mating-system modifier genes unlinked to the $M$ locus.

**Introduction**

Heterostyly is a floral polymorphism characterized by two (distyly) or three (tristyly) style morphs with a reciprocal arrangement of stigmas and anthers (Darwin 1877; Ganders 1979; Barrett 1992). The polymorphism has evolved in at least 28 flowering plant families and promotes animal-mediated cross-pollination. Compatible mating in most heterostylous populations occurs between anthers and stigmas of equivalent height. Although there has been considerable work on the function and adaptive significance of heterostyly (reviewed in Lloyd and Webb 1992; Barrett and Shore 2008), little is known about its underlying genetic architecture. Most heterostylyous species are distylyous with populations comprised of long- and short-styled morphs (hereafter L- and S-morphs) and this polymorphism is governed by the $S$ locus (Bateson and Gregory 1905; reviewed in Lewis and Jones 1992). Tristylyous species are restricted to six flowering plant families and populations are composed of long-, mid- and short-styled morphs (hereafter L-, M- and S-morphs) reviewed in Barrett (1993). The polymorphism is governed by two diallelic loci, $S$ and $M$, with the $S$ locus epistatic to the $M$ locus (Barlow 1923; Fisher and Mather 1943; Fyfe 1950; reviewed in Lewis and Jones 1992).
Investigations of the molecular genetic architecture of heterostyly have focused solely on distylous species. The S locus is presumed to be a supergene complex composed of a cluster of linked genes with co-ordinated function maintained by balancing selection (Ernst 1928; Charlesworth and Charlesworth 1979a,b; Lewis and Jones 1992; Charlesworth 2016). Similar supergene complexes have been found to control mimicry in Heliconius butterflies (Joron et al. 2011; Kunte et al. 2014), social behavior in ants (Wang et al. 2013), banding patterns in snails (Richards et al. 2013) and pigmentation and social behavior in white-throated sparrows (Tuttle et al. 2016). Inferences about the order and number of linked genes within the S locus have been made in Primula species, based on the detection of rare recombinant genotypes that give rise to mutant phenotypes (Ernst 1928, 1955; Dowrick 1956). Three loci have been identified in the Primula linkage group (hereafter LG) and the order of these loci is G, P and A: with G controlling style length and incompatibility, P controlling pollen size and incompatibility, and A determining anther height (reviewed in Charlesworth and Charlesworth 1979b; Barrett and Shore 2008). The development of recent molecular genotyping approaches has facilitated the dissection of the genetic architecture of the heterostylish polymorphism and the goal of this study is to apply these approaches to a tristyly species.

Currently, a handful of studies have identified putatively linked markers and genes to the S locus in distylous species. Yoshida et al. (2011) reported that genetic regions associated with style length and anther height in distylous Primula sieboldii were both found in the same ~30 cM interval on a single LG. Furthermore, a 1.5 Mbp S-linked region in P. vulgaris with a 1 Mbp interval containing 82 candidate genes (Li et al. 2015), a 1.6 Mbp S-linked region in Turnera subulata containing several genes (Labonne and Shore 2011), and two genes completely linked to the S locus in Fagopyrum esculentum (Yasui et al. 2012) have each been identified. Ushijima
et al. (2012) found four genes that were differentially expressed or differentiated between the L- and S-morphs in *Linum grandiflorum* and molecular analysis suggested they might be related to the distylosus syndrome. Most recently, Nowak et al. (2015) assembled the draft genome of distylosus *P. veris* and performed bulk segregant analysis on L- and S-morph pools. They identified 13 variants in candidate S-linked regions and 113 genes with morph-biased expression and one that was completely silenced in flowers of the L-morph. In contrast, almost nothing is known about the number, order and location of genes governing the S and M loci in tristylosus species.

A common feature of heterostylosus groups is the breakdown of the polymorphism and the evolution of a range of derived mating systems. The most frequent transition involves the evolution of selfing variants with stigmas and anthers at a similar height within a flower (Charlesworth and Charlesworth 1979b; Barrett 1989; Weller 1992). In distylosus and tristylosus groups these are referred to as homostylosus and semi-homostylosus variants or morphs, respectively. In the latter case, usually only a single anther level is involved in floral modifications hence the term “semi-homostyly” (see Ornduff 1972, Barrett 1988 for illustrations of various semi-homostylosus variants). The evolutionary breakdown of tristyly is especially well documented in *Eichhornia* (Pontederiaceae), where in each of the three tristylosus species semi-homostylosus variants have established in natural populations (reviewed in Barrett 1988). Studies of *E. paniculata* indicate multiple independent transitions to selfing via semi-homostyle evolution (Husband and Barrett 1993; Barrett et al. 2009). The loss of herkogamy (stigma-anther separation) in semi-homostylosus variants is governed by recessive modifiers non-allelic to the S or M loci (Fenster and Barrett 1994, Vallejo-Marín and Barrett 2009). However, the location,
genetic basis and effect size of the loci underlying the breakdown of the polymorphism in *E. paniculata* remains unknown.

Here, we investigate the genetic architecture of tristyly and its breakdown to selfing in *E. paniculata*. The species is a diploid (*n*=8) emergent, annual aquatic in the Pontederiaceae (Monocotyledoneae: Commelinales, Cantino et al. 2007; Angiosperm Phylogeny Group 2009), a small, largely Neotropical, family of freshwater aquatics. The tristylos polymorphism in *E. paniculata* is associated with a cryptic trimorphic incompatibility system, in which inter-morph cross-pollination is favored over intra-morph cross-pollination and self-pollination (Cruzan and Barrett 1993, 2015) thus promoting phenotypic disassortative mating (Barrett et al. 1987). In tristylos species of Pontederiaceae, the flowers of each style morph have upper- and lower-stamen levels each composed of three stamens (Richards and Barrett 1984, 1992). The L-morph has mid- and short-level stamens, the M-morph has long- and short-level stamens and the S-morph has long- and mid-level stamens. Tristyly in *E. paniculata* is governed by two diallelic loci with the *S* locus epistatic to the *M* locus. The S-morph is governed by a dominant *S* allele (genotypes: *SsMM, SsMm, Ssmm*), the mid-styled morph (M-morph) by a dominant *M* allele (genotypes: *ssMM or ssMm*) and the L-morph is of genotype *ssmm* (S.C.H. Barrett unpubl. data). Controlled crosses and progeny tests indicate that the two loci are linked and separated by a genetic distance of 2.7 cM (Chapter 5). Given the close linkage between the two loci, *E. paniculata* is useful for uncovering the overall genetic architecture of the tristylos syndrome as identifying the location of one locus should facilitate the discovery of the second locus by assessing closely linked genetic markers. Thus, in this study of *E. paniculata* we restrict our attention to a study of the genetic architecture of the *M* locus.
The largest concentration of *E. paniculata* populations is in the dry, arid, caatinga region of N.E. Brazil, with a smaller number of populations in Cuba and Jamaica, and a few isolated populations in Nicaragua and Mexico. Brazilian populations are largely tristyloous and outcrossing whereas populations in the Caribbean and Central America are predominantly selfing (Barrett et al. 1989; Barrett and Husband 1990; Barrett et al. 2009). The occurrence of these selfing populations is consistent with Baker’s law (Baker 1955), and it has been estimated that colonization of the Caribbean from Brazil occurred ~120,000 years ago (Ness et al. 2010). Semi-homostylous variants of the M-morph (hereafter M’) occur sporadically in N.E. Brazil, but predominate in Jamaica and Cuba (Barrett 1985; Barrett et al 1989). In the M’ semi-homostyle, various stamens of the short-level have elongated to the mid-level of the flower resulting in contact between the anthers and mid-level stigmas (Richards and Barrett 1992; Barrett et al. 2009). In Nicaragua and Mexico, a different semi-homostylous variant occurs which is absent from the Caribbean and Brazil. This variant is a semi-homostylous long-styled morph (hereafter L’) resulting from a separate breakdown of tristyly (Barrett et al. 2009). Unlike the M’ semi-homostyle, several or all of the three mid-level stamens of this variant can be in contact with the long-level stigma, also due to stamen elongation. A recombinational origin for homostyles occurs in several distyloous species (Dowrick 1956; Shore and Barrett 1985; reviewed in Barrett and Shore 2008). Although semi-homostyles have often been reported in tristyloous taxa (e.g. *Lythrum salicaria* – Stout 1925, Esser 1953; *Oxalis* spp. – Ornduff 1972; *Eichhornia* spp. – Barrett 1978, 1979, 1988), there is no evidence that these have originated rapidly through recombination in the heterostyly linkage group (see Charlesworth 1979). Molecular studies indicate that samples of the semi-homostyloous L’ morph from Nicaragua share more SNPs with outcrossing populations from Brazil than with M’ semi-homostyles from Jamaica (Ness et al.
2011, 2012), a pattern consistent with their separate origins. Also, controlled crosses suggest that the transitions to selfing in the M’ semi-homostyle involves one or two recessive modifier genes whereas both major and minor genes appear to control stamen modification in L’ semi-homostyle (Fenster and Barrett 1994; Barrett et al. 2009), although this hypothesis remains to be investigated using molecular approaches. Thus, morphological, geographic and genetic evidence are in accord with the hypothesis that the two semi-homostyous variants have independent origins. Therefore, *E. paniculata* provides an excellent opportunity to study the genetic architecture of morphological adaptations associated with higher selfing rates and investigate if the same or different genes are involved in independent transitions to predominant self-fertilization in a single species of a flowering plant.

The primary goal of this study is to map the broad genomic region containing the *M* locus and investigate the genetic architecture of independent mating-system transitions in *E. paniculata*. As highly inbred lines were not available for this species, we used autogamous genotypes from Central America and the Caribbean for genetic mapping, both of which were fixed for alternative alleles at the *M* locus (*ssmm* and *ssMM*, respectively). An assumption of our study is that these highly selfing populations would have largely homozygous genomes. This expectation is supported by evidence from studies of both allozyme and nucleotide diversity which indicate relatively low levels of heterozygosity in Caribbean and Central America populations (Barrett and Husband 1997; Ness et al. 2010). We crossed the homozygous L’ semi-homostyle to the M’ semi-homostyle and backcrossed the F₁ to the homozygous recessive parent (L’) to generate an F₂ population that segregated for the *M/m* allele. We measured floral and vegetative traits including style length and anther height of the resulting F₂ progeny and analyzed the correlations among traits using principal component analyses to control for dependencies.
We then genotyped F2 progeny using the genotyping-by-sequencing (GBS) approach (Elshire et al. 2011) and performed composite interval mapping (CIM) (Zeng 1994; Jansen and Stam 1994) to identify the regions governing style length and anther height. We also estimated population genetic parameters, obtained from a previously generated dataset from outcrossing *E. paniculata* populations (Chapter 3). Given that negative frequency-dependent selection maintains the tristylos polymorphism, we expected regions associated with the *M* locus to show evidence of balancing selection including elevated levels of polymorphism and Tajima's *D*, and possibly be associated with regions of low recombination (Yasui et al. 2012; Charlesworth 2016). We bulked progeny with long- versus mid-length styles and sequenced RNA from these two pools to generate a list of potential candidate genes governing style length, and we sequenced and assembled the parental genomes to better understand the genomic context of candidate genes and markers and estimated the genome size to facilitate the assembly. We also assessed patterns of polymorphism, gene expression and molecular function of candidate genes to infer their associations to the *M* locus.

Our study addressed the following specific questions: 1) What is the number, size and relative influence of genetic regions governing the *M* locus in *E. paniculata*? We identified the genetic region governing style length and stamen levels, traits distinguishing the style morphs in tristylos species, and scanned for evidence of balancing selection along this region. 2) Are there genes that segregate for alternate SNPs, or that show differential gene expression between the L’ and M’ semi-homostyles in F2 progeny and, if so, what is their molecular function? Such genes are strong candidates for being associated with the *M* locus. 3) Are different genetic regions responsible for the independent transitions to selfing in the two semi-homostylos morphs? Given their independent origins and contrasting patterns of stamen modification we predicted
that the regions of the genome governing selfing in the two variants would likely differ. Our study represents the first attempt to localize the genetic region governing style length and anther height and to investigate whether independent transitions to selfing in a plant species are associated with contrasting suites of modifier genes governing floral morphology.

**Materials and methods**

*F₁ and F₂ backcross populations*

We performed controlled crosses to generate individuals segregating for the \( M/m \) allele to investigate the genomic architecture of the \( M \) locus. We crossed a semi-homostylous \( M' (ssMM) \) individual from Slipe, St. Elizabeth, Jamaica with a semi-homostylous \( L' (ssmm) \) individual from Oaxaca, San Mateo del Mar, Nr. Tehuantepec, Mexico. The cross involved the \( M' \) morph as the female parent and the \( L' \) morph as the male parent. The female parent was emasculated prior to cross-pollination. Seed from this cross was germinated to produce an \( F₁ \) of mid-styled plants (all \( ssMm \)) and a single plant was chosen to backcross to the \( L' (ssmm) \) parent with the \( ssmm \) plant as the female parent. In late July 2014, we soaked the \( F₂ \) seeds in a glass beaker for one week and then transferred young seedlings into individual 7 cm plastic pots filled with 1:1 ratio of Permier 3 potting soil and sand in a single glasshouse at the University of Toronto with 26-31°C temperature and a 16-hour day length. Plants were arranged with equal spacing on a single flooded bench. The \( F₂ \) backcross population was comprised of ~600 plants of which ~550 flowered with the first plant commencing anthesis after 55 days and all measurements were completed within a 40-day period. We phenotyped all flowering progeny, but were only able to obtain near complete genotype data for 462 progeny (see supplementary methods for genotype quality controls).
**Floral measurements**

We performed several floral measurements to measure the extent of phenotypic variation in the F2 backcross progeny. At the start of flowering for each plant we recorded the date, and two weeks after the flowering of each plant commenced we measured flower (perianth) breadth, width and length, style length and the height of each of the six stamens within a flower (Appendix IV - Figure 6S1). Flower breadth and width were the longest distances along the vertical and horizontal axes, respectively, of the top view of the flower using the nectar guide to orient the flower. Flower length, style length and anther height were the distances from the base of the floral tube to the tip of the flower or the floral organ, as shown from the side view of the flower. We sequentially labeled anthers from 1 to 6 in each morph corresponding to the position relative to the base of the flower with 6 being furthest away from the base. We measured the phenotypes of two flowers per plant and used the average value in our analyses. We also measured the height of the plant from the surface of the soil to the tip of the tallest leaf. We estimated the mean and 95% confidence intervals (CI) for plant height and all floral traits. We compared measurements of plant height, flower width, breadth and length and style length between progeny of the L´ and M´ morphs using two sample t-tests assuming unequal variances. We performed principle component analyses on all measured floral traits. We used a screeplot to determine the number of meaningful components among the total measurements.

**Genotyping-by-sequencing**

We used the GBS approach to identify and generate genetic markers for F1 and F2 progeny. For all F2 progeny and the ssMM, ssMm and ssmm parental genotypes, we collected floral bud tissue for DNA extractions. We extracted DNA using the Qiagen DNeasy Plant Mini Kit. The DNA material was shipped to the Biotechnology Resource Center Genomic Diversity Facility at
Cornell University to create genotyping-by-sequencing libraries made with the PstI enzyme (Elshire et al. 2011). At the same facility, the libraries were sequenced using the 100 bp paired-end approach on the Illumina HiSeq2000. After sequencing, we used the denovo_map.pl pipeline part of the STACKS software to identify informative SNPs (Catchen et al. 2013). We set 20 as the minimum coverage to create a stack and removed highly repetitive tags. We used the default values for the remaining parameters for SNP calling. At the termination of SNP calling we identified a total of ~20,000 markers.

**Composite interval mapping**

For each trait that we measured, we performed composite interval mapping (CIM) to identify the interval regions containing quantitative trait loci (QTL) using the R/QTL software (Broman et al. 2003; Broman and Sen 2009). First, we constructed a genetic map for *E. paniculata* (detailed in supplementary methods) and calculated conditional genotype probabilities for each cM in the genetic map assuming a genotyping error rate of 0.03. Further, we performed CIM using the imputation method. In some cases, we compared the interval scans using the expectation–maximization algorithm (EM) and extended Haley-Knott regression (EHK). For each LOD peak, we performed 1000 permutations to attain its genome-scan-adjusted *P*-value. We estimated the percent of variation in phenotype explained under a single QTL model using the formula $1 - 10^{-2 \text{LOD} / n}$, where *n* was the sample size. For a multiple QTL model, we used the "makeqtl" and "fitqtl" functions, using the imputation method for the latter function, in the R/QTL software to estimate the proportion of phenotypic variances explained by each QTL. We used the "addint" function to consider if there was any evidence for interactions among QTLs in a multiple QTL model. For the markers with the highest LOD scores underneath each LOD peak, we identified their corresponding effect sizes.
Identifying genes within QTLs

We generated a genomic assembly of *E. paniculata* using DNA read data generated for the parents and integrated that with the genetic map to identify the candidate genes within the QTLs identified for each trait. Details of the genomic assembly are found within the supplementary methods. We previously generated a 65.53 Mbp *de novo* transcriptome assembly (16416 contigs, $N_{50} = 2.2$ Kbp) from six selfing transcriptomes from Caribbean plants using the programs Velvet 1.2.08 (Zerbino and Birney 2008) and Oases 0.2.08 (Schulz et al. 2012) and predicted coding regions by BLAST searches to plant databases (Chapter 3). Further, we filtered paralogous loci by removing contigs that contained sites polymorphic across multiple selfing populations and those that had a genomic coverage greater than the sequencing depth. Here, we performed a custom nucleotide BLAST search (Altschul et al. 1990) to identify the scaffolds containing markers with LOD scores above the significance threshold in each QTL. We used these markers as the query, the genomic assembly as the database and extracted the top hit. Further, we performed another custom nucleotide BLAST search, with these scaffolds as the query and the transcriptome assembly as the database, to identify the genes found within each of these scaffolds.

Population genetic estimates

We estimated population genetic parameters for candidate genes underlying each QTL to investigate evidence for balancing selection, an expectation under the hypothesis that the *M locus* governing tristyly is maintained by frequency-dependent selection. For the candidate genes, we used a population genetic dataset for outcrossing previously reported in Chapter 3. Briefly, we used a scattered sampling approach to collect one plant per population from ten outcrossing populations from N.E. Brazil. We sequenced the transcriptomes, mapped the reads and called
SNPs in each plant using the aforementioned transcriptome reference and using a similar approach to this study. Here, using the SNP dataset we generated estimates of nucleotide diversity ($\Pi$) and Tajima's $D$ at synonymous sites using the program Polymorphorama (Andolfatto 2007; Haddrill et al. 2008). We also estimated population recombination rate ($\rho$) using the program DnaSP v5 (Librado and Rozas 2009). Along a given genomic interval, we selected the longest scaffold to visualize the distribution of genes and their population genetic estimates within the specified region. We generated scaffold-specific and genome-wide mean values for $\Pi$, Tajima's $D$ and $\rho$ across the 200 bootstrap replicates by randomly resampling across genes. We estimated genome-wide 95% CIs for each parameter estimate by excluding the top and bottom 2.5% of the bootstrap replicates and assessed if scaffold-specific means were significantly different from the genome-wide values by performing a two-tailed permutation test on the bootstrap replicates.

**Bulk segregant analysis**

To identify candidate transcripts putatively linked to the $M$ locus we performed bulk segregant analysis (BSA) on the $L'$ and $M'$ semi-homostyloous $F_2$ backcross progeny to identify differentially expressed genes. We collected a single floral bud from each $F_2$ progeny of approximately the same size. We pooled floral buds of 10 $F_2$ progeny of the same morph and extracted RNA using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich). We then pooled equal amounts of RNA from five extractions to generate six $L'$- and six $M'$-morph pools each containing RNA from ~50 $F_2$ progeny. The extracted RNA samples were used to make Illumina TruSeq RNA libraries that were sequenced using the 100 bp paired end protocol on two lanes of the Illumina HiSeq 2000 at the Génome Québec Innovation Centre at McGill University. Three $L'$- and three $M'$-morph pooled libraries were sequenced on each lane. Note that we previously
sequenced the floral bud transcriptomes of the L’ and M’ parents (see Chapter 3) and the raw
sequence data are available under accession number SRP049636 at the Sequence Read Archive

Candidate SNPs for style-length variation

To integrate our bulk segregant analyses with the GBS genetic map, we mapped the RNA-Seq
short reads from the F2 progeny, and the L’- and M-morph grandparents and genomic reads from
the F1 parents, to both the previously generated transcriptome assembly (Chapter 3), and the
genomic assembly generated in this study. The read mapping and variant calling approaches are
detailed in the supplementary methods. We identified candidate genes governing style morph by
assessing the SNP segregation in the bulked F2 progeny transcriptomes. We expected genes from
the transcriptome assembly would contain sites that were homozygous across all F2 L’-morph
progeny pools and heterozygous in all F2 M’-morph progeny pools. We only included: (1) sites
for which both genomic and transcriptome data indicated the maternal L’-morph parent (ssmm)
was homozygous, (2) sites for which genomic data indicated the paternal M’-morph parent
(ssMm) was heterozygous and (3) sites for which transcriptome data indicated the M’-morph
grandfather (ssMM) was homozygous for the alternate allele. Hereafter, these sites are termed
candidate SNPs. For the genes containing one or more candidate SNPs, we performed a custom
nucleotide BLAST search (Altschul et al. 1990) with the genes as the query and the genomic
assembly as the database, to identify the genomic regions containing these genes. For the
genomic regions matching these genes, we validated the variant calls for these candidate SNPs
by ensuring the genomic scaffold data also indicated that the mother and grandmother (L’-morph
parent - ssmm) were homozygous and the father (F1 M’-morph parent - ssMm) was heterozygous
for the candidate SNPs. We performed another custom nucleotide blast search, with these scaffolds as the query and the marker sequences from the genetic map as the database, to identify LGs in which these scaffolds were found. We also inferred the molecular function and expression intensity of candidate genes and these approaches are detailed in the supplementary methods.

Results

*Phenotypic variation of the L′ and M′ progeny*

There were 227 L′ and 235 M′ F₂ progeny in the backcross generation, which is not significantly different from the expected 1:1 ratio for a cross between *ssmm* x *ssMm* genotypes (Deviation from 1:1; *G*-test statistic: 0.201, *P*=0.647). On average, plants took 72.5 days (CI ± 2 day) after seed germination to flower. There was no significant difference in flowering time between the L′ and M′ progeny. L′ progeny were significantly taller than M′ progeny but had smaller flowers (Appendix IV - Figure 6S2). Two sample *t*-tests using unequal variances indicated that plant height and flower size were significantly different between the L′ and M′ progeny at the 1% threshold. There were significant differences in style length and anther position between the style morphs. Style lengths for the L′ and M′ F₂ progeny were 15.66 ± 0.10 mm and 11.65 ± 0.12 mm (mean ± CI), respectively (Figure 6.1), and this difference was significant (two-sample t-test: *P* < 0.001). The upper three L′ anthers (anthers 4-6) were close to each other and were 0.9-3 mm below the stigma. This range in stigma-anther separation reflected the occurrence of two distinguishable phenotypes, one in which anther 6 was very close or touching the stigma and one which was separated from the stigma. Thus the average value for anther 6 indicates a small separation. The two phenotypes could not be distinguished in the mapping study (see below).
The lower three L´ anthers (anthers 1-3) were 5-8 mm below the stigma. In the M´-morph the upper three anthers (anthers 4-6) were close to each other and were 4-5.5 mm above the stigma whereas M´ anthers 1-3 were ~3 mm below and above the stigma, respectively, with anther 2 very close to or touching the stigma. Principle component analyses of the 10 floral traits and plant height allowed assessments of possible trait associations within style morphs. The screeplot suggested two main dimensions for both style morphs (Appendix IV - Figure 6S3A,B). Style length and anther heights were not strongly associated with plant height or with flower width, breadth and length within each morph (Appendix IV - Figure 6S3B). The comparisons of style length and anther height differences within and between L´- and M´-morph progeny were not affected when each measure was first divided by the flower size, calculated as flower width x breadth x length or by plant height.

Figure 6.1. Mean anther heights of 227 L´ and 235 M´ F$_2$ Eichhornia paniculata progeny. We crossed a semi-homostyous M´ (ssMM) with a semi-homostyous L´ (ssmm) and backcrossed
the resulting F₁ M’ plant (ssMm) to the L’ parent to generate the F₂ population. We sequentially labeled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower with 6 being furthest away from the base. The bars represent 95% confidence intervals. The black triangles indicate the position of the stigma in each morph.

Genomic reference and genetic map

Figure 6.2. Genetic map of *Eichhornia paniculata*; the number of markers in each linkage group is indicated on the top of the diagram.

We generated a genomic assembly and genetic map for *E. paniculata* to assist with identifying the genetic regions associated with the *M* locus. The genomic assembly of parental genotypes, assisted by genome size estimation, indicated a large section of the genome contained repetitive content. The haploid genome size of *E. paniculata* based on flow cytometry was 0.61 Gbp. Assemblies with k-mer values > 40 had a large number of < 100 bp contigs resulting in a
fragmented assembly. For example, with a k-mer of 81 there were over a million contigs totaling 1.1 Gbp of assembled sequences. Repeat content analyses indicated that the majority of these small contigs matched known repetitive elements. With a k-mer of 31, there were 24,226 contigs totaling about 310 Mbp of sequence that were annotated as containing repetitive content and these contigs were excluded prior to scaffolding. Note that this k-mer value of 31 was used for the final assembly. After excluding these contigs, there were 40,320 scaffolds totaling 573 Mbp of assembled sequences with 4.6% of "N" bases (Appendix IV - Table 6S1). 40.3% and 0.9 % of scaffolds were greater than 10 Kbp and > 100 Kbp respectively with an $N_{50}$ 31.7 Kbp. Given the sequencing depth and genome size of *E. paniculata*, we expected a 25-33X coverage for most contigs in each individual. After mapping genomic reads from each individual back to the genomic reference, most scaffolds had coverage ranging between 5-35X but some had far greater coverage suggestive of collapsing of repetitive sequence in the assembly (Appendix IV - Figure 6S4). We removed high coverage scaffolds (> 40X) prior to further analyses as they are likely to contain poorly assembled and highly repetitive regions. We generated genetic markers along regions with low repeat content to construct the genetic map of *E. paniculata*. Using 1450 genetic markers, the genetic map had expected eight linkage groups, consistent with the number of chromosomes in *E. paniculata*, with 90-250 markers in each LG (Figure 6.2). The sizes of the LGs ranged from 10-50 cM with a total length of 255 cM. However, larger LGs, particularly groups 1, 2, 3 and 6, had long stretches containing few markers.

*Interval mapping of the M locus*

Joint assessments of the genotypes and phenotypes of the F2 progeny provided insight into the genetic regions associated with style length and anther height. CIM using the imputation method indicated there was a single peak on LG5 with LOD scores of > 35 shared by style length (Figure
6.3A) and the upper three anthers which distinguish the two style morphs (L´ and M´ anthers 4-6; Figure 6.3B). There were no other peaks shared between style length and anther heights and this shared peak on linkage group 5 was the only one evident for style length. Analyzing the style length peak indicated there were 47 markers located between 7.5 - 11.5cM (Figure 6.3A, Appendix IV - Table 6S2). The genotypes for the marker with the highest LOD score underneath this peak were strongly associated with the observed difference in style lengths between the L´ and M´ progeny (Appendix IV - Figure 6S5). Finally, CIM using EHK and EM approaches also yielded similar results (Appendix IV - Figure 6S6).

Figure 6.3. Composite interval mapping of: A) style length variation and B) variation in the anthers of the upper-level stamens in 227 L´ and 235 M´ F2 backcross progeny of *Eichhornia paniculata*. We crossed a semi-homostylos M´ (ssMM) with a semi-homostylos L´ (ssmm) and
backcrossed the resulting F$_1$ M’ plant ($ssMm$) to the L’ parent to generate the F$_2$ population. Shown are map distances and LOD scores along all linkage groups. The upper-level anthers in the L’ and M’ correspond to mid- and long-level stamens, respectively. The horizontal dashed lines indicate the threshold for significance (LOD > 3) at the 0.1% level based on 1000 permutations.

Figure 6.4. Patterns of polymorphism and Tajima's $D$ for *Eichhornia paniculata* genes found within linkage group 5. Illustrated are estimates of nucleotide diversity ($\pi$) for genes: (A) within the interval associated with style-morph variation and (B) five cM away from the interval associated with style-morph variation and (C) estimates of Tajima's $D$ within the interval associated with style-morph variation. Each interval break denotes a different genomic scaffold.
We chose the longest scaffold among those that matched a given interval position. Each point corresponds to the $\pi$ or Tajima's $D$ of a gene based on samples from eight outcrossing tristyloous populations from N.E. Brazil. The shaded boxes indicate the confidence interval of the genome-wide average for each parameter estimate generated from 200 bootstrap replicates randomly resampling across genes.

Comparison of genetic mapping results to the genomic assembly revealed the genomic context for markers within an identified genetic interval. The markers matched 35 non-overlapping scaffolds of various sizes totaling 2.39Mbp of genomic sequence (Appendix IV - Table 6S2). The scaffolds found within this interval contained 252 genes and the position near 8.7cM contained the greatest density of genes. We identified a cluster of genes with relatively high nucleotide diversity in outcrossing populations near the highest LOD score at 8.9cM in LG5 (Figure 6.4A). Although scaffold specific means were not significantly different from the genome-wide average, similar patterns were not observed 5cM either side of the identified interval (Figure 6.4B). To assess whether this elevated gene density and nucleotide diversity was simply a function of being near the centre of the LG, we identified scaffolds matching to the centre of LGs 3 and 7. Both scaffolds were ~100Kbp, comparable in size to the longest scaffold size on LG5, but only contained 4-5 genes of one Kbp length. Also, these few genes along LGs 3 and 7 exhibited low $\Pi$ in outcrossing populations (Appendix IV - Figure 6S7) suggesting the patterns we saw on LG 5 were unique. For the remaining LGs we were unable to identify scaffolds of sufficient length.

Population genomic scans along the 8.5-9.7cM interval on LG5 identified unique patterns of nucleotide diversity consistent with the action of balancing selection. The confidence interval
of genome-wide estimate of Tajima's $D$ overlapped 0 (Figure 6.4C). In contrast, a number of genes found near the interval between 8.5-8.7cM on LG5 had positive Tajima's $D$ and this value was significantly greater than the genome-wide average at the 0.1% threshold when assessed using a permutation test. $\rho$ estimates for genes along the interval between 8.5-8.7cM on LG5 were significantly lower than the genome-wide average at the 0.1% threshold (Appendix IV - Figure 6S8). Finally, while there was no evidence that any of the Arabidopsis orthologs for the genes identified within the interval were enriched for a particular molecular function, six genes were characterized as being involved in reproductive development.

**Transcriptomes of $L'$ and $M'$ pools**

We also performed BSA using RNA from pooled F$_2$ progeny to identify candidate genes governing style morph. 340 of 16147 contigs in the transcriptome assembly had one or more sites that were homozygous across all L' pools and heterozygous across all M' pools, as expected if such SNPs were linked to the style-morph locus. Of this set, only 48 and 8 contigs had greater than five or ten sites respectively showing the previously described patterns of variants. There were no contigs that contained sites homozygous in all M' pools and heterozygous in L' pools, suggesting a low rate of false positives. However, we were unable to identify any genes with morph-limited gene expression, or those with at least a 10-fold difference in expression between the two progeny morph pools.

BSA was generally congruent with our results from linkage mapping. Approximately 10% of the 340 BSA genes ($n=32$) containing style morph differentiating candidate SNPs matched markers along the 8.7-11.3 cM interval on LG5 and contained style morph differentiating candidate SNPs (Appendix IV - Table 6S2). Custom BLAST searches indicated that the other LGs contained far fewer candidate genes (Appendix IV - Table 6S3). However, we
were unable to ascertain the map position and location for the majority of these candidate genes. Of the 334 genes with the style morph differentiating candidate SNPs, there was evidence to indicate significant functional enrichment in post-embryonic development among the *Arabidopsis* orthologs (Appendix IV - Figure 6S9). 21 genes were involved in this function and one gene was found within the 8.5-9.7cM interval on LG5.

*Regions governing anther height variation in L’ and M’ variants*

![Figure 6.5. LOD scores on linkage groups 2 and 4 following composite interval mapping of anther heights of L’ and M’ variants in F2 progeny of *Eichhornia paniculata*. Shown are LOD](image)
scores for: A) the lowest three and B) the upper three L´ anthers along linkage group 2 and LOD scores for C) the lowest three and D) the upper three M´ anthers along LG 4. We sequentially labeled anthers from 1 to 6 for each morph, corresponding to their positions relative to the base of the flower with 6 being furthest away from the base. The horizontal dash lines indicate the threshold for significance at the 0.1% level based on 1000 permutations.

Anther height variation in the L´ and M´ semi-homostylyous variants was governed by QTLs separate from the region governing style length, likely reflecting anther height modifier loci associated with the transition to selfing. In each semi-homostyle there were peaks that were unique to anthers with the smallest degree of physical separation from the stigma, corresponding to the modified anthers causing autonomous self-pollination. For L´, the upper three anther levels (anthers 4-6) had a significant peak on LG2 (Figure 6.5A, B). The interval spanned ~8cM and contained 54 markers. The markers with the highest LOD score underneath this peak were of genotypes AA, BB and AB in the L´ parent, M´ parent and F1 M´, respectively. An AA genotype at this marker was associated with ~0.6 mm increase in anther height when compared to AB genotypes (Appendix IV - Figure 6S10). This QTL explained 6-8.5% of the variation in the height of upper-level anthers (Appendix IV - Table 6S4). Very few genes were identified in the scaffolds containing these markers. This interval contained 39 non-overlapping scaffolds totaling a size of 1.68 Mbp of genomic sequence (Appendix IV - Table 6S5). The genomic scaffolds contained a total of 197 genes.

Similarly, the M´ variant had a QTL not found in the L´ variant. The second and third lowest-level anthers (M´ anthers 2-3), but not the other anthers, had a peak on LG4 (Figure 6.5C, D). The identified interval was at least 8cM but LOD scores did not fall below the significance
threshold at the terminal markers. Further, the 21 markers underneath this peak were widely distributed. The marker with the highest LOD score underneath this peak were of genotypes BB, AA and AB in the L´ parent, M parent and F1 M´, respectively. An AA genotype at this marker was associated with ~1.5mm (Appendix IV - Figure 6S11A) and ~0.5 mm (Appendix IV - Figure 6S11B) increase in the height of M´ anthers 2 and 3, respectively, when compared to AB genotypes. Genotypic differentiation was not accompanied by differences in the heights of L´ anthers 2 and 3 (Appendix IV - Figure 6S11C, D). This QTL explained 18.1% and 8.6% of the variation in the heights of M´ anthers 2 and 3, respectively. This interval matched 31 non-overlapping scaffolds totaling a size of 1.05 Mbp of genomic sequence (Appendix IV - Table 6S6). Different markers along 7-11 cM of the interval matched the same genomic scaffolds and matched a total of 96 genes. CIM using EHK and EM approaches yielded similar results for the uppermost anther level in the L´ progeny (anther 6, Appendix IV - Figure 6S12) and the second lowest anther level in the M´ progeny (anther 2, Appendix IV - Figure 6S13).

L´- and M´-morphs possessed other unique and shared QTLs governing anther-height variation. The multiple QTLs governing anther heights explained 60-70% of the variation in L´-morph progeny and 8-42% of the trait variation in M´-progeny. There was a lack of evidence to indicate any epistatic interactions among the identified QTLs in each morph. CIM indicated that all anther levels in the L´-morph and the lowest two anthers in the M- morph (M´ anthers 1-2) had a peak on LG3 (Figure 6.6). This interval spanned 8.3 cM and there were 178 markers underneath this peak (Appendix IV - Table 6S7). Under a completely additive multiple QTL model, this QTL explained 42-55% of variation in the height of the L´-morph anthers (Appendix IV - Table 6S4) and the lowest M´-morph anther level (M´ anther 1) but only 14.5% of variation in the second lowest M- morph anther level (M´ anther 2). Note that the effect size of the change
in anther height and the percentage of trait variation explained due to this marker for M’ anther 2 was smaller than the level due to the aforementioned marker on LG4 (Appendix IV - Figure 6S11B). This interval matched 685 Kbp of non-overlapping genomic sequence containing 256 genes. All anther levels in L’-morph progeny also had peaks in LGs 1 and 6, but the LOD scores for L’-morph anthers 1 and 4 were below the significance threshold for the QTL on LG6 (Appendix IV - Figure 6S14). Each QTL explained 6-8% of the variation in the height of anthers (Appendix IV - Table 6S4). The interval of LG1 spanned ~7 cM with two markers underneath the peak and the interval of LG6 spanned ~4 cM with a single marker underneath the peak. The upper three anther levels in the M- morph progeny (M’ anthers 4-6) did not show any significant peaks across the LGs.

Figure 6.6. LOD scores on linkage group 3 of *Eichhornia paniculata* following composite interval mapping of anther heights of the L’ and M’ semi-homostyloous morphs in F2 progeny.
Shown are LOD scores for: A) the lowest three L’ anthers, B) the upper three L’ anthers, C) the lowest three M’ anthers and D) the upper three M’ anthers. We sequentially labeled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower with 6 being furthest away from the base. The horizontal dashed lines indicate the threshold for significance at the 0.1% level based on 1000 permutations.

Discussion

Our study investigated the genetic architecture of variation in style length and anther height characterizing the tristylos syndrome of *E. paniculata* and also its breakdown to selfing. We found that a single large genetic interval near the center of linkage group 5 governed style length and the stamen level that differentiates the L- and M-morphs. This interval also showed signatures of balancing selection and was associated with regions of low recombination providing evidence that we have broadly mapped the *M* locus in this species. We also found 334 genes with contrasting patterns of SNPs in the L’ and M’ F2 progeny, although we found no evidence for morph-specific gene expression. In both style morphs, modified stamens causing increased self-fertilization contained unique QTLs that were not associated with the *M* locus occurring elsewhere in the genome. Significantly, different QTLs were involved with stamen modifications in the two semi-homostylous variants consistent with their independent origins. Below we discuss insights gained from mapping style length and anther height in *E. paniculata* on the genetic architecture of tristyly and the breakdown of this outcrossing polymorphism to selfing.
Broad mapping of the style morph locus

Our mapping study indicated that the $M$ locus responsible for style-length differentiation between the L- and M-morph localized to a single linkage group. This result is similar to the patterns observed from genetic mapping studies of distyloous *Primula* which also identified that the style length and anther heights mapped to the same linkage group (Yoshida et al. 2011). The identified size of the region containing the $M$ locus was large and this limited our ability to identify candidate genes associated with the tristyloous syndrome. Our genotyping approach was specifically chosen to maximize density along the regions with low repeat content (Elshire et al. 2011). The genomic assembly and genome size estimates suggested that a large section of the genome consists of highly repetitive regions. Long stretches of the genetic map without any markers could therefore indicate regions with highly repetitive content. The genetic distance estimate of the interval identified as governing the $M$ locus suggested that it could contain $\sim 11\text{Mbp}$ of sequence or more. This interval was larger than the distance presumed to be separating the $M$ and $S$ loci in this species (Chapter 5), suggesting that the $S$ locus could also lie in this interval. We mapped $\sim 2.4\text{Mbp}$ of non-overlapping genomic sequence to this region. At least 252 genes were expressed along this interval and the region near the 8.9cM region of LG5 had a particularly high gene density. However, it is unclear if any of these genes were causally associated with the tristyloous syndrome as most lacked strong evidence for morph-biased or morph-specific expression. Even so, the interval on LG5 contained the greatest number of genes with candidate SNPs in our transcriptome BSA analysis compared to all other LGs. This serves as an independent genotyping validation that the identified genetic region contains the style morph loci in *E. paniculata*. 21 genes in this region were significantly enriched for functions
associated with reproductive and floral development and had morph-biased expression and in future may be useful to fine map the $M$ locus.

Even though the interval containing the $M$ locus was broad, our analysis of outcrossing populations from Brazil identified genes within this interval that had unique population genetic signatures compared to the remainder of the genome. Genes matching to markers along the 8.7-9.7cM portion of LG5 had higher levels of polymorphism, elevated Tajima's $D$ and lower $\rho$ suggesting linkage to the $M$ locus. These patterns are consistent with the operation of balancing selection associated with the maintenance of tristyly in outcrossing populations. However, a few genes along 8.7-9.7cM portion of LG5 remained polymorphic in the selfing populations. While the extent of polymorphism observed across each interval was limited by the longest scaffold identified, our conclusions are unlikely to be altered because the size of scaffolds compared was similar. As this region is located at the centre of LG5 and contains higher marker density compared to surrounding regions, it is possible that the signal of elevated Tajima's $D$ might be, in part, due to an association with centromeric regions (Carneiro et al. 2009). Elevated polymorphism is generally unexpected in centromeric regions, but in the outcrossing plant *Arabidopsis lyrata*, diversity was not reduced in pericentromeric regions although gene density was low (Wright et al. 2006; Kawabe et al. 2008). Similarly, Bensasson (2011) reported high levels of polymorphism in the centromeres in multiple yeast populations even though there was little evidence for crossing-over or gene conversion within centromeres. Results from both taxa suggest selection and mutational forces can countervail this expectation. Similar patterns of marker density and polymorphism were not observed near the centre of other LGs suggesting the pattern was unique to LG5. Patterns of elevated polymorphism and Tajima's $D$ and associations to low recombining regions are characteristic of the region containing the $M$ locus and these
broad mapping results and our list of candidate markers and genes should allow more explicit future characterization of the *M* locus.

*Genetic architecture of stamen modifications promoting self-pollination*

The stamens promoting self-pollination in each semi-homostyloous variant had QTLs unlinked to the *M* locus and these QTLs were also not shared between the variants. This suggests that the evolution of the two semi-homostyles occurred independently and in neither case were stamen modifications the result of recombination within the *M* locus. The upper three anthers in L’ and the second and third shortest anthers in M’, corresponding to the modified stamen levels in each morph, had QTLs not found in the remaining stamen levels of each variant. These QTLs had the largest influence on anther position, and with anthers exhibiting the lowest degree of physical separation from the stigma in each semi-homostyle. The QTL on LG4 for M’ anthers 2-3 was very large and our genetic map may not have had sufficient marker density to capture the entire size of this QTL. The large size of the identified regions and low marker density within each interval made it difficult to identify candidate genes, but several were identified that were differentially expressed between the semi-homostyloous morphs. The genotype associated with stamen elongation of these modified anthers in the L’semi-homostyle was the same as that of the L’ parent and the allele found in the homozygous genotype associated with stamen elongation of these modified anthers in M’progeny was fixed in the M’ parent; both patterns confirmed that the modifier alleles associated with stamen elongation in each morph were also found in the parents. Interestingly, the allele associated with modification of the short-level stamens in the M’semi-homostyle did not have a similar effect in modifying the short-level stamens of the L’ semi-homostyloous progeny suggesting the effect of this allele is morph-limited in its expression. Lloyd and Webb (1992b) proposed that some genes associated with heterostyly only altered phenotypic
traits in one floral morph but not the other and thus exhibited morph-limited expression. Based on their crossing studies of *E. paniculata*, Fenster and Barrett (1994) found no evidence to indicate that short-level stamens of L-morph plants in segregating families containing M´morph plants exhibited modified stamens also suggesting morph-limited gene expression is associated with stamen modifications in the M´semi-homostyle.

In addition to the one QTL shared with style length, anther heights in the L´ and M´ semi-homostyles were governed by different genetic regions, some of which were shared between the two selfing variants and others that were unique to each morph. These QTLs are significant because they provide insight into the genetic architecture of evolutionary transitions from outcrossing to selfing in *E. paniculata*. Anthers that were positioned below the stigma in each style morph (L´ anthers 1-6 and M´ anthers 1-2) shared the same QTL on LG3. This locus exhibited the most influence on anther-height variation compared to the other identified QTLs. It is probable that this region contains the modifier genes that are non-allelic to the *S* and *M* loci governing stamen elongation and causing the loss of herkogamy, as predicted from crossing studies (Fenster and Barrett 1994; Barrett et al. 2009). The lack of evidence that the QTL on LG3 also governs the upper three anthers (long-level stamens) that are only found in M´ plants, could be because any further elongation of this stamen level has no functional significance since any modification would be unlikely to promote self-pollination. All stamen levels in the L´ plants had additional QTLs modifying anther position on LGs 1 and 6. In contrast, there were only one or two significant QTLs governing most of the anther-height variation in M´ plants. Our results conform to expectations from crossing studies indicating that multiple genes are associated with the evolution of L´ semi-homostyles in Central America whereas only a few genes govern the transition to selfing in M´ semi-homostyles in the Caribbean (Fenster and Barrett 1994; Barrett et
al. 2009). Fenster and Barrett (1994) reported that the major genes associated with M’ semi-homostyle evolution were recessive. It is likely that we did not detect QTLs associated with recessive modifiers due to our backcrossing design. More generally, our mapping results indicate that the loss of herkogamy in the tristylos L- and M-morphs of *E. paniculata* resulting in the evolution of selfing is governed by genetic regions separate from the *M* locus. This provides evidence against a recombinational origin for homostyles, as occurs in several distylos species (Dowrick 1956; Shore and Barrett 1985; reviewed in Barrett and Shore 2008). Rather, it seems more probable that the evolution of selfing in tristylos species may be more commonly achieved through gradual changes to stamen position governed by modifiers elsewhere in the genome.

**Appendix IV**

**Supporting information for Chapter 6**
Supplementary methods

Genetic map construction

We generated a genetic map to assess where the sequenced genetic markers were positioned along the *E. paniculata* genome. We used R/QTL software (Broman et al. 2003) to construct a genetic map using the approach described by Karl Broman (www.rqtl.org/tutorials/geneticmaps.pdf). We first removed markers where more than 90% of the F2 progeny were not typed and omitted progeny that were not typed for 90% or more of the markers. Such markers and progeny might introduce more noise during genetic map construction than be informative. We only retained SNPs conforming to expected segregation patterns from the genetic cross where the mother and grandmother (*ssmm*) were homozygous, the father (*ssMm*) was heterozygous, and the grandfather (*ssMM*) was homozygous for the alternate allele, using custom Perl scripts. However, this step removed only 1-2% of SNPs. We also removed markers that showed significantly distorted segregation patterns compared to those expected from a backcross assessing significance at the 5% level after performing a Bonferroni correction for multiple tests. In this process, we removed ~50 F2 progeny and ~18,000 markers. We formed linkage groups (hereafter LGs) assuming the recombination fraction between markers was 0.35, with a minimum LOD score of 20, as this resulted in 20X lower values for pairwise recombination fractions for markers within the same LGs compared to markers from different LGs. For each LG, we used the 'checkAlleles' and 'ripple' functions to identify the marker order that resulted in the highest likelihood. Once the markers were ordered, we re-estimated the genetic map assuming a genotyping error rate of 0.03 to get more accurate map distances. To visualise the genetic map, we used the 'pullRF' and 'pullMap' functions to retrieve the pairwise
recombination fractions between markers, maps distances and marker order on each LG, respectively. The final map was constructed using 1,450 markers from 462 progeny.

Genomic sequencing of parental genotypes

We sequenced the genomes of the parental L´ and M´ semi-homostyles to identify the genomic context of QTLs identified for measured traits. To obtain the optimal parameters for the genomic assembly, we estimated the genome size of *E. paniculata*. We collected leaf tissue from the parental plants used for genome size estimation using flow cytometry conducted at the Plant Cytometry Services in the Netherlands (http://www.plantcytometry.nl/). Two measurements were made and we report the average. We then extracted DNA from floral bud tissue using the Qiagen DNeasy Plant Mini Kit. Illumina TruSeq DNA libraries with 400 bp insert sizes were prepared from both samples and an additional library with a 5 Kbp insert size was prepared from DNA extracted from the L´-morph parent. Each library was sequenced on individual lanes using the 100 bp paired end protocol on Illumina HiSeq 2000 at the McGill University and Génome Québec Innovation Centre. After sequencing, we removed reads < 50 bp and reads that had more than 10% of "N" bases using a custom Perl script. We used SOAPdenovo v2.04-r240 (Luo et al. 2012) to assemble the genomes of the L´ and M´ parental genotypes. We subsampled the sequenced reads from the 5 Kbp insert size to a tenth of its original value prior to assembly, as using all reads would have been a memory intense process. We allowed for a maximum read length of 155 bp and assigned a higher rank for the libraries with 400 bp insert sizes than those with 5 Kbp insert sizes for scaffolding. We performed assemblies under a range of k-mer values from 30 to 90. We annotated and characterised repetitive regions with RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015)
Molecular functional characterisation

We inferred the function of candidate genes using a comparative genomics approach to assess evidence as to whether the genes may be functionally associated with the tristyloous syndrome. We performed custom translated nucleotide BLAST searches with *E. paniculata* genes as the query against the *Arabidopsis* protein database obtained from The Arabidopsis Information Resource (TAIR) (Lamesch et al. 2012). We performed a singular enrichment analysis using the genes expressed in each tissue as the query and the TAIR10 *Arabidopsis* gene model as the reference using the agriGO toolkit and database (Du et al. 2010). We found significant gene ontology terms were found after performing hypergeometric tests with the Yekutieli multiple test correction at a significance level of 0.05.

Read mapping and variant calling in sequenced transcriptomes

To integrate our bulk segregant analyses with the GBS genetic map, we mapped the RNA-Seq short reads from the F₂ progeny, and the L´ and M´grandparents and genomic reads from the F₁ parents, to both the previously generated transcriptome assembly and the genomic assembly generated in this study. First, we mapped reads with Burrows-Wheeler Aligner v0.7.8-r455 (Li and Durbin 2009) using default parameters and further used the "Sampe" command to combine the paired end read mapping results. Then, we used the Stampy 1.0.23 software (Lunter and Goodson 2011) with default parameters to map more divergent reads and identify insertions and deletions (indels). We processed the read mapping output using the SamFormatConverter, ReorderSam, AddOrReplaceReadGroups and BuildBamIndex programs part of the Picard tools package 1.124 using default settings (http://picard.sourceforge.net, last accessed April 18,
We also used the MarkDuplicates program, also part of the Picard tools package 1.124, to identify and tag duplicate reads. As erroneous mismatches might have occurred near indels, we used the RealignerTargetCreator and IndelRealigner programs, part of the Genome Analysis Toolkit (GATK) 3.3-0-g37228af (DePristo et al. 2011), with default parameters to realign sequences within ~3 kb of an indel. We estimated the number of reads mapping to each contig using HTSeq-count 0.6.1p1 (Anders et al. 2014) under the union mode. We ran the UnifiedGenotyper program (also part of GATK 3.3-0-g37228af; DePristo et al. 2011) to call invariants sites, SNPs and indels from all samples using default parameters with the BadCigar read filter, which uses a Bayesian model. From the UnifiedGenotyper output, we retained sites for which the Phred scaled quality score was > 60 and depth in each individual sequenced was > 20. We retained a SNP if the Phred scaled genotype quality for all samples was > 60 and excluded those from 5 bp on either side of an indel where spurious SNP calls are likely to be made. In addition, we mapped the genomic reads for the F1 parents to the genomic reference constructed in this study using the aforementioned approach and called variants.
Supplementary figures

Figure 6S1. Phenotypic measurements of the flowers of the L´ and M´semi-homostyles of *Eichhornia paniculata*.

Figure 6S2. Plant height and flower size of semi-homostylyous morphs in the backcross F2 of *Eichhornia paniculata* progeny. Error bars represent 95% confidence intervals. Based on 227 and 235 plants of the L´ and M´ semi-homostyles.
Figure 6S3. Principle component analysis of correlations among plant height, flower size, style length and anther height of 227 and 235 L’ and M’ F2 progeny, respectively, of *Eichhornia paniculata*. Shown are: A) screeplot and B) plots of loadings along the first two principle component axes for L’ progeny. Also shown are: C) screeplot and D) plots of loadings along the first two principle component axes for M’ progeny.

Figure 6S4. Number of genomic reads that mapped to the genomic reference assembly for the L’ (ssmm) and M’ (ssMm) parental genotypes of *Eichhornia paniculata*. 
Figure 6S5. Effect plot for the marker with the highest LOD score on linkage group 5 identified during interval mapping of style length in F2 progeny of *Eichhornia paniculata*.

Figure 6S6. Comparison of composite interval mapping of the style length of L′ and M′ semi-homostyles in F2 progeny of *Eichhornia paniculata* using the imputation method (IMP), expectation-maximization algorithm (EM) and extended Haley-Knott regression (EHK). The horizontal dash lines indicate the threshold for significance (LOD > 3) at the 0.1% level based on 1000 permutations.
Figure 6S7. Average pairwise differences per nucleotide site at genes along the centre of linkage groups 3 and 7 in *Eichhornia paniculata*. We chose the longest scaffold among those that matched a given interval position. Each point corresponds to the $\pi$ of a gene based on a scattered sample of eight outcrossing populations from N.E. Brazil. The dash line indicates the genome-wide average estimate of nucleotide diversity ($\pi$).

Figure 6S8. Population recombination rate estimates for genes found along the genomic interval on linkage group 5 of *Eichhornia paniculata*. We chose the longest scaffold among those that matched a given interval position. Each interval break denotes a different genomic scaffold. Each point corresponds to the recombination rate of a polymorphic gene based on a scattered sample of eight outcrossing *Eichhornia paniculata* populations from N.E. Brazil. The dash line indicates the genome-wide average recombination rate.
Figure 6S9. Functional enrichment of *Eichhornia paniculata* genes with candidate SNPs suggesting linkage to the style-morph locus. We performed singular enrichment analysis with the *Arabidopsis thaliana* orthologs of the *Eichhornia paniculata* genes using the TAIR10 Arabidopsis genemodel as the background and categorised them according to biological function. We found significant GO terms by performing hypergeometric tests with the Yekutieli multiple-test correction at a significance level of 0.05 indicated by the yellow boxes.
Figure 6S10. Effect plots for the marker with the highest LOD score on linkage group 2 identified during interval mapping of anther height for the upper three anther levels: A) 4, B) 5, and C) 6, in the L’ semi-homostyle of *Eichhornia paniculata* in F$_2$ progeny. We sequentially labelled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower, with 6 being furthest away from the base.
Figure 6S11. Effect plots for the marker with the highest LOD score on linkage group 4 identified during interval mapping of the anther height for modified anther levels: A) 2, B) 3, in the M’-morph and the anther height: A) 2, B) 3, in the L’ semi-homostyle of Eichhornia paniculata in F₂ progeny. We sequentially labelled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower, with 6 being furthest away from the base.
Figure 6S12. Comparison of composite interval mapping of anther 6 of the L’ semi-homostyle of *Eichhornia paniculata* in F$_2$ progeny using the imputation method (IMP), expectation-maximization algorithm (EM) and extended Haley-Knott regression (EHK). The horizontal dash lines indicate the threshold for significance (LOD > 3.2) at the 0.1% level based on 1000 permutations.
Figure 6S13. Comparison of composite interval mapping of the second lowest anther of the M’ semi-homostyle of *Eichhornia paniculata* in F$_2$ progeny using the imputation method (IMP), expectation-maximization algorithm (EM) and extended Haley-Knott regression (EHK). The horizontal dash lines indicate the threshold for significance (LOD > 3.05) at the 0.1% level based on 1000 permutations.
Figure 6S14. Composite interval mapping of anther heights of the L’ semi-homostyle of *Eichhornia paniculata* in F2 progeny. Shown are LOD scores for: A) lowest three anther levels along linkage group 1 and B) linkage group 6 and C) upper three anther levels along linkage group 1 and D) linkage group 6. We sequentially labelled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower with 6 being furthest away from the base. The horizontal dash lines indicate the threshold for significance (LOD > 3.2) at the 0.1% level based on 1000 permutations.
**Supplementary tables**

Table 6S1. Summary metrics of genomic assembly for *Eichhornia paniculata*.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Contigs</th>
<th>Scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number</td>
<td>65,688</td>
<td>40,320</td>
</tr>
<tr>
<td>Total size</td>
<td>546 Mbp</td>
<td>573 Mbp</td>
</tr>
<tr>
<td>Longest</td>
<td>226 Kbp</td>
<td>249 Kbp</td>
</tr>
<tr>
<td>Shortest</td>
<td>1 Kbp</td>
<td>1 Kbp</td>
</tr>
<tr>
<td>Number &gt; 10 Kbp</td>
<td>17,448 (26.6%)</td>
<td>16,250 (40.3%)</td>
</tr>
<tr>
<td>Number &gt; 100 Kbp</td>
<td>17 (0.02%)</td>
<td>354 (0.9%)</td>
</tr>
<tr>
<td>Mean size</td>
<td>8.3 Kbp</td>
<td>14.2 Kbp</td>
</tr>
<tr>
<td>Median size</td>
<td>4 Kbp</td>
<td>6.5 Kbp</td>
</tr>
<tr>
<td>N_{50} length</td>
<td>16.7 Kbp</td>
<td>31.7 Kbp</td>
</tr>
<tr>
<td>Percent A</td>
<td>32.6</td>
<td>31.1</td>
</tr>
<tr>
<td>Percent C</td>
<td>17.4</td>
<td>16.6</td>
</tr>
<tr>
<td>Percent G</td>
<td>17.4</td>
<td>16.6</td>
</tr>
<tr>
<td>Percent T</td>
<td>32.6</td>
<td>31.1</td>
</tr>
</tbody>
</table>
Table 6S2. Number of markers and genes within the interval on linkage group 5 that had high LOD scores following composite interval mapping of style length, and number of genes with candidate SNPs suggesting linkage to the style-morph locus in that interval in F₂ *Eichhornia paniculata* progeny.

<table>
<thead>
<tr>
<th>Location (cM)</th>
<th>Number of markers</th>
<th>Number of scaffolds</th>
<th>Total size of genomic sequence (kbp)</th>
<th>Number of genes</th>
<th>Number of genes with candidate SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.7</td>
<td>1</td>
<td>1</td>
<td>79</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>8.7</td>
<td>5</td>
<td>1</td>
<td>81</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>8.9</td>
<td>33</td>
<td>27</td>
<td>1728</td>
<td>118</td>
<td>17</td>
</tr>
<tr>
<td>9.5</td>
<td>4</td>
<td>2</td>
<td>119</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>9.7</td>
<td>2</td>
<td>1</td>
<td>107</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>10.2</td>
<td>1</td>
<td>2</td>
<td>132</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>148</td>
<td>25</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 6S3. Number of genes with candidate SNPs suggesting linkage to the style morph locus in each linkage group in *Eichhornia paniculata* F$_2$ progeny.

<table>
<thead>
<tr>
<th>Linkage group</th>
<th>Number of genes with candidate SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6S4. Percent of variation in height of anthers in the L’ semi-homostyle explained by each identified QTL in *Eichhornia paniculata* F$_2$ progeny.

<table>
<thead>
<tr>
<th>Anther$^a$</th>
<th>QTL location (linkage group)</th>
<th>Percent of variation in phenotype explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>8.4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>53.3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9.5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6.4</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>56.9</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6.2</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>9.2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8.4</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>54.2</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>8.1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>6.8</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>53.3</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>8.1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>7.1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>6.2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>47.3</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>

*We sequentially labelled anthers from 1 to 6 for each morph corresponding to their positions relative to the base of the flower with 6 being furthest away from the base.*
Table 6S5. Number of markers and genes within the interval on linkage group 2 that had high LOD scores following composite interval mapping of anthers in the L’ semi-homostyle of *Eichhornia paniculata* F₂ progeny.

<table>
<thead>
<tr>
<th>Location (cM)</th>
<th>Number of markers</th>
<th>Number of scaffolds</th>
<th>Total size of genomic sequence (kbp)</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.9</td>
<td>3</td>
<td>1</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>40.9</td>
<td>2</td>
<td>1</td>
<td>62</td>
<td>3</td>
</tr>
<tr>
<td>41.9</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>42.5</td>
<td>1</td>
<td>1</td>
<td>65</td>
<td>2</td>
</tr>
<tr>
<td>44.4</td>
<td>15</td>
<td>10</td>
<td>527</td>
<td>62</td>
</tr>
<tr>
<td>45.7</td>
<td>17</td>
<td>14</td>
<td>315</td>
<td>47</td>
</tr>
<tr>
<td>46.5</td>
<td>15</td>
<td>11</td>
<td>630</td>
<td>72</td>
</tr>
</tbody>
</table>
Table 6S6. Number of markers and genes found within the interval on linkage group 4 that had high LOD scores following composite interval mapping of the M´ semi-homostyle anthers in *Eichhornia paniculata* F₂ progeny.

<table>
<thead>
<tr>
<th>Location (cM)</th>
<th>Number of markers</th>
<th>Number of scaffolds</th>
<th>Total size of genomic sequence (kbp)</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.5</td>
<td>14</td>
<td>18</td>
<td>790</td>
<td>65</td>
</tr>
<tr>
<td>12.2</td>
<td>4</td>
<td>11</td>
<td>108</td>
<td>5</td>
</tr>
<tr>
<td>15.5</td>
<td>3</td>
<td>2</td>
<td>156</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 6S7. Number of markers within the interval on linkage group 3 that had high LOD scores following composite interval mapping of anther height in *Eichhornia paniculata* F₂ progeny.

<table>
<thead>
<tr>
<th>Location (cM)</th>
<th>Number of markers</th>
<th>Number of scaffolds</th>
<th>Total size of genomic sequence (Kbp)</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.7</td>
<td>37</td>
<td>1</td>
<td>222</td>
<td>96</td>
</tr>
<tr>
<td>7.9</td>
<td>105</td>
<td>1</td>
<td>223</td>
<td>98</td>
</tr>
<tr>
<td>10.1</td>
<td>10</td>
<td>1</td>
<td>111</td>
<td>46</td>
</tr>
<tr>
<td>11.6</td>
<td>11</td>
<td>1</td>
<td>129</td>
<td>12</td>
</tr>
</tbody>
</table>
Chapter 7

Concluding discussion

In this thesis, I investigated the genetic architecture of tristyly and the genomic consequences of the breakdown of this floral polymorphism from outcrossing to selfing in the annual, neotropical plant *Eichhornia paniculata* (Pontederiaceae). I found that linked selection can result in the rapid loss of genetic diversity following transitions to selfing, and that the resulting reduction in the effective population size ($N_e$) of selfers is associated with a genome-wide accumulation of weakly deleterious mutations. At the same time, the increased homozygosity of selfing genomes exposes strongly deleterious mutations to selection. A small fraction of alleles in selfing genotypes of *E. paniculata* were down-regulated following the shift to self-fertilization, but the underlying mechanism for this remains unknown. I also found that the $S$ and $M$ loci governing tristyly were strongly linked in *E. paniculata* and I broadly mapped the $M$ locus and provided evidence for balancing selection consistent with frequency-dependent selection maintaining the polymorphism. I also discovered that separate modifier regions unlinked to the $M$ locus were involved in the evolutionary breakdown of tristyly through their influence in reducing herkogamy in flowers causing autonomous self-pollination. In the following sections, I summarize the major finding of each chapter in my thesis and suggest future work that could be profitably undertaken.

Chapter 2 - *The loss of genetic diversity following the transition from outcrossing to selfing.*

Here, I investigated how demographic and genetic processes accompanying the transition from outcrossing to selfing shape patterns of genetic diversity. Principally, I investigated how quickly linked selection reduces genome-wide diversity following the evolution of predominant selfing. A key finding, from multi-locus simulations and comparative molecular data from related
outcrossing and selfing lineages, was that linked selection is very strong even at the very early stage of the transition to selfing. This complicates the proposal (see Schoen et al 1996) that molecular population genetic data can be used to disentangle the demographic context in which selfing evolves and infer the selective mechanisms responsible.

Chapter 3 - The accumulation of weakly deleterious and exposure of strongly deleterious mutations following the transition from outcrossing to selfing.

In this investigation, I examined the frequency and strength of selection acting on deleterious mutations in outcrossing and selfing populations of *E. paniculata*. A reduction in the *N*<sub>e</sub> of selfing populations can result in the accumulation of weakly deleterious mutations following the transition from outcrossing to selfing (Charlesworth and Wright 2001). At the same time, the more homozygous selfing genome offers fewer chances for the masking of strongly deleterious mutations (Glémin 2007). Using computer simulations and polymorphism data generated from *E. paniculata*, I found selfing genomes had a greater frequency of effectively neutral and strongly deleterious mutations when compared to outcrossing genomes. The simulations also indicated that both types of mutations accumulate in selfers, when there are mutations with varying dominance levels segregating in outcrossing progenitors. The differences in patterns of mutations in outcrossing and selfing genomes were only observed early during the transition to selfing. Strongly deleterious mutations are likely to be purged quickly following this transition and recurrent selfing is likely to homogenize genomes. The resulting reduction in the frequency of segregating polymorphisms over time may have resulted in a reduced ability to detect differences in patterns of polymorphism in outcrossing and selfing genomes. Nevertheless, recently derived selfing lineages provide evidence for both the accumulation of slightly deleterious mutations and for more efficient removal of strongly deleterious recessive mutations.
Chapter 4 - Down-regulation of selfing alleles following the recent shift to selfing

In this study, I investigated patterns of regulatory differentiation accompanying the recent transition from outcrossing to selfing in *E. paniculata*. Through controlled crosses I examined differential allele expression in hybrid plants with both selfing and outcrossing alleles. I found that selfing alleles were down-regulated for <10% of genes and that these genes were under lower selective constraints in selfing populations compared to outcrossing populations. Taken together, these results suggest that down-regulation of genes in selfing populations may have contributed towards the patterns of reduced selection efficacy and an accumulation of nonsynonymous mutations.

Chapter 5 - Genetic linkage between the S and M loci governing tristyly

Two loci (*S* and *M*) govern the tristylos polymorphism in diverse tristylos families (Lewis and Jones 1992). In this chapter, I investigated the role of genetic linkage between the two loci through controlled crosses and progeny testing in *E. paniculata*. I found that the *S* and *M* loci were strongly linked and were separated by a map distance of 2.7cM. However, a literature survey indicated that linkage between the two loci is not present in all tristylos species. The presence of linkage did not seem to be associated with plant family, taxon age or ploidy level, although sample sizes of species and family were small thus preventing a rigorous comparative analysis. It is possible that linkage evolved independently in some species, or that it was present in a common ancestor and was subsequently lost. Nevertheless, the presence of linkage might have important consequences for the origin of at least one of the two loci governing tristyly and this possibility remains to be thoroughly explored.

Chapter 6 - Genetic architecture of M locus and modifiers associated with multiple transitions to selfing.
In this chapter, I investigated the genetic architecture of the \( M \) locus governing the tristyloous polymorphism, and genetic regions associated with multiple transitions from outcrossing to selfing in \( E. \) paniculata. By crossing long- and mid-styled semi-homostyles resulting from separate transitions to selfing, I generated progeny that segregated for the \( M/m \) allele. I then used genetic mapping approaches to uncover the interval regions associated with the style morph polymorphism. Style length and anther height, traits that characterize the tristyloous syndrome, localized to the same interval on a single linkage group containing highly polymorphic genes with elevated Tajima's \( D \). Further, I found that stamen modifications associated with separate transitions to selfing were governed by different interval regions. My results indicated that the broadly mapped \( M \) locus showed evidence for balancing selection, consistent with frequency-dependent selection maintaining tristyly (Barrett 1993), and that a contrasting set of modifiers unlinked to the \( M \) locus were associated with the independent origins of selfing in \( E. \) paniculata.

**Future directions**

*How does genic and regulatory variation change accompanying the transition from outcrossing to selfing?*

Much remains to be known about the distribution of deleterious and beneficial mutations in outcrossing and selfing genomes and their impact of genome-wide diversity following the evolution of selfing. Genome-wide diversity can be reduced following the transition to selfing due to background selection acting on deleterious mutations in selfing genomes (Chapter 2). Given accumulating evidence of the major role of positive selection in plant genomes (reviewed in Hough et al. 2013), selective sweeps of beneficial mutations could also result in the loss of genetic diversity following the transition to selfing. At the same time, linkage between beneficial
and deleterious mutations might prolong the time taken for selective sweeps to occur (Hartfield and Glémin 2014). Although beneficial mutations are likely to be fixed quickly, tracing patterns of genetic diversity and linkage disequilibrium in the genome can be useful to detect historical or ongoing selective sweeps (reviewed in Nielsen 2005). Using such approaches, sweeps have been suggested to fix the locus associated with the loss of self-incomparability in *Arabidopsis thaliana* (Shimizu et al. 2004), and have also been detected genome-wide in the partial selfer *Caenorhabditis elegans* (Anderson et al. 2012). Allele frequency spectra generated from polymorphism data can also be used to detect positive selection in the genome, as such a pattern of selection would result in larger fraction of nonsynonymous compared to synonymous fixations (Eyre-Walker and Keightley 2009). Given that there have been multiple independent origins of selfing from outcrossing in flowering plants (Stebbins 1957; Takebayashi and Morrell 2001; Igic and Busch 2013; Wright et al. 2013; Barrett et al. 2014), it would be interesting to investigate the strength of positive and negative selection associated with mating system differentiation in separate transitions.

While polymorphism data is a powerful means to infer the genomic consequences of selfing, one should be cautious about the inferences drawn from such data. Glémin (2007) noted that relaxed selection following the evolution of predominant selfing can be detected under a range of mutation and recombination parameters, but this pattern was less visible in polymorphism data. I found that the ability to detect differences in levels of deleterious mutations of varying dominance levels in outcrossing and selfing genomes depended on the timing of the transition to selfing (Chapter 3). One concern with current methods to infer the distribution of mutations in the genome from polymorphism data is that they assume random mating among populations (e.g. Keightley and Eyre-Walker 2007). Development of model-
fitting approaches to account for the effects of linkage between sites would facilitate more accurate profiling of the distribution of mutations in selfing populations. A further consideration with the use of polymorphism data is that pooling separate selfing transitions is likely to erroneously inflate the number of segregating polymorphisms that are detected. Selfing has evolved on multiple occasions both within and among species, but the precise number of times it has evolved is obviously unknown. Global estimates of genetic diversity when considering all selfing populations jointly are often greater than when considering independent transitions to selfing, or transitions that involve different colonization histories (e.g. *Arabidopsis* - Foxe et al. 2010; *Clarkia* - Pettengill and Moeller 2011). Complementing polymorphism metrics with ecological and demographic information on the number of times that selfing has evolved, combined with the development of tools to account for linkage between sites and the dominance of mutations, might allow future studies to make more robust inferences on the genomic consequences of selfing from polymorphism data.

The role of differential regulation and expression of genes following transitions from outcrossing to selfing and the evolution of traits characteristic of the 'selfing syndrome' are poorly understood. Currently, only a small number of studies have assessed patterns of regulatory differentiation between contrasting mating systems (e.g. *Arabidopsis* - He et al. 2012; *Capsella* - Steige et al. 2015; *Eichhornia* - Chapter 4). Whereas insertion of transposable elements has resulted in silencing of selfing alleles in *Arabidopsis* (He et al. 2012), a similar molecular mechanism has resulted in the silencing of outcrossing alleles in *Capsella* (Steige et al. 2015). Documenting the molecular mechanism of regulatory differentiation in multiple outcrossing and selfing lineages would allow a better comparison of the variation associated with gene regulation following the evolution of selfing. Also, uncovering the role and function of
differentially expressed genes may provide insight into whether patterns of regulatory
differentiation are associated with floral traits promoting autogamy, or features associated with
local adaptation following the colonization of new environments by selfing populations. An
extended regulatory profile of outcrossing and selfing genomes sampled from multiple
populations and species would provide a better understanding of how mating systems influence
gene regulation.

What is the genetic architecture of the maintenance and breakdown of the tristylos
polymorphism?

The role of genetic linkage between the $S$ and $M$ loci governing tristyly remains unclear,
especially as it is not a consistent feature of all species exhibiting this polymorphism. Linkage is
well known as a genetic mechanism maintaining favorable gene combinations (Darlington and
Mather 1949; Mather 1950), and might have contributed to the origin of the second locus after
the establishment of the first during the evolution of tristyly. However, the parameters under
which linkage is a critical factor remain to be rigorously assessed using evolutionary simulations.
Key parameters to consider when performing such simulations are the distribution of mutations
in the genome, mutation, recombination and outcrossing rates, strength of heteromorphic
incompatibility reactions of novel alleles and population size. Such detailed explorations should
provide greater insights into the evolutionary origin of the tristylos polymorphism.

Very little is currently known about the genetic architecture of the tristylos
polymorphism, particularly evidence for whether supergenes control the $S$ or $M$ loci, as they
appear to do for the $S$ locus controlling distyly (Charlesworth 1979; Barrett 1993). I broadly
mapped the $M$ locus in *E. paniculata* (Chapter 6), and also found that it is closely linked to the $S$
locus in this species (Chapter 5). Future investigations of the architecture of the $S$ locus in this
species would be informative and could be complemented by studies of the markers and genes that are associated with the $M$ locus. I discovered that the $M$ locus showed evidence for balancing selection. This finding was uninformative with respect to the supergene hypothesis as this pattern could result from either supergene control, or from a model in which one or two genes regulate the remaining genes associated with the polymorphism (Charlesworth 1979). However, the cluster of high diversity within the $M$ locus spanned 50kb and 5-6 genes, possibly reflecting the presence of a supergene cluster. One approach to uncover evidence for the presence of supergenes would be to investigate genomic evidence for chromosomal inversions surrounding the $S$ or $M$ loci, as this is a means of maintaining favorable gene combinations through suppression of recombination (Charlesworth 2016). Investigating evidence for the presence of mechanisms that suppress recombination for the two loci governing tristyly should aid in assessing if supergenes do indeed control this polymorphism.

While the evolutionary breakdown of tristyly in *E. paniculata* was accompanied by the spread of modifier genes unrelated to the $M$ locus, how these modifiers were fixed in populations is an open question. Conventional thought was that most genes involved in adaptation would be dominant due to forces such as Haldane's Sieve (reviewed in Coyne and Orr 2004). However, Fenster and Barrett (1994) suggested that the major genes associated with $M'$ semi-homostyle evolution in *E. paniculata* were likely recessive. Moreover, Charlesworth (1992) demonstrated theoretically that inbreeding can aid in facilitating the fixation of recessive modifiers in the genome. In Chapter 6, I was unable to assess this possibility for the loci associated with stamen elongation in the $M'$ morph due to the backcross crossing design used in the study. Further genetic mapping with a F$_2$ crossing design could assess this thoroughly and also help to identify genes associated with the evolutionary transition to selfing, as performed in a range of other
species (e.g. *Turnera* - Shore and Barrett 1985, 1990; *Mimulus* - Fishman et al. 2002; *Capsella* - Sicard et al. 2011; Slotte et al. 2011). Using this approach, one could also investigate the genetic architecture of the more recent origin of M’ semi-homostyles in Brazil (Husband and Barrett 1993). The transition to semi-homostyly in Brazil occurred more recently when compared to the colonization of the Caribbean by *E. paniculata*. The semi-homostyloous variants in Brazil are phenotypically identical to the tristyloous mid-styled morph, except for one modified stamen position that causes autonomous self-pollination (see Richard and Barrett 1992). It would be interesting to explicitly compare the molecular genomics of semi-homostyles in the Caribbean and Brazil. Doing so would enrich our understanding of the comparative molecular framework for how selfing populations originate and evolve and how this transition is associated with the gradual build-up of floral modifications characteristic of the 'selfing syndrome'.

**Main conclusions**

In my thesis, I have used a variety of approaches to study the maintenance and breakdown of tristyly, an adaptive multi-locus floral polymorphism assumed to be maintained by negative frequency-dependent selection (and see Eckert et al. 1996a). These approaches have included controlled crosses of Mendelian and quantitative traits and, using glasshouse studies, the investigation of phenotypic and genetic variation, population genomic approaches, computer simulations of evolutionary models, and studies of gene expression. My work on the evolutionary breakdown of this polymorphism has demonstrated how this change in mating system is associated with different patterns of molecular diversity, the distribution of deleterious mutations, differential gene expression and the spread of modifiers promoting autogamy. Further, I uncovered that the *S* and *M* loci governing tristyly are strongly linked in *E. paniculata*.
and I isolated the linkage group containing the $M$ locus. Population genomic patterns associated with the genetic interval containing the $M$ locus were consistent with the expectation that the floral polymorphism is maintained by negative frequency-dependent selection in natural populations. Below, I list five topics identified in my concluding discussion that could extend our understanding of the molecular evolutionary genomics of the maintenance and breakdown of tristyly.

- Determine if there is evidence for selective sweeps in *E. paniculata*, by studying genetic diversity in narrow windows along contiguous genomic sequences.
- Develop model-fitting approaches to infer selection efficacy across the genome and account for the effects of non-random mating in populations.
- Compare gene expression differences and the distribution of non-coding genetic elements among outcrossing and selfing populations from Brazil, the Caribbean and Central America to generate a global profile of regulatory variation in *E. paniculata*.
- Perform mapping studies and compare marker order of $ssmm$, $ssMm$ and $ssMM$ genotypes to investigate if chromosomal inversions are found in regions contain the $S$ and $M$ loci.
- Perform test crosses using tristyloous and semi-homostyloous mid-styled morphs from Brazil and generate a F$_2$ progeny to investigate the genetic architecture of the origin of semi-homostyles in this region.

Studies on the origin of selfing continue to fascinate evolutionary biologists and work on *E. paniculata* represents one of the most extensively studied model systems concerned with the evolutionary transition from outcrossing to selfing. My thesis exploited the substantial geographical variation in floral morphology and mating systems that exist in wild populations of this species and provided novel insights into the tristyloous polymorphism and
its breakdown to selfing. Future research could extend the genetic and genomic approaches I have used in my thesis to address the many unresolved questions that remain.
Bibliography


Carneiro M, Ferrand N, Nachman MW. 2009. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics* 181:593-606.


Darwin C. 1877. The different forms of flowers on plants of the same species. London: John Murray.


46:573-588.


