

GENETICS OF SAFFLOWER (*CARTHAMUS TINCTORIUS* L.; ASTERACEAE):

DOMESTICATION, DIVERSITY, AND EVOLUTION OF *CARTHAMUS*

by

STEPHANIE ANNE PEARL

(Under the Direction of John M. Burke)

ABSTRACT

The study of domestication, the process by which wild species are selected for human consumption and use, can be used as a model for understanding the genetics of adaptation. Many economically important crops (e.g., sunflower and lettuce) are in the Asteraceae, the largest family of flowering plants. This family also includes a number of underutilized crops, which can serve as valuable resources for meeting the increasing food demands of the world's growing population. *Carthamus tinctorius* L. (safflower) is an example of one such underutilized crop. Safflower is an attractive crop for further development, given that it is capable of growing in moisture-limited areas and is commercialized for its seed oils rich in unsaturated fatty acids. This dissertation has investigated the genetic diversity within and among wild, cultivated, and commercial safflowers. Additionally, the research presented here has characterized the genetic architecture of domestication traits in safflower and compared patterns of selection across crop and weeds species in the Cardueae. Population genetic analyses identified a significant decrease in allelic richness that occurred as a result of the safflower domestication bottleneck and identified useful sources for the future introgression of novel diversity from closely related, wild safflower species and parts of the safflower germplasm collection. The investigation of safflower

genetic architecture revealed that, similar to the case in sunflower and unlike many other crops, the genetics of safflower domestication is complex. Moreover, comparative mapping results suggested that parallel trait transitions in these independent crop lineages may have been driven by parallel genotypic changes. Finally, molecular evolutionary analyses among safflower and its weedy relatives showed that Cardueae crop and weed species shared similar patterns of selection. Taken together, this dissertation has contributed to a body of knowledge on the genetics of adaptation, using safflower domestication as a model.

INDEX WORDS: Asteraceae, comparative genomics, domestication, molecular evolution, population genetics, safflower

GENETICS OF SAFFLOWER (*CARTHAMUS TINCTORIUS* L.; ASTERACEAE):
DOMESTICATION, DIVERSITY, AND EVOLUTION OF *CARTHAMUS*

by

STEPHANIE ANNE PEARL

BS, Vanderbilt University, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

© 2013

Stephanie Anne Pearl

All Rights Reserved

GENETICS OF SAFFLOWER (*CARTHAMUS TINCTORIUS* L.; ASTERACEAE):
DOMESTICATION, DIVERSITY, AND EVOLUTION OF *CARTHAMUS*

by

STEPHANIE ANNE PEARL

Major Professor:	John M. Burke
Committee:	Michael Arnold
	Shu-Mei Chang
	Katrien Devos
	James L. Hamrick

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2013

DEDICATION

I dedicate this dissertation to Mom and Dad, for always providing me with their love and support no matter what I pursue, and to my husband, for his endless patience, maintaining my sanity, and filling my life with love and laughter throughout these last five and a half years.

ACKNOWLEDGEMENTS

I would like to thank my advisor, John M. Burke, for training and supporting me throughout my graduate career. Working with John has opened many doors for me and has improved my science writing and problem solving abilities. My research in the Burke lab has been made possible by the NSF-funded *Compositae* Genome Project, and I am grateful to have been a part of this collaboration. I would also like to thank my committee members Michael Arnold, Shu-Mei Chang, Katrien Devos, and Jim Hamrick, who have all been very encouraging as they have challenged and helped me to become a scientist. Much of my research has also been made possible by several members of the safflower community, including Vicki Bradley and Richard Johnson at the USDA, Art Weisker at SeedTec, Jerry Bergman at Safflower Technologies International, and Amram Ashri of the Hebrew University of Jerusalem. I am exceptionally grateful to the wonderful staff in the Department of Plant Biology, especially Susan Watkins, for lighting the path through graduate school, Shannon Kennedy, for helping me throughout my grant applications, and Mike Boyd, for maintaining a wonderful greenhouse space. I also thank our graduate coordinator, Lisa Donovan, and department head, Michelle Momany, for leading a great department and offering their continued support to the Plant Biology graduate student body. I thank all of my fellow Burke Lab members, especially Jennifer Mandel, Mark Chapman, Jennifer Dechaine, Jessica Barb, John Bowers, Jonathan Corbi, Savithri Nambeesan, Ed McAssey, and Adam Bewick, whose expertise and friendship have helped me to surmount many hurdles throughout my graduate career. Finally, I am grateful for the friendships that I have developed with Lisa Kanizay, Michael McKain, and Jeremy Rentsch, from whom I

have learned much about science and even more about myself. I cherish the fact that, together, we brought outreach to UGA Plant Biology.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
 CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
REFERENCES	10
2 GENETIC ANALYSIS OF SAFFLOWER DOMESTICATION REVEALS PARALLELS WITH SUNFLOWER DOMESTICATION	16
ABSTRACT.....	17
BACKGROUND	18
METHODS	20
RESULTS	28
DISCUSSION	32
REFERENCES	36
3 ANALYSIS OF GENETIC DIVERSITY IN <i>CARTHAMUS TINCTORIUS</i> L. (SAFFLOWER), AN UNDERUTILIZED CROP	57
ABSTRACT.....	58
INTRODUCTION	59
MATERIALS AND METHODS.....	61

RESULTS	64
DISCUSSION	67
REFERENCES	70
4 MOLECULAR EVOLUTIONARY ANALYSES OF SAFFLOWER AND ITS WEEDY RELATIVES IN THE CARDUEAE (ASTERACEAE).....	83
ABSTRACT.....	84
INTRODUCTION	85
METHODS	87
RESULTS	91
DISCUSSION	95
REFERENCES	98
5 CONCLUSIONS.....	113
REFERENCES	116
APPENDICES	
A SUPPORTING INFORMATION FOR CHAPTER II	117
B SUPPORTING INFORMATION FOR CHAPTER III	126
C SUPPORTING INFORMATION FOR CHAPTER IV	142

LIST OF TABLES

	Page
Table 2.1: Average trait values of mapping parents	44
Table 2.2: Quantitative trait locus positions, modes of gene action, and magnitudes of effect for 19 out of the 24 traits studied.....	45
Table 3.1: Collection information of wild safflower samples used in this study	74
Table 3.2: Genetic diversity statistics of wild, cultivated (Old World and New World), and commercial (CO and STI) safflower groupings.....	74
Table 3.3: Genetic diversity statistics comparing wilds with USDA cultivated safflower accessions (Old World and New World combined)	75
Table 4.1: Collection information and tissue types used in the RNA extraction of the five species used in this study.....	104
Table 4.2: Sequencing information regarding each of the species used in this study	105
Table 4.3: Functional annotations of genes exhibiting positive selection	106

LIST OF FIGURES

	Page
Figure 2.1: Trait distributions of the F ₂ mapping population and mapping parent representatives ..	
.....	47
Figure 2.2: Genetic map of the safflower genome and corresponding quantitative trait locus	
positions	50
Figure 2.3: Comparative mapping of the safflower, lettuce, and sunflower genomes	55
Figure 3.1: STRUCTURE plots of 190 <i>Carthamus</i> individuals	76
Figure 3.2: Map of sampling locations of Old World and New World USDA accessions	78
Figure 3.3: STRUCTURE plot of 142 <i>Carthamus</i> individuals in which wild, New World, CO,	
and STI accessions are assigned to one of four Old World populations	81
Figure 3.4: Principal coordinates analysis of 190 <i>Carthamus</i> individuals	82
Figure 4.1: Phylogenetic relationships among the species used in this study	109
Figure 4.2: Example orthogroup in which diversifying selection is acting upon two codon sites	
.....	110
Figure 4.3: Venn diagram depicting positively selected genes that are shared among and unique	
to <i>C.tinctorius</i> , <i>C. palaestinus</i> , <i>C. oxyacanthus</i> , and the <i>Centaurea/Cirsium</i> outgroup ..	111
Figure 4.4: Example orthogroup in which positive selection is acting jointly on <i>C. oxyacanthus</i>	
and <i>C. palaestinus</i> , which share the same protein sequence.....	112

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

“I have seen great surprise expressed in horticultural works at the wonderful skill of gardeners, in having produced such splendid results from such poor materials; but the art has been simple, and as far as the final result is concerned, has been followed almost unconsciously. It has consisted in always cultivating the best-known variety, sowing its seeds, and, when a slightly better variety chanced to appear, selecting it, and onwards.”

--Charles Darwin [1]

Understanding domestication, the process by which humans have cultivated wild species and caused them to change into a form more useful for human purposes, has fascinated biologists since the time of Darwin. Domestication can be viewed as a special case of adaptation, given that crops are adapted for cultivation in human settlements in much the same way that wild plant populations are adapted to natural environments [3]. This is true regardless of whether selection for agronomically important traits has resulted from artificial (intentional) or automatic (unconscious) selection [4]. Domesticated species have also been considered the products of “long-term selection experiments” [5] and are therefore useful as models for studying the molecular consequences of phenotypic selection. Such studies can provide insight into the genetic consequences of selective sweeps as well as the limitations that the genetic architecture of a particular crop imposes on its response to selection [6–8], thereby providing knowledge that will benefit both the evolutionary biology and crop breeding communities.

History of the study of the genetics of adaptation. Comprehensive analyses of the genetic architecture of adaptive phenotypes have only been conducted for the past few decades. Prior to the availability of molecular markers for the study of plant domestication and adaptation, evolutionary biologists developed mathematical models to describe the number and effect size of mutations involved in selective response. The most influential work in this area for much of the twentieth century was the *Genetical Theory of Natural Selection* [9]. This work presented the “geometric model,” which predicted the likelihood that any random mutation conferring a phenotypic effect of any magnitude would be beneficial. This became the basis of what was later coined the “infinitesimal model” [10], which predicted that “infinitesimally small” mutations have a 50% chance of being favorable while large effect mutations have a very small likelihood of being beneficial. Therefore, the thought that small mutations are the basis of adaptation shaped much of the Modern Synthesis. This did little to inspire further investigation into the genetics of adaptation, given that mutations would be too infinitesimally small and numerous to detect (reviewed in [11–13]).

The advent of quantitative trait locus (QTL) mapping provided an opportunity to investigate the number of genes involved in the evolution of adaptations as well as the distribution of their phenotypic effects. Early mapping studies suggested that domestication traits were largely controlled by just a few genes with large phenotypic effects (e.g., fruit weight in tomato [14], seed dormancy and seed weight in common bean [15], and branching in maize [16]). However, these observations alone did not disprove the infinitesimal model [17], as it is possible to overestimate QTL effect sizes and fail to identify small-effect QTL as an artifact of small sample sizes (the “Beavis Effect” [18]) or as a result of linkage disequilibrium and epistatic interactions among QTL [8]. More recent QTL studies with larger sample sizes and

more extensive genotyping have shown that, in many cases, domestication traits are controlled by a few large-effect QTL plus other QTL with moderate or small effects [19]. Further, recent crop population genomic studies have shown that large numbers of genes are typically under selection during crop domestication and/or improvement ([20, 21]; reviewed in [22, 23]). These findings indicate that the genetic architecture of domestication traits is likely to be complex, even for crops in which initial QTL-based approaches have suggested otherwise.

Parallel domestication. Comparison of QTL analyses can provide insight into the frequency with which parallel genotypic changes (changes in the same genes across species yielding the same traits) respond to selection for homologous traits comprising the “domestication syndrome” (e.g., increased fruit size, loss of seed dormancy, uniform flowering time, loss of shattering; [24, 25]) among different crop species. Cases in which homologous QTL in different crops map to the same conserved regions of the genome suggest that changes at homologous loci were selected upon during domestication. Comparative QTL mapping among crops in the Fabaceae [26], Poaceae [27], and Solanaceae [28] have provided evidence suggesting that many domestication traits, such as seed weight, fruit weight, short-day flowering, and perenniality, may indeed be controlled by the same genes. Other approaches have been used to identify parallel evolution in several other systems, suggesting that parallel evolution occurs at all taxonomic levels (reviewed in [29]).

Understanding the extent to which parallel genotypic evolution occurs will help evolutionary biologists understand the dynamics governing the genetics of adaptation and extent to which artificial and natural selection are similar [4]. Examples of parallel evolution suggest that there are rate-limiting, key elements of evolution that constrain phenotypic potentials [30]. Parallel genotypic evolution may result from genetic or developmental constraints [31–33]

(reviewed in [34]). For example, genes with unique roles within developmental pathways are constrained and therefore provide smaller targets for mutation [35]. Genetic biases might also occur, either because certain mutations yield a greater phenotypic response than others or because they enjoy greater selective benefits due to their pleiotropic effects, therefore leading to disproportionately strong positive selection favoring these genetic changes [34].

Why are so few plants domesticated? Given that domestication is arguably the most important innovation of human societies in the past 13,000 years [36], it is a wonder that so few species have become domesticated. Out of the estimated 300,000 species of plants that exist today [25], fewer than 1% have been domesticated by humans [2, 5, 36]. Historical evidence suggests that this paucity of domesticated species is not due to a lack of human effort. For example, multiple crops have been rapidly, and in some cases, repeatedly (e.g., rice [37], *Phaseolus* beans [19, 38], and barley [39]) domesticated in multiple parts of the world. Further, ancient discoveries have revealed that thousands of species have been routinely harvested in the wild and cultivated as an early crop, but are not domesticated today [36]. For example, goosefoot, sumpweed, and knotweed were among the early crops cultivated in the eastern United States [40, 41]. However, goosefoot and knotweed seeds were too small. Further, sumpweed produced hay fever-inducing pollen and emitted a foul odor. These undesirable characteristics hindered further efforts to domesticate any of these three species. Finally, “Our failure to domesticate even a single major new food plant in modern times suggests that ancient people really may have explored virtually all useful wild plants and domesticated all the ones worth domesticating” [2].

Why have humans domesticated so few species? Theory predicts that adaptation can occur only in species with sufficient genetic variation (e.g., [42, 43]), and simulations conducted

on *Zea* species supports the hypothesis that effective population size is a preadaptation for domestication [44]. Further, genetic architecture may play a role in predisposing wild species to domestication. For example, domestication QTL are commonly clustered throughout the genomes of many crops (e.g., pearl millet [45], pepper [46], and safflower (Pearl et al., in revision), and it has been speculated that species with clustered domestication genes may have been inherently easier to domesticate [5]. Additionally, the complexity of the genetic architecture may have limited a wild species' potential for domestication (but see [23, 47], and Pearl et al., in revision). Consider wild almonds and oaks: both tree nuts contain compounds that are toxic to humans. In almonds, the toxic compound amygdalin is controlled by a single gene, while the toxins in acorns are polygenic. The non-poisonous condition in almonds is dominant, therefore improving the ease with which almonds were domesticated for human consumption (reviewed in [36]).

The importance of leveraging germplasm collections and underutilized crops.

Society today relies overwhelmingly on just four species: rice, maize, wheat, and soybean. These crops offer only a limited set of micronutrients and require a very specific set of environmental conditions to produce high yields. To improve the economic viability and competitiveness of American agriculture, it is necessary to improve the quality of agricultural resources in order to benefit society as a whole [48]. This is challenging, given the pressures resulting from growing population sizes and the consequently reduced availability of natural resources. Further, our increasingly variable and extreme climate places additional stress on these natural resources.

The American food and agricultural industry would also benefit from developing crops that offer consumers healthier choices and a more diverse array of micronutrients. This is especially important for helping to reduce obesity rates, which have doubled in adults and tripled

in children over the past 30 years [49]. These issues are especially pressing for low-income individuals, who often face limited access to healthy food options and are at greatest risk for obesity. One strategy for increasing food availability and enhancing American nutrition is to alter the food supply by increasing the diversity of crops produced by American agriculture [49, 50]. As recently emphasized by the chief of the United Nations Food and Agriculture Organization at the International Crops for the 21st Century seminar, increased research and development of underutilized crops can help address food security issues. Underutilized crops are defined as those domesticated species whose genetic potential has not been fully realized [51]. These non-commodity crops are part of a “larger biodiversity portfolio” that tends to be underused by farmers and consumers for a variety of agronomic, economic, and cultural factors [51]. Specifically, by increasing the diversity of crops produced by our agriculture, important micronutrients will become increasingly available to the public [52] and the soil health of agroecosystems will improve via the introduction of complex crop rotations [53]. Therefore, by leveraging currently underutilized crops, the American food and agricultural industry can protect our food security and enhance quality of life for our society as a whole[54].

The establishment and subsequent genetic characterization of germplasm collections is an essential component of leveraging and securing the resource base of underutilized crops [55]. These germplasm collections are often comprised of cultivated materials obtained from throughout the world and may include closely related, wild species. Materials acquired from a crop’s center of origin (i.e., region of initial cultivation) are especially valuable, given that these regions typically harbor higher levels of diversity. Leveraging the diversity maintained within these materials can help reintroduce to crops the genetic variation on which breeding depends [56], as crops have lost the majority of this diversity during the “domestication bottleneck” that

results from the breeding process [57]. Although wild germplasm may not physically appear to be a desirable source of variation, these accessions may be the source of desirable alleles for agriculturally important, multi-gene traits that are not yet present within breeding lines. Additionally, wild and exotic plants tend to be resilient in the face of environmental stresses (such as disease) and can thus serve as the source of valuable stress resistant genes [58]. By “harnessing the wealth” of genetic variation stored in gene banks and present throughout nature [56], it can become possible to further improve crops and support the demands of our growing population.

The Asteraceae. One important family for the focus of crop evolution studies is the Asteraceae (e.g., the Compositae), the largest family of flowering plants. Sunflower (*Helianthus annuus* L.) and lettuce (*Lactuca sativa* L.) are the two most economically important crops within the Asteraceae and have been the subject of multiple population genetic and genomic analyses over the past 10-15 years (e.g., [59–61], among many others since then). In addition to sunflower and lettuce, the Asteraceae includes underutilized and many weedy and economically costly, noxious species. Therefore, closer examination of the members within this family can provide resources for the development of underutilized crops while also shedding light into why certain species become crops and why others become invasive.

The Cardueae tribe is one of the largest tribes within the Asteraceae [62] and is known for its weedy thistles. Only a few species have become domesticated within the Cardueae, and therefore it can serve as a study system for comparing crop and weed evolution. Members of this tribe include *Cirsium arvense* L. (spotted knapweed) and *Centaurea solstitialis* L. (yellow star thistle), which are among the most damaging weeds in the Asteraceae [63], and *Carthamus*

oxyacanthus Bieb., a federally listed noxious weed that is closely related to the underutilized crop *Carthamus tinctorius* L. (safflower).

Safflower is an annual, self-compatible, thistle-like, diploid ($2n = 2x = 24$) crop believed to have a single origin of domestication in the Fertile Crescent region dating to approximately 4,500 years ago [64, 65]. Its haploid genome size is approximately 1.4 Gb [66]. Safflowers have long taproots that facilitate water uptake in even the driest environments, enabling this crop to be grown on marginal lands where moisture would otherwise be limiting. Initially, safflower was cultivated for its flowers, which were largely used for dyes as well as in teas and as a food additive. Safflower florets have also been used for medicinal purposes in some parts of the world. For example, extracts from these florets have been shown to reduce hypertension and reduce blood cholesterol levels [67].

More recently, safflower has been selected for its high quality, healthy seed oil which is rich in unsaturated fatty acids. Commercialization of safflower as an oilseed crop first began in North America during the 1950s, approximately 50 years after it was brought to North America [68]. Safflower breeding programs that began in the 1940s and 1950s investigated the inheritance of various traits, such as flower color, spininess, and oil content. Meanwhile, Paulden F. Knowles, “the father of California safflower,” played a key role in introducing and developing safflower as a crop within the United States. During the 1950s and 1960s, Knowles traveled throughout North Africa, the Middle East, and South Asia in an effort to collect germplasm representative of safflower and its wild relatives found throughout their native range. The germplasm that Knowles collected and developed is among the more than 2,300 accessions of safflower currently maintained by the USDA.

Crops and weeds, at first, seem strikingly different from one another: they vary in their levels of desirability to humans, and many of the traits comprising the domestication syndrome (e.g., loss of seed dormancy and shattering) are maladaptive in non-crop species. Further, relative to weeds, crops have an increased dependence on man for success in permanently disturbed, man-made habitats [69]. Despite these differences, crops and weeds also share a number of similarities: rapid growth, short generation times, and when considering annual seed crops, a relatively increased investment into reproduction. The traits that both crops and weeds have developed are the response to periods of strong, directional selection. Analyses of molecular evolution among homologous sets of genes among closely related crops and weeds, such as those in the Cardueae, can yield insight into how positive selection across the genome has compared during the evolution and domestication of weeds and crops.

Purpose of Study. This work uses safflower as a model for studying genome and trait evolution within the Asteraceae. The QTL mapping of domestication traits in safflower is an important first step for generating genomic resources for this underutilized crop and providing the basis for comparative genomic analyses across three major subfamilies in the Asteraceae. Here, it has also enabled an investigation into parallel domestication within the Asteraceae. Second, the estimation of genetic diversity within wild safflower individuals and across safflower germplasm and breeding lines informs breeders of sources of novel diversity that may prove beneficial in the advancement of their breeding programs. Finally, our investigation of the rates of nonsynonymous vs. synonymous substitutions within gene families encompassing five Cardueae species (three weeds, a crop, and its progenitor) reveal the targets of positive selection, providing insights into what has driven these different species down different evolutionary

trajectories. As a whole, this dissertation contributes to a body of knowledge on the genetics of adaptation, using domestication as a model.

REFERENCES

1. Darwin C: *On the Origin of Species by Natural Selection*. 1859.
2. Diamond J: *Guns, Germs, and Steel*. New York, New York: W.W. Norton & Company, Inc.; 1997.
3. Ross-Ibarra J, Morrell PL, Gaut BS: **Plant domestication, a unique opportunity to identify the genetic basis of adaptation**. *PNAS* 2007, **104**:8641–8648.
4. Purugganan MD, Fuller DQ: **The nature of selection during plant domestication**. *Nature* 2009, **457**:843–848.
5. Gepts P: **Crop domestication as a long-term selection experiment**. *Plant Breeding Reviews* 2004, **24.2**:1–44.
6. Stuber C, Moll R, Goodman M, Schaffer H, Weir B: **Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays*L.)**. *Genetics* 1980, **95**:225–236.
7. Wang RL, Stec A, Hey J, Lukens L, Doebley J: **The limits of selection during maize domestication**. *Nature* 1999, **398**:236–239.
8. Bost B, De Vienne D, Hospital F, Moreau L, Dillmann C: **Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects**. *Genetics* 2001, **157**:1773–1787.
9. Fisher RA: *The Genetical Theory of Natural Selection*. Oxford, UK: Oxford University Press; 1930.
10. Bulmer M: *The Mathematical Theory of Quantitative Genetics*. Oxford, England: Clarendon Press; 1980.
11. Orr HA, Coyne JA: **The genetics of adaptation : a reassessment**. *The American Naturalist* 1992, **140**:725–742.
12. Orr HA: **The population genetics of adaptation: the distribution of factors fixed during adaptive evolution**. *Evolution* 1998, **52**:935–949.
13. Orr HA: **The genetic theory of adaptation: a brief history**. *Nature Reviews Genetics* 2005, **6**:119–27.

14. Frary A, Nesbitt TC, Frary A, Grandillo S, Van der Knapp E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD: **fw2.2: A quantitative trait locus key to the evolution of tomato fruit size.** *Science* 2000, **289**:85–88.
15. Koinange EMK, Singh SP, Gepts P: **Genetic control of the domestication syndrome in common bean.** *Crop Science* 1996, **36**:1037–1045.
16. Doebley J, Stec A: **Genetic analysis of the morphological differences between maize and teosinte.** *Genetics* 1991, **129**:285–295.
17. Mackay TFC: **The genetic architecture of quantitative traits.** *Annual Review of Genetics* 2001, **35**:303–339.
18. Beavis W: **The power and deceit of QTL experiments: lessons from comparative QTL studies.** In *49th Annual corn and sorghum research conference*. Washington, DC: American Seed Trade Association; 1994:250–266.
19. Poncet V, Sarr TRA, Gepts P: **Quantitative trait locus analyses of the domestication syndrome and domestication process.** In *Encyclopedia of Plant and Crop Science*. New York: Marcel Dekker, Inc.; 2004:1069–1074.
20. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS: **The effects of artificial selection on the maize genome.** *Science* 2005, **308**:1310–4.
21. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright R a, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner R a, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J: **Comparative population genomics of maize domestication and improvement.** *Nature Genetics* 2012, **44**:808–811.
22. Morrell PL, Buckler ES, Ross-Ibarra J: **Crop genomics: advances and applications.** *Nature Reviews Genetics* 2012, **13**:85–96.
23. Wallace JG, Larsson SJ, Buckler ES: **Entering the second century of maize quantitative genetics.** *Heredity* 2013(January):1–9.
24. Harlan JR, De Wet JMJ, Price EG: **Comparative evolution of cereals.** *Evolution* 1973, **27**:311–325.
25. Harlan JR: *Crops and Man*. 2nd ed. Madison, WI: American Society of Agronomy; 1992.
26. Fatokun CA, Menancio-hautea DI, Danesh D, Young ND: **Evidence for orthologous seed weight genes in cowpea and mung bean based on RFLP mapping.** *Genetics* 1992, **132**:841–846.

27. Paterson A, Lin YR, Li Z, Schertz KF, Doebley JF, Pinson SR, Liu SC, Stansel JW, Irvine JE: **Convergent domestication of cereal crops by independent mutations at corresponding genetic loci.** *Science* 1995, **269**:1714–8.
28. Doganlar S, Frary A, Daunay M-C, Lester RN, Tanksley SD: **Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant.** *Genetics* 2002, **161**:1713–1726.
29. Wood TE, Burke JM, Rieseberg LH: **Parallel genotypic adaptation: when evolution repeats itself.** *Genetica* 2005, **123**:157–170.
30. Elmer KR, Meyer A: **Adaptation in the age of ecological genomics: insights from parallelism and convergence.** *Trends in ecology & evolution* 2011, **26**:298–306.
31. Wake D: **Homoplasy: The result of natural selection, or evidence of design limitations?** *The American Naturalist* 1991, **138**:543–567.
32. Shubin N, Wake DB, Crawford A: **Morphological variation in the limbs of *Taricha granulosa* (Caudata: Salamandriidae): Evolutionary and phylogenetic implications.** *Evolution* 1995, **49**:874–884.
33. West Eberhard M: *Developmental Plasticity and Evolution*. New York: Oxford University Press; 2003.
34. Schluter D, Clifford E a, Nemethy M, McKinnon JS: **Parallel evolution and inheritance of quantitative traits.** *The American Naturalist* 2004, **163**:809–822.
35. Futuyma DJ: **Evolutionary constraint and ecological consequences.** *Evolution* 2010, **64**:1865–1884.
36. Diamond J: **Evolution, consequences and future of plant and animal domestication.** *Nature* 2002, **418**:700–707.
37. Londo JP, Chiang Y-C, Hung K-H, Chiang T-Y, Schaal B a: **Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*.** *Proceedings of the National Academy of Sciences* 2006, **103**:9578–9583.
38. Sonnante G, Stockton T, Nodari RO, Becerra Velásquez VL, Gepts P: **Evolution of genetic diversity during the domestication of common-bean (*Phaseolus vulgaris* L.).** *Theoretical and applied genetics* 1994, **89**:629–635.
39. Morrell PL, Clegg MT: **Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent.** *Proceedings of the National Academy of Sciences* 2007, **104**:3289–3294.

40. Delcourt PA, Delcourt HR, Ison CR, Sharp WE, Gremillion KJ: **Prehistoric human use of fire, the eastern agricultural complex, and Appalachian oak-chestnut forests: Paleoecology of Cliff Palace Pond, Kentucky.** *American Antiquity* 1998, **63**:263–278.
41. Gremillion KJ: **Plant husbandry at the Archaic/Woodland transition: Evidence from the Cold Oak shelter, Kentucky.** *Midcontinental Journal of Archaeology* 1993:161–189.
42. Avise JC, Hamrick J (Eds): *Conservation Genetics*. Springer; 1996.
43. Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits*. 1998.
44. Ross-Ibarra J: **Recombination, genetic diversity, and plant domestication.** The University of Georgia; 2006.
45. Poncet V, Lamy F, Devos KM, Gale MD, Sarr A, Robert T: **Genetic control of domestication traits in pearl millet (*Pennisetum glaucum* L., Poaceae).** *Theoretical and Applied Genetics* 2000, **100**:147–159.
46. Chaim A Ben, Paran I, Grube RC, Jahn M, Van Wijk R, Peleman J: **QTL mapping of fruit-related traits in pepper (*Capsicum annuum*).** *Theoretical and Applied Genetics* 2001, **102**:1016–1028.
47. Burke JM, Tang S, Knapp SJ, Rieseberg LH: **Genetic analysis of sunflower domestication.** *Genetics* 2002, **161**:1257–1267.
48. *Toward Sustainable Agricultural Systems in the 21st Century*. Washington, DC: The National Academies Press; 2010.
49. USDA: *Nutrition and Childhood Obesity Science White Paper*. 2012.
50. Wallinga D: **Agricultural policy and childhood obesity: a food systems and public health commentary.** *Health affairs* 2010, **29**:405–10.
51. Padulosi S, Hoeschle-Zeledon I: **Underutilized plant species : what are they ?** *Leisa Leusden* 2004:5–6.
52. **FAO Media Centre: Neglected crops need a rethink - can help world face the food security challenges of the future, says Graziano da Silva**
[<http://www.fao.org/news/story/en/item/166368/icode/>]
53. Merrill SD, Tanaka DL, Hanson JD: **Root length growth of eight crop species in haplustoll soils.** *Soil Science Society of America Journal* 2002, **66**:913–923.
54. Mayes S, Massawe FJ, Alderson PG, Roberts J a, Azam-Ali SN, Hermann M: **The potential for underutilized crops to improve security of food production.** *Journal of experimental botany* 2012, **63**:1075–9.

55. Padulosi S, Hodgkin T, Williams JT, Haq N: **Underutilized crops : trends, challenges and opportunities in the 21 st Century**. In *Perspectives on New Crops and New Uses*. Edited by Janick J. Alexandria, VA: ASHS Press; 1999:140–145.
56. Tanksley SD, McCouch SR: **Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild**. *Science* 1997, **277**:1063–1066.
57. Eyre-Walker A, Gaut R, Hilton H, Feldman D, Gaut B: **Investigation of the bottleneck leading to the domestication of maize**. *Proceedings of the National Academy of Sciences* 1998, **95**:4441–4446.
58. Smedegaard-Petersen V: **The Limiting Effect of Disease Resistance on Yield**. *Annual Review of Phytopathology* 1985, **23**:475–490.
59. Rieseberg L, Choi H, Chan R, Spore C: **Genomic map of a diploid hybrid species**. *Heredity* 1993, **70**:285–293.
60. Kesseli R V, Paran I, Michelmore RW: **Analysis of a Detailed Genetic Linkage Map of Lactuca sativa (Lettuce) constructed from RFLP and RAPD markers**. *Genetics* 1994, **136**:1435–1466.
61. Arias D, Rieseberg LH: **Genetic relationships among domesticated and wild sunflowers (Helianthus annuus, Asteraceae)**. *Economic Botany* 1995, **49**:239–248.
62. Susanna A, Garcia-jacas N: **Cardueae (Carduoideae)**. In *Systematics, evolution, and biogeography of the Compositae*. Edited by Funk VA, Susanna T, Stuessy T, Bayer R. Vienna, Austria: International Association for Plant Taxonomy; 2009:293–313.
63. LeJeune KD, Seastedt TR: **Centaurea Species : the Forb That Won the West**. *Conservation Biology* 2001, **15**:1568–1574.
64. Knowles P, Ashri A: **Safflower -- Carthamus tinctorius (Compositae)**. In *Evolution of Crop Plants*. Edited by Smartt, J. and Simmonds NW. Harlow, UK: Longman; 1995:47–50.
65. Van Zeist W, Rooijen Waterbolk-Van W: **Two interesting floral finds from third millenium BC Tell Hammam et-Turkman, northern Syria**. *Vegetation History and Archaeobotany* 1992, **1**:157–161.
66. Garnatje T, Garcia S, Vilatersana R, Vallès J: **Genome size variation in the genus Carthamus (Asteraceae, Cardueae): systematic implications and additive changes during allopolyploidization**. *Annals of botany* 2006, **97**:461–7.
67. Guimiao W, Yili L: **Clinical application of safflower (Carthamus tinctorius)**. *Traditional Chines Medical Science Journal* 1985, **1**:42–43.
68. Knowles P: **Safflower**. *Advances in Agronomy* 1958, **10**:289–323.

69. De Wet MJM, Harlan JR: **Weeds and domesticates: evolution in the man-made habitat.**
Economic Botany 1975, **29**:99–108.

CHAPTER II

GENETIC ANALYSIS OF SAFFLOWER DOMESTICATION REVEALS PARALLELS
WITH SUNFLOWER DOMESTICATION¹

¹ Pearl, S.A., J.E. Bowers, S. Reyes-Chin-Wo, R. Michelmore, J.M. Burke. Submitted to BMC Plant Biology, 9/19/13.

ABSTRACT

Background

Safflower (*Carthamus tinctorius* L.) is an oilseed crop in the Asteraceae that is valued for its oils rich in unsaturated fatty acids. Here, we present an analysis of the genetic architecture of safflower domestication and compare our findings to those from sunflower (*Helianthus annuus* L.), an independently domesticated oilseed crop within the same family.

We mapped quantitative trait loci (QTL) underlying 24 domestication-related traits in progeny from a cross between safflower and its wild progenitor, *Carthamus palaestinus* Eig. Also, we compared QTL positions in safflower against those that have been previously identified in crop x wild crosses in sunflower to identify instances of colocalization.

Results

We mapped 61 QTL, the vast majority of which (59) exhibited minor or moderate phenotypic effects. The two large-effect QTL corresponded to one each for flower color and leaf spininess. A total of 15 safflower QTL colocalized with previously reported sunflower QTL for the same traits. These traits included flowering time, disc diameter, number of heads, number of selfed seed, achene weight, length, and width, percent oil content, and percent oleic and linoleic acid.

Conclusions

Similar to the case in sunflower and unlike many other crops, our results suggest that the genetics of safflower domestication is complex. Moreover, our comparative mapping results suggest that parallel trait transitions in these independent crop lineages may have been driven by parallel genotypic changes.

BACKGROUND

The process of domestication, which has long been considered to be a form of “applied evolution,” inspired some of the earliest studies of evolution in response to natural selection [1]. Indeed, given the parallels between the adaptation of domesticated species to human-disturbed environments and the adaptation of wild populations to survival in natural environments [2], evolution under domestication is viewed by many as a valuable opportunity for studying the genetics of adaptation. Because many crop species share a common suite of traits (e.g. loss of seed dormancy, uniform flowering time, and fruit size) that evolved in response to selection during domestication (referred to as the “domestication syndrome”; [3]), comparative analyses across independent crop lineages also hold great promise for studying the genetic basis of parallel phenotypic evolution.

Over the years, quantitative trait locus (QTL) mapping has been used to investigate the genetic architecture of traits comprising the domestication syndrome in numerous crop species. Although early QTL-based studies suggested that domestication traits were predominantly controlled by a small number of large-effect QTL (e.g. [4–6]), other studies have revealed a higher level of genetic complexity (e.g. [7, 8]). Comparisons among QTL analyses can also provide insight into the extent to which parallel phenotypic changes across independent crop lineages are driven by selection on homologous genes, or at least genomic regions. For example, comparative QTL mapping across crops in the Fabaceae [9], Poaceae [10, 11], and Solanaceae [12] has provided evidence that many domestication traits, including increased seed weight, increased fruit size, and changes in flowering time and life history may be conditioned by independent changes in homologous genes in different lineages. Beyond providing fundamental evolutionary insights, such comparative genomic analyses also have the potential to aid in the

improvement of other crops about which less is known. For example, knowledge that the *Arabidopsis* dwarfing gene, *GAI*, is structurally and functionally homologous to the wheat and maize dwarfing genes *RHT-B1*, *RHT-D1*, and *D8*, led to the transformation of the *GAI* gene into basmati rice to produce dwarf varieties [13].

In the present study, we investigate the genetic basis of the domestication syndrome in the oilseed crop safflower (*Carthamus tinctorius* L.; Carduoideae). Safflower is an annual, self-compatible, diploid ($2n = 2x = 24$; [14]) crop believed to have had a single origin of domestication in the Fertile Crescent region approximately 4000 years ago [15]. This species is well-adapted to growth in dry environments, having a long taproot (reported to grow over 1.5m; [16, 17]), that enables water uptake even when surface moisture is limiting. Originally, safflower was cultivated for its floral pigments (carthamine; [18]). Since its initial domestication, safflower cultivation has spread to other parts of the world, including many underdeveloped countries (e.g. Ethiopia, Afghanistan, and Sudan). Commercialization of safflower in the Americas began in the 1950s, where it has largely been used as an oilseed crop, in bird seed mixes, and as an ornamental species. Safflower is especially attractive as an oilseed crop, given that its seed oils are rich in mono- and polyunsaturated fatty acids. Though safflower possesses many of the standard traits that comprise the domestication syndrome (e.g. loss of seed dormancy, uniform flowering time, increased seed production, and increased seed oil quality and content), most cultivated safflower varieties have retained certain weed-like characteristics of their wild relatives (e.g. branching and leaf spines).

Safflower is a member of the Asteraceae, which is currently recognized as the largest family of flowering plants [19, 20]. This family contains ca. 10% of all flowering plant species [20] and includes over 40 economically important crops grown for a variety of uses, such as

safflower, lettuce (*Lactuca sativa* L.; Cichorioideae), and sunflower (*Helianthus annuus* L.; Asteroideae). These three crops represent the three major subfamilies within the Asteraceae, which collectively account for 95% of the species diversity within the family. Like safflower, sunflower is primarily grown as an oilseed crop. Given this, along with the wealth of available information on the origin and evolution of cultivated sunflower (e.g. [7, 8, 21–25]), our work also provides an opportunity to study the genetic basis of parallel phenotypic changes during domestication within this important family.

Here, we describe a genetic map-based study of domestication-related traits in a population derived from a cross between safflower and its wild progenitor (*C. palaestinus* Eig.; see below). Our results indicate that the genetic architecture of safflower domestication is complex, with the majority of traits being controlled by multiple QTL with small to moderate phenotypic effects. Moreover, a comparison of our results to those derived from similar analyses in sunflower provides evidence of QTL colocalization, suggesting that parallel trait transitions may have been driven by parallel genotypic changes in these independent crop lineages.

METHODS

Mapping population. Seeds obtained from the USDA for both safflower (cv AC Sunset; PI 592391) and *C. palaestinus* (PI 235663) were germinated in the University of Georgia greenhouses during the summer of 2009. AC Sunset is an inbred, dual-purpose birdseed and oilseed cultivar developed in Canada (Mündel *et al.*, 1996). Like many other high oil varieties, the leaf margins of AC Sunset plants have prominent spines. Genetic analyses based on nuclear and chloroplast markers [27] as well as archaeological [28] and geographic evidence [18, 29] all point to the predominantly selfing *C. palaestinus* ($2n = 2x = 24$; [30]) as the wild progenitor of safflower. This species is native to the Middle East in the area around Israel and is fully cross-

compatible with safflower. Though it exhibits considerable morphological variation for a variety of traits including leaf spininess, leaf shape, duration of rosette habit, and flower color, *C. palaestinus* can be distinguished from safflower based on its tendency toward non-uniform germination, an extended rosette habit, and reduced seed size. Also, contrary to the expectation based on most crop-wild comparisons, *C. palaestinus* exhibits more limited branching than safflower ([31]; unpublished observation).

A single safflower plant served as a pollen donor in a cross between safflower and its progenitor. The F_1 seeds from this cross were germinated and the resultant plants were selfed to produce F_2 families, the largest of which was chosen for study in the QTL analysis described here. A mapping population consisting of 276 F_2 individuals was grown and phenotyped in the greenhouse during the summer of 2010. Additionally, nine plants of the inbred AC Sunset and nine selfed offspring of the *C. palaestinus* mapping parent were grown in the greenhouse alongside the mapping population to provide estimates of parental trait means under the same conditions as the mapping population. The mapping population and parental plants were transplanted into 12-inch tall treepots (Stuewe and Sons, Tangent, OR) and grown in a completely randomized fashion within a single greenhouse room. All pots were moved weekly throughout the duration of the study to minimize the effects of micro-environmental variations within the greenhouse.

Phenotyping. Plants were checked daily and dates were recorded for estimates of root growth rate and the initiation of flowering. Root growth rate was based on the number of days until the roots reached the bottom of each pot. Leaf size, shape, perimeter, and spininess were estimated and averaged across three leaves collected from each plant (the most recent fully expanded leaf, the leaf directly below the primary capitulum, and the longest rosette leaf) and

scanned for analysis with ImageJ v1.43u [32]). Spininess was measured using a modified version of the spine index, which was initially described in [33] as the number of spines on a leaf multiplied by the length of the longest spine on that leaf. Here, a “standardized spine index” was used, taken as a measure of the number of spines per centimeter of leaf margin multiplied by the length of the longest spine on that leaf.

Heads were bagged on the day of anthesis to prevent cross-pollination and potential seed loss. The height and diameter of the primary head and disc were measured using digital calipers on the day of first flowering of each plant. Stem height was measured as the length of the stem from the base of the plant to the base of the primary head. Fresh florets were collected from the third flowering head on the day that it opened and mature florets were collected from the primary head after all flowering had ceased. These florets were flash frozen in liquid nitrogen to preserve their pigments, which were later measured with a Gardner XL20 colorimeter (Bethesda, MD) using coordinates from CIE $L^*a^*b^*$ color space (in which L^* represents luminosity and a^* and b^* represent the coordinates of each pigment’s hue on the red/green and blue/yellow axes, respectively; [34, 35]). For each measurement, differences in hue due to differences in light intensity were controlled for by holding L^* constant at 30 units. Because floret color changes from yellow (anthesis) to red (senescence) in AC Sunset (but not in the wild progenitor), we recorded the magnitude of floret color change in each plant by calculating the difference between the a^* value at flowering and the a^* value at maturity. Smaller a^* values correspond to yellower flowers and larger values correspond to redder flowers, and b^* values changed marginally between these two colors.

Heads were harvested at physiological maturity (i.e. when the bracts were no longer green). Seven days after harvest, 12-16 achenes were sterilized using a 10% bleach solution and

planted at a 1.5 cm depth into small pots maintained in a growth chamber for the assessment of seed dormancy and viability. Plants with primary heads lacking sufficient seed set (N=119) were omitted from this analysis. Pots were monitored daily for up to 60 days and dates of seedling emergence were recorded and used to estimate the fraction of achenes that germinated within the 60 day window and calculate the average time to germination for each F₃ family.

At the conclusion of flowering, the number of senesced heads, the height (above the soil) of the lowest branching point, the number of internodes, and the length of the second internode were recorded for each plant. As the remaining heads reached maturity, selfed achenes were harvested, counted, weighed, and measured. Average achene weight was based on a random subset of 50 achenes. However, for plants that produced less than 50 achenes, the average achene weight was based on the total number of achenes.

Seed oil content and composition were estimated for all plants with sufficient seed set. These measurements were taken following previously established protocols (percent oil content [21, 36]; seed oil composition [24]). Briefly, a Bruker MQ20 minispec NMR analyzer (The Woodlands, TX) was used to determine percent oil content. The standard protocol was modified to accommodate measurements based on small seed sets by placing ca. one cm of tissue paper into the flat-bottomed tubes and adding ca. one cm of cleaned safflower seeds on top of the tissue paper. Percent oil content was estimated using a calibration curve using commercial safflower oil as a standard (Hollywood, Boulder, CO). Oil composition was determined based on gas chromatography of fatty acid methyl esters. A total of ten achenes from each plant were hand ground and fatty acids were extracted and then analyzed using an Agilent 6890N gas chromatograph (Santa Clara, CA).

SNP identification. Single nucleotide polymorphisms (SNPs) that differentiated the parents of our population were identified from expressed sequence tag (EST) and transcriptome data generated from each parent. For the cultivated mapping parent, we used the AC Sunset EST data produced via Sanger sequencing of cDNAs described in Chapman et al. [37]. For the wild mapping parent, we produced transcriptome sequence data as follows: RNA was extracted from mature leaves, bracts, florets, and developing ovules collected from a single *C. palaestinus* plant using a combined trizol (Invitrogen, Carlsbad, CA) RNeasy mini column (Qiagen, Valencia, CA) method. RNA extracted from each tissue type was pooled in equal proportions, normalized prior to 454 library preparation following the protocols described by Lai et al. [38] and Meyer et al. [39], sequenced, and assembled using MIRA v3.0.3 [40] (see Text A1).

The assembled sequences from the mapping parent assemblies were aligned to each other using Mosaik [41]. SNPs were identified using SAMtools [42] and run through the Illumina GoldenGate “assay design tool” (San Diego, CA; <http://illumina.com>), which identified SNPs free of other polymorphisms within 60 bp of the targeted SNP site and assigned a quality score predicting the success with which a SNP would be assayed. To facilitate downstream comparative genomic analyses, SNPs used in this study were preferentially selected from *Carthamus unigenes* with mapped homologs in the high-density sunflower “consensus” map [43]. A total of 384 SNPs meeting design requirements were targeted for genotyping, and a subset of these were validated via genetic mapping (see below).

Genotyping and map construction. Whole genomic DNA was extracted from leaf tissue from each F₂ plant as well as the mapping parents using a modified CTAB protocol [44]. DNA concentrations were estimated using the Quant-iT PicoGreen kit (Invitrogen) using a Biotek Synergy 2 plate reader (Winooski, VT). The Illumina GoldenGate Assay described above

was then used to genotype each sample on the BeadXpress Reader (Illumina) at the Georgia Genomics Facility. Allele calls were obtained using the Illumina GenomeStudio software v2011.1.

A genetic linkage map was constructed using MapMaker 3.0/EXP [45, 46]. Briefly, initial linkage groups (LGs) were identified using the “group” command with a minimum LOD score of 5.0 and a maximum frequency of recombination of 0.4 between adjacent markers. Preliminary map orders were determined using the “compare” command on a subset of markers within a LG and the remaining markers were placed using the “try” command. For each linkage group, marker orders were confirmed using the “ripple” command and the final marker orders, presented here, represent the most likely marker orders given the data.

Statistical analyses and QTL mapping. Histograms and trait means of the mapping population and mapping parents were plotted for visualization using the R Statistical Package [47]. Estimated parental trait values were further analyzed to test for significant differences using either Welch’s *t*-tests or, when trait distributions deviated significantly from normality (as determined by the Shapiro-Wilk test for normality), Wilcoxon signed-rank tests. Spearman correlation coefficients were calculated among all traits measured in the F₂ mapping population using the “hmisc” package [48] in R [47]. Significance was determined using the sequential Bonferroni method with $\alpha = 0.05$ [49].

QTL were identified using QTL Cartographer v1.17j [50, 51] following established approaches (e.g. [7, 8, 21, 24]). Briefly, composite interval mapping (CIM) was performed in ZmapQTL with a 10 cM window and a maximum of five background cofactors identified using SRmapQTL with forward/backward stepwise regression, and tests were performed at 2 cM intervals. Permutation thresholds ($\alpha = 0.05$ and 0.1) for declaring QTL significance were

estimated based on 1000 permutations for each trait [52, 53]. Secondary peaks were not considered as separate QTL unless there was a 2-LOD decline between adjacent peaks.

The results generated from CIM were then used as an initial model for multiple interval mapping (MIM), as implemented by MIMapQTL [54]. This analysis was used to confirm QTL identified via CIM. Following the authors' recommendations, the information criterion was set as $IC(k) = -2(\log(L) - kc(n)/2)$, where $c(n) = \log(n)$ was the penalty function and the threshold was set at 0. Epistasis was investigated at a genome-wide level using EPISTACY v2 [55] to test for interactions between all possible pairs of codominant markers that exhibited unique segregation patterns (i.e. redundant markers that showed identical segregation patterns were joined into a single haplotype to reduce the number of pairwise comparisons). As suggested by the author, significance was determined by dividing the comparison-wise error rate ($\alpha = 0.05$) by $g(g-1)/2$, where g is the haploid number of LGs in safflower ($n = 12$).

The mode of gene action for each QTL was estimated by dividing the dominance effect of the cultivar allele by its additive effect (d/a), such that cultivar alleles that are completely recessive have a value of -1 and those completely dominant have a value of +1. Following the cutoffs employed by Burke et al. [7], the mode of gene action of the cultivar allele at each locus was categorized as follows: underdominant $\leq -1.25 < \text{recessive} \leq -0.75 < \text{partially recessive} \leq -0.25 < \text{additive} < 0.25 \leq \text{partially dominant} < 0.75 \leq \text{dominant} < 1.25 \leq \text{overdominant}$.

Additionally, the magnitude of the effect of each QTL was considered to be “large” if the percentage of segregating phenotypic variance explained (PVE) was greater than 25%, “small” if the PVE was less than 10%, and “intermediate” if in between.

Comparative genomic analyses. In order to identify homologous loci between the safflower and sunflower genomes, all ESTs harboring safflower SNPs mapped in this study as

well as all loci from the 10,000 feature sunflower consensus map [43] were compared via BLAST to the lettuce genome, v4 [56]. As noted above, lettuce is a member of the Cichorioideae, which falls at an intermediate phylogenetic position between the Carduoideae and the Asteroideae. The use of the lettuce genome as an intermediary greatly simplified these analyses because it is the same ploidy level as safflower (though functionally diploid, sunflower is a paleopolyploid [tetraploid relative to safflower and lettuce] due to an ancient whole genome duplication at the base of the Heliantheae; [57]) and because it dramatically increased the number of mapped loci bridging the safflower and sunflower genomes (i.e. virtually all of the ESTs from which the safflower and sunflower markers were derived could be matched to corresponding sequences in the lettuce genome). The top two BLASTN hits with an e-value better than 1×10^{-6} were recorded and any sunflower and safflower contigs with a shared lettuce BLAST hit were considered homologous. We then surveyed the literature to identify all previously mapped sunflower domestication QTL. Because many of the markers used to map these QTL were included in the sunflower consensus map [43], it was possible to project the positions of these QTL onto that map for comparative QTL mapping. For instances in which the bounds of 1-LOD intervals could not be projected onto the consensus map (due to an absence of shared markers at the 1-LOD boundaries), we estimated the distance from shared markers within the 1-LOD interval to the boundaries based on relative map lengths. To determine the probability that instances of QTL colocalization were due to chance alone, we used the hypergeometric probability distribution function ('sampling without replacement'; [58]); as described in Paterson et al. [11] and Paterson [59] as follows:

$$p = \frac{\binom{l}{m} \binom{n-l}{s-m}}{\binom{n}{s}}$$

where n is the number of intervals that can be compared (e.g. genome size divided by average QTL size for a given trait), m is the number of colocalizing QTL, l is the total number of QTL in the larger sample, and s is the number of QTL in the smaller sample for a given trait.

RESULTS

Phenotypic Analyses. Of the 24 traits analyzed, 15 differed significantly between the mapping parents when grown alongside the mapping population (Table 2.1). Comparisons of the means and standard deviations of the mapping parent representatives to the F_2 trait distributions revealed transgressive segregation for the majority of the traits analyzed (i.e. there are F_2 individuals with trait values exceeding one standard deviation in either direction of the mapping parents; Figure 2.1). The most extreme examples of transgressive segregation were for traits related to vegetative growth: capitulum height, disc diameter, stem height, leaf size, and achene size. In other words, many of the F_2 plants and their achenes were larger than either of the mapping parents.

Approximately one fourth of the mapping population had florets that developed a deep red color at maturity. The flower color ratio did not differ significantly from a 3:1 distribution ($\chi^2 = 0.16$, $P = 0.6$; data not shown). This suggests that flower color variation in the mapping population is controlled by a single gene, with the ability to turn red being recessive.

Many of the traits under study were correlated within the F_2 mapping population (Table A1) in a way that was largely consistent with the observed parental trait combinations. Several of these correlations, however, were no longer significant after Bonferonni correction. Interestingly, the total number of selfed achenes produced was positively correlated with achene oil content ($\rho = 0.573$, $P < 0.001$) and stem height ($\rho = 0.420$, $P < 0.001$) and negatively correlated with

achene dimensions (achene weight: $\rho = -0.562$, $P < 0.001$; achene width: $\rho = -0.558$, $P < 0.001$; achene length: $\rho = -0.730$, $P < 0.001$).

Genetic Mapping. Of the 384 SNPs designed for the Illumina GoldenGate assay, 244 exhibited interpretable polymorphisms that could be used for genetic mapping (data not shown). The remaining 140 markers were omitted due to monomorphism in the mapping population, overly complex segregation patterns (presumably due to paralogy), or failure of the assay probes to hybridize with the target DNA. Although SNPs are typically scored as codominant markers, 26 of the SNPs included in this study were scored as dominant markers due to the lack of a clear distinction between one of the homozygote classes and the heterozygote class (these markers are flagged with an asterisk in Figure 2.2).

The 244 markers coalesced into 12 LGs, consistent with the haploid number of $n = 12$ chromosomes that has previously been reported for safflower (Figure 2.2; [30]). These LGs ranged in size from 30.7 to 105.3 cM (average = 71.5), with each group comprising 6 to 40 markers (average = 20.3). The total map length summed to 858.2 cM, resulting in an average intermarker distance of 3.7 cM (range = 0.0-39.6 cM).

The segregation patterns of 14 out of the 244 mapped markers deviated from Mendelian expectations (i.e. they exhibited significant segregation distortion after Bonferroni correction; [60–62]. Eleven of these loci were found in two distorted regions spanning 13 cM (on LG K) and 12 cM (on LG L; Figure 2.2). Distortion occurred in both directions: in some cases, the wild allele was overrepresented in the mapping population while in other cases, the cultivar allele was overrepresented. Two markers exhibited an underrepresentation of both homozygote classes, yielding a heterozygote excess. Within the two aforementioned distorted regions, the direction of deviation remained consistent, though the segregation of the markers on LG K were skewed in

the direction of the wild parent and those on LG L were in the direction of the cultivar parent. Further, the magnitude of the distortion was significantly greater on LG L ($P = 0.01$).

QTL Mapping. A total of 61 QTL were identified for 21 of the 24 traits studied (Table 2.2, Figure 2.2). Only LG F lacked QTL. Eight of the sixty-one QTL were marginally significant, having been identified at the $\alpha = 0.1$ permutation threshold during CIM; the remainder exceeded the $\alpha = 0.05$ permutation threshold and all of these QTL were confirmed via MIM. The 1-LOD confidence intervals for these QTL averaged 13.5 cM, ranging from 1.5 to 31.9 cM. Of the 21 traits, 16 traits had multiple QTL (range = 2-7). Nearly all mapped QTL 1-LOD intervals overlapped at least 1 of the 244 mapped markers, with the lone exception of a QTL for % oil that mapped between markers on LG L. One trait had two instances of antagonistic QTL on the same LG (capitulum height; on both LGs A and D), where the cultivar-derived allele for one QTL conferred a cultivar-like phenotype and the cultivar-derived allele for the other QTL conferred a wild-like phenotype. The majority of all QTL identified mapped to one of seven QTL clusters on seven different chromosomes, with each cluster harboring three to twelve QTL. Also, 23 of the QTL identified in this study mapped to 3 of 7 genomic regions that exhibited marker clustering (LGs C, E, and H).

The PVE for the identified QTL ranged from 4.2% to 63.4% (Table 2.2), with the majority of the QTL having small effects (PVE <10%). There were 13 QTL with intermediate effects and just 2 QTL with large effects (spiniess and flower color, 32.7% and 63.4% respectively). For traits that differed significantly between the mapping parents, it was possible to investigate whether the respective QTL had allelic effects in the expected direction (i.e. where the cultivar allele produced a more cultivar-like phenotype). Examination of the 44 QTL identified for the 15 traits with means differing significantly in the parents revealed 12 QTL for 9

traits conferring phenotypes in the “wrong” direction. For the remaining three traits for which multiple QTL were identified, the effects of all QTL were consistent with the observed parental trait differences (average leaf size, spininess, and number of internodes; Table 2.2). Interestingly, the majority of the seed width QTL had cultivar alleles conferring a wild-like phenotype. The mode of gene action of each QTL ranged from -21.7 (average seed weight) to 7.3 (total number of seeds per plant), though the majority fell in the -1 to 1.25 range (average 0.72 ± 0.40). Traits with overdominant QTL included flowering time, total number of seeds, seed weight, seed length, seed width, and seed oil content; those with underdominant QTL included leaf size, number of capitula, and seed weight.

The genome-wide scan for epistasis detected a total of 105 significant epistatic interactions at the $\alpha < 0.05$ level (after correcting for multiple comparisons) for 19 traits (Table A2). Note that multiple interactions detected among loci on the same linkage groups were counted as a single interaction, accounting for the non-independence of these loci due to linkage. Not all traits with QTL were found to be influenced by epistatic loci and some traits only appeared to be influenced by loci with epistatic effects. On average, 5.5 ± 1.27 interactions were identified for the traits with epistasis, ranging from 1 to 19 significant interactions per trait. Additive-by-additive and additive-by-dominant interactions comprised the majority of the interactions (40 and 42 interactions, respectively). Upon closer inspection of the EPISTACY results, we noticed two traits that had QTL $\times$ QTL interactions: leaf shape (LG H $\times$ L, additive-by-additive) and seed length (LG K $\times$ C, additive-by-dominant and dominant-by-dominant).

QTL Colocalization. Though we were somewhat limited in our ability to directly identify instances of QTL colocalization between safflower and sunflower due to a limited number of shared markers between the maps, the lettuce genome sequence served as an effective

intermediary in bridging the gaps among these maps. We ultimately identified 15 QTL corresponding to 10 different domestication-related traits in safflower (1 to 3 QTL per trait) that colocalized with previously identified QTL in sunflower (Figure 2.3, Figure A1; [7, 8, 22–25]). Specifically, we saw highly significant evidence of QTL colocalization for % oil content (LG I; $P = 0.001$; Figure 2.3; [23]) as well as marginally significant evidence of QTL colocalization for days to flower (LG I; $P = 0.08$), achene weight (LGs C and H; $P = 0.07$), % linoleic acid (LG H; $P = 0.08$), and achene width (LG C; $P = 0.09$). The significance of the remaining traits that exhibited QTL colocalization between safflower and sunflower was less compelling (P -values ranged from 0.12 to 0.21 for disc diameter, number of heads, number of selfed seeds, achene length, and % oleic acid). Nonetheless, when applied across all ten traits with evidence of QTL colocalization, Fisher’s combined probability test was highly significant ($P = 0.0001$) and remained significant even when excluding the highly significant result for % oil content ($P = 0.004$).

DISCUSSION

Genetic architecture of safflower domestication. Our results indicate that domestication-related traits in safflower are largely controlled by multiple genes of small to moderate effect. Only two traits (flower color and spininess) had “major” QTL (i.e. $PVE > 25\%$). As such, the genetic architecture of safflower domestication is similar to that of sunflower, which is the only other crop in which QTL analyses have revealed such a paucity of major effect QTL [7, 8, 24]. More commonly, QTL mapping has suggested that domestication-related traits are conditioned by a relatively small number of loci with large phenotypic effects (reviewed in [63, 64]). However, population genomic analyses have recently shown that much larger numbers of genes are typically under selection during crop domestication and/or improvement ([65, 66];

reviewed in [67, 68]). These findings indicate that the genetic architecture of domestication traits is likely to be complex, even for crops in which initial QTL-based approaches have suggested otherwise.

The single largest QTL identified in our investigation of safflower explained 63.4% of the phenotypic variance in flower color. Further, our observation of 3:1 segregation of flower color suggests that the difference in the production of carthamine (the quinochalcone pigment responsible for the production of red florets) within our mapping population is due to one or multiple alleles at a single locus. Earlier crossing studies of safflower suggested that multiple genes influence flower color [69, 70] and more recent studies have shown that there are at least two interacting genes differentiating orange and yellow florets [71]. The fact that we identified just a single QTL suggests that the mapping parents in our population differ primarily in terms of the production of carthamine as opposed to the other floral pigments. More generally, the findings of single, large effect QTL for the presence of a particular floral pigment as well as for leaf spines are consistent with the views of Gottlieb [72], who argued that presence/absence characters and major or structural differences in plants are commonly controlled by just one or two genes. In contrast, differences in more continuously varying traits tend to be controlled by multiple genes, as we have also found to be true for safflower.

Map features and QTL distribution. All but 14 of the markers analyzed exhibited Mendelian segregation ratios. Interestingly, 11 of these 14 markers occurred within two distorted regions spanning 12 cM and 13 cM on LGs K and L, respectively. Though the cause of this distortion remains unknown, it may be due to genomic divergence between the mapping parents in these regions. In this light, it is worth noting the distorted region on LG K harbors QTL for seed-related traits and the distorted region on LG L harbors the large effect QTL for spininess as

well as other QTL for internode length, number of internodes, disc diameter, capitulum height, and leaf shape. Though distortion within 40 cM of a QTL can make it difficult to detect dominant QTL in F_2 mapping populations, it can aid the detection of additive QTL [73, 74].

In terms of overall marker distribution, we observed numerous tight clusters across multiple LGs. These clusters could be a byproduct of an uneven distribution of genes across the genome (recall that all SNPs employed in this study were selected from transcribed sequences), chromosomal rearrangements that differentiate the mapping parents, and suppress recombination in affected regions (though the F_1 hybrids did not seem to suffer reduced fertility), or – perhaps more likely – the suppression of recombination in and near centromeres. Clustering has also been reported in other genetic maps generated from *C. tinctorius* x *C. tinctorius* crosses, though to a lesser extent [75, 76].

We likewise observed a number of QTL clusters across the genome. In some cases, these clusters appeared to co-occur with the aforementioned marker clusters, suggesting that they may be mapping to gene dense regions or to regions with suppressed recombination. It has been argued that species with clustered genes underlying domestication-related traits may have been inherently easier to domesticate [77]. In this context, it is worth noting that clustering of domestication-related loci has previously been documented in a number of other crops, including maize [5], common bean [6], pearl millet [78], pepper [79] and sunflower [7] (reviewed in [80]). While Pernès [81] predicted that the linkage of domestication genes can aid cross-pollinated crops in maintaining trait complexes that comprise the domestication syndrome, and further modeling has supported this prediction [82], empirical studies (including the present study) have indicated that these QTL clusters are also found in highly selfing crops [6, 83–85]. It is possible that these crops are more allogamous than they seem, or perhaps that increased allogamy was

predominant earlier during domestication in order to “assemble the domestication syndrome” [77].

Transgressive segregation. In general terms, transgressive segregation can be produced by complementary gene action, overdominance, and/or epistasis. The former, in which parents possess alleles with opposing (i.e. antagonistic) effects and hybridize to generate offspring carrying an excess of alleles with effects in the same direction [86, 87], has been implicated as the most common cause of this phenomenon (summarized in [87, 88]). Consistent with this view, traits with the most extreme transgressive segregation in our population (capitulum height, disc diameter, achene length and width, and achene oil content), were conditioned by multiple QTL (two to seven per trait) and, in many cases, these traits were conditioned by alleles with antagonistic parental effects. However, we also detected evidence of overdominant QTL effects for three of the traits exhibiting transgressive segregation (total seeds, achene length, achene width, and seed weight). Therefore, overdominance cannot be discounted. Finally, many of these same traits were influenced by multiple genetic interactions, suggesting that epistasis could have played a role in producing the observed transgressive segregation (though instances of epistasis were not limited to traits exhibiting transgressive segregation).

QTL colocalization. The colocalization of QTL for individual traits observed between safflower and sunflower supports the view that selection may, in some cases, have acted on the same genes during the independent domestication of safflower and sunflower. While additional work, including fine-mapping, positional cloning, and functional analyses, will be required to establish with certainty that the same underlying genes are responsible for these instances of colocalization, our findings are consistent with parallel trait evolution having been driven by parallel genotypic changes, as has been documented in other animal and plant systems. These

include the evolution of red and green color vision in multiple vertebrate species [89] (reviewed in [90]), the ability of bats, dolphins, and whales to echolocate [91, 92] (reviewed in [90]), herbicide resistance in maize and cocklebur [93] (reviewed in [94]), and the glutinous phenotype in rice [95–97], Chinese waxy maize [98], and foxtail millet [99] (reviewed in [64]).

Interestingly, some of the traits with evidence of parallel genotypic evolution in safflower and sunflower have exhibited similar patterns in other crops. These cases include seed weight in cowpea and mung bean [9] and across maize, rice, and sorghum [11, 100] (reviewed in [94]) as well as fruit length and shape in eggplant and tomato [12]. Going forward, an improved understanding of the genes underlying parallel trait transitions will provide key insights into the repeatability of evolution, helping us to better predict the phenotypic effects of genotypic changes across a broad array of crops.

REFERENCES

1. Darwin C: *The Variation of Animals and Plants Under Domestication*. 1868.
2. Ross-Ibarra J, Morrell PL, Gaut BS: **Plant domestication, a unique opportunity to identify the genetic basis of adaptation**. *Proceedings of the National Academy of Sciences* 2007, **104**:8641–8648.
3. Harlan JR: *Crops and Man*. 2nd ed. Madison, WI: American Society of Agronomy; 1992.
4. Doebley J: **Molecular evidence and the evolution of maize**. *Economic Botany* 1990, **44**:6–27.
5. Doebley J, Stec A: **Genetic analysis of the morphological differences between maize and teosinte**. *Genetics* 1991, **129**:285–295.
6. Koinange EMK, Singh SP, Gepts P: **Genetic control of the domestication syndrome in common bean**. *Crop Science* 1996, **36**:1037–1045.
7. Burke JM, Tang S, Knapp SJ, Rieseberg LH: **Genetic analysis of sunflower domestication**. *Genetics* 2002, **161**:1257–1267.

8. Wills DM, Burke JM: **Quantitative trait locus analysis of the early domestication of sunflower.** *Genetics* 2007, **176**:2589–2599.
9. Fatokun CA, Menancio-hautea DI, Danesh D, Young ND: **Evidence for orthologous seed weight genes in cowpea and mung bean based on RFLP mapping.** *Genetics* 1992, **132**:841–846.
10. Hu FY, Tao DY, Sacks E, Fu BY, Xu P, Li J, Yang Y, McNally K, Khush GS, Paterson A, Li Z-K: **Convergent evolution of perenniality in rice and sorghum.** *Proceedings of the National Academy of Sciences* 2003, **100**:4050–4054.
11. Paterson A, Lin YR, Li Z, Schertz KF, Doebley JF, Pinson SR, Liu SC, Stansel JW, Irvine JE: **Convergent domestication of cereal crops by independent mutations at corresponding genetic loci.** *Science* 1995, **269**:1714–8.
12. Doganlar S, Frary A, Daunay M-C, Lester RN, Tanksley SD: **Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant.** *Genetics* 2002, **161**:1713–1726.
13. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F, Sudhakar D, Christou P, Centre JI, Snape JW, Gale MD, Harberd NP: **“Green revolution” genes encode mutant gibberellin response modulators.** *Nature* 1999, **400**:256–261.
14. Patel J, Narayana G: **Chromosome numbers in safflower.** *Current Science* 1935, **4**:412.
15. Weiss EA: **Safflower.** In *Oilseed crops*. 2nd ed. London: Blackwell Science Ltd; 2000:93–129.
16. Kar G, Kumar A, Martha M: **Water use efficiency and crop coefficients of dry season oilseed crops.** *Agricultural Water Management* 2007, **87**:73–82.
17. Merrill SD, Tanaka DL, Hanson JD: **Root length growth of eight crop species in haplustoll soils.** *Soil Science Society of America Journal* 2002, **66**:913–923.
18. Knowles P, Ashri A: **Safflower -- *Carthamus tinctorius* (Compositae).** In *Evolution of Crop Plants*. Edited by Smartt, J. and Simmonds NW. Harlow, UK: Longman; 1995:47–50.
19. **Angiosperm Phylogeny Website** [<http://www.mobot.org/MOBOT/research/APweb/>]
20. Funk VA, Bayer RJ, Keely S, Chan R, Watson L, Gemeinholzer B, Schilling E, Panero JL, Baldwin BG, Garcia-Jacas N, Susanna A, Jansen RK: **Everywhere but Antarctica: using a supertree to understand the diversity and distribution of the Compositae.** *Biologische Skrifter* 2005, **55**:343–374.

21. Burke JM, Knapp SJ, Rieseberg LH: **Genetic consequences of selection during the evolution of cultivated sunflower.** *Genetics* 2005, **171**:1933–1940.
22. Dechaine JM, Burger JC, Chapman MA, Seiler GJ, Brunick R, Knapp SJ, Burke JM: **Fitness effects and genetic architecture of plant-herbivore interactions in sunflower crop-wild hybrids.** *New Phytologist* 2009, **184**:828–841.
23. Dechaine JM, Burger JC, Burke JM: **Ecological patterns and genetic analysis of post-dispersal seed predation in sunflower (*Helianthus annuus*) crop-wild hybrids.** *Molecular Ecology* 2010, **19**:3477–3488.
24. Wills DM, Abdel-Haleem H, Knapp SJ, Burke JM: **Genetic architecture of novel traits in the hopi sunflower.** *The Journal of Heredity* 2010, **101**:727–736.
25. Chapman MA, Burke JM: **Evidence of selection on fatty acid biosynthetic genes during the evolution of cultivated sunflower.** *Theoretical and Applied Genetics* 2012, **125**:828–841.
26. Mundel H, Huang HC, Braun JP, Kiehn F: **AC Sunset safflower.** *Canadian Journal of Plant Science* 1996, **76**:469–471.
27. Chapman MA, Burke JM: **DNA sequence diversity and the origin of cultivated safflower (*Carthamus tinctorius* L.; Asteraceae).** *BMC Plant Biology* 2007, **7**:60.
28. Van Zeist W, Rooijen Waterbolk-Van W: **Two interesting floral finds from third millenium BC Tell Hammam et-Turkman, northern Syria.** *Vegetation History and Archaeobotany* 1992, **1**:157–161.
29. Ashri A, Rudich J: **Unequal reciprocal natural hybridization rates between two *Carthamus* L. species.** *Crop Science* 1965, **5**:190–191.
30. Ashri A, Knowles PF: **Cytogenetics of safflower (*Carthamus* L.) species and their hybrids.** *Agronomy Journal* 1960, **52**:11–17.
31. Hanelt P: **Monographische ubersicht der gattung *Carthamus* L. (Compositae).** *Feddes Repertorium* 1963, **67**:41–180.
32. Schneider CA, Rasband WS, Eliceiri KW: **NIH Image to ImageJ: 25 years of image analysis.** *Nature Methods* 2012, **9**.
33. Claassen CE, Ekdahl WG, Severson GM: **The estimation of oil percentage in safflower seed and the association of oil percentage with hull and nitrogen percentages, seed size, and degree of spininess of the plant.** *Agronomy Journal* 1950, **42**:478–482.
34. **Colorimetric fundamentals: CIE 1976 L*a*b* (CIELAB)**
[<http://industrial.datacolor.com/support/wp-content/uploads/2013/01/Color-Fundamentals-Part-II.pdf>]

35. **Technical report: colorimetry** [<https://law.resource.org/pub/us/cfr/ibr/003/cie.15.2004.pdf>]
36. Tang S, Leon A, Bridges WC, Knapp SJ: **Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower.** *Crop Science* 2006, **46**:721–734.
37. Chapman MA, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore RW, Tang S, Knapp SJ, Burke JM: **Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (*Carthamus tinctorius* L.).** *Theoretical and Applied Genetics* 2009, **120**:85–91.
38. Lai Z, Kane NC, Kozik A, Hodgins KA, Dlugosch KM, Barker MS, Matvienko M, Yu Q, Turner KG, Pearl SA, Bell GDM, Zou Y, Grassa C, Guggisberg A, Adams KL, Anderson J V, Horvath DP, Kesseli R V, Burke JM, Michelmore RW, Rieseberg LH: **Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression.** *American Journal of Botany* 2012, **99**:209–218.
39. Meyer E, Aglyamova G, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz M: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GS-Flx.** *BMC Genomics* 2009, **10**:219.
40. Chevreux B, Wetter T, Suhai S: **Genome sequence assembly using trace signals and additional sequence information.** In *German Conference on Bioinformatics*; 1999.
41. Lee W-P, Stromberg M, Ward A, Stewart C, Garrison E, Marth G: **MOSAIK: A hash-based algorithm for accurate next-generation sequencing read mapping.** *arXiv:13091149* 2013.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.
43. Bowers JE, Nambeesan S, Corbi J, Barker MS, Rieseberg LH, Knapp SJ, Burke JM: **Development of an ultra-dense genetic map of the sunflower genome based on single-feature polymorphisms.** *PLoS ONE* 2012, **7**:e51360.
44. Doyle JL, Doyle JR: **A rapid DNA isolation procedure for small quantities of fresh leaf tissue.** *Phytochemistry Bulletin* 1987, **19**:11–15.
45. Lander E, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L: **MAPMAKER : An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations.** *Genomics* 1987, **1**:174–181.
46. Lincoln SE, Lander ES: **Systematic detection of errors in genetic linkage data.** *Genomics* 1992, **14**:604–610.
47. **R: A language and environment for statistical computing** [<http://www.r-project.org>]
48. Harrell Jr FE: **Hmisc: Harrell Miscellaneous.** 2012.

49. Rice WR: **Analyzing tables of statistical tests.** *Evolution* 1989, **43**:223–225.
50. Basten CJ, Weir BS, Zeng Z-B: **QTL Cartographer.** 2004.
51. Basten C, Weir B, Zeng Z-B: **Zmap-a QTL cartographer.** In *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software.* Edited by Smith C, Gavora J, Benkel B, Chesnais J, Fairfull W, Gibson J, Kennedy B, Burnside E. Guelph, Ontario, Canada: Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production; 1994:65–66.
52. Churchill GA, Doerge RW: **Empirical threshold values for quantitative trait mapping.** *Genetics* 1994, **138**:963–971.
53. Doerge RW, Churchill GA: **Permutation tests for multiple loci affecting a quantitative character.** *Genetics* 1996, **142**:285–294.
54. Kao CH, Zeng ZB, Teasdale RD: **Multiple interval mapping for quantitative trait loci.** *Genetics* 1999, **152**:1203–1216.
55. Holland JB: **EPISTACY: A SAS Program for Detecting Two-Locus Epistatic Interactions Using Genetic Marker Information.** *Journal of Heredity* 1998, **89**:374–375.
56. **Lettuce Genome Resource** [<http://lgr.genomecenter.ucdavis.edu/>]
57. Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH: **Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years.** *Molecular Biology and Evolution* 2008, **25**:2445–2455.
58. Larsen RJ, Marx ML: *An Introduction to Probability and Its Applications.* Englewood Cliffs, NJ: Prentice Hall Inc.; 1985.
59. Paterson AH: **What has QTL mapping taught us about plant domestication?** *New Phytologist* 2002, **154**:591–608.
60. Sandler L, Novitski E: **Meiotic drive as an evolutionary force.** *American Naturalist* 1957, **91**:105–110.
61. Sandler L, Golic K: **Segregation distortion in *Drosophila*.** *Trends in Genetics* 1985, **1**:181–185.
62. Lyttle TW: **Segregation distorters.** *Annual Review of Genetics* 1991, **25**:511–557.
63. Burger JC, Chapman MA, Burke JM: **Molecular insights into the evolution of crop plants.** *American Journal of Botany* 2008, **95**:113–122.

64. Gross BL, Olsen KM: **Genetic perspectives on crop domestication.** *Trends in Plant Science* 2010, **15**:529–537.
65. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS: **The effects of artificial selection on the maize genome.** *Science* 2005, **308**:1310–4.
66. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright R a, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner R a, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J: **Comparative population genomics of maize domestication and improvement.** *Nature Genetics* 2012, **44**:808–811.
67. Morrell PL, Buckler ES, Ross-Ibarra J: **Crop genomics: advances and applications.** *Nature Reviews Genetics* 2012, **13**:85–96.
68. Wallace JG, Larsson SJ, Buckler ES: **Entering the second century of maize quantitative genetics.** *Heredity* 2013(January):1–9.
69. Narkhede BN, Deokar AB: **Inheritance of corolla colour in safflower.** *Journal of Maharashtra Agricultural Universities* 1986, **11**:278–281.
70. Rao M: **Inheritance of characters in safflower--*Carthamus tinctorius* L.** *Madras Agricultural Journal* 1943, **31**:141–148.
71. Pahlavani MH, Mirlohi AF, Saeidi G: **Inheritance of flower color and spininess in safflower (*Carthamus tinctorius* L.).** *Journal of Heredity* 2004, **95**:265–267.
72. Gottlieb LB: **Genetics and morphological evolution in plants.** *The American Naturalist* 1984, **123**:681–709.
73. Xu S: **Quantitative trait locus mapping can benefit from segregation distortion.** *Genetics* 2008, **180**:2201–2208.
74. Zhang L, Wang S, Li H, Deng Q, Zheng A, Li S, Li P, Li Z, Wang J: **Effects of missing marker and segregation distortion on QTL mapping in F2 populations.** *Theoretical and Applied Genetics* 2010, **121**:1071–82.
75. Mayerhofer R, Archibald C, Bowles V, Good AG: **Development of molecular markers and linkage maps for the *Carthamus* species *C. tinctorius* and *C. oxyacanthus*.** *Genome* 2010, **53**:266–276.
76. Hamdan YA, García-Moreno MJ, Fernández-Martínez JM, Velasco L, Pérez-Vich B: **Mapping of major and modifying genes for high oleic acid content in safflower.** *Molecular Breeding* 2012, **30**:1279–1293.

77. Gepts P: **Crop domestication as a long-term selection experiment.** *Plant Breeding Reviews* 2004, **24**:1–44.
78. Poncet V, Lamy F, Devos KM, Gale MD, Sarr A, Robert T: **Genetic control of domestication traits in pearl millet (*Pennisetum glaucum* L., Poaceae).** *Theoretical and Applied Genetics* 2000, **100**:147–159.
79. Chaim A Ben, Paran I, Grube RC, Jahn M, Van Wijk R, Peleman J: **QTL mapping of fruit-related traits in pepper (*Capsicum annuum*).** *Theoretical and Applied Genetics* 2001, **102**:1016–1028.
80. Poncet V, Sarr TRA, Gepts P: **Quantitative trait locus analyses of the domestication syndrome and domestication process.** In *Encyclopedia of Plant and Crop Science*. New York: Marcel Dekker, Inc.; 2004:1069–1074.
81. Pernes J: **La genétique de la domestication des cereales.** *La Recherche* 1983, **14**:910–919.
82. Le Thierry D'Ennequin, M. Toupance B, Robert T, Godelle B, Gouyon PH: **Plant domestication: a model for studying the selection of linkage.** *Journal of Evolutionary Biology* 1999, **12**:1138–1147.
83. Claassen CE: **Natural and controlled crossing in safflower, *Carthamus tinctorius* L.** *Agronomy Journal* 1950, **42**:381–384.
84. Ibarra-Perez FJ, Ehdaie B, Waines JG: **Estimation of outcrossing rate in common bean.** *Crop Science* 1997, **37**:60–65.
85. Verdugo-Hernández S, Reyes-Luna R, Oyama K: **Genetic structure and differentiation of wild and domesticated populations of *Capsicum annuum* (Solanaceae) from Mexico.** *Plant Systematics and Evolution* 2001, **226**:129–142.
86. DeVicente MC, Tanksley SD: **QTL analysis of transgressive segregation in an interspecific tomato cross.** *Genetics* 1993, **134**:585–596.
87. Rieseberg LH, Archer MA, Wayne RK: **Transgressive segregation, adaptation and speciation.** *Heredity* 1999, **83**:363–372.
88. Rieseberg LH, Widmer A, Arntz AM, Burke JM: **The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations.** *Philosophical Transactions of the Royal Society of London B* 2003, **358**:1141–1147.
89. Yokoyama S, Radlwimmer FB: **The molecular genetics and evolution of red and green color vision in vertebrates.** *Genetics* 2001, **158**:1697–1710.
90. Christin P-A, Weinreich DM, Besnard G: **Causes and evolutionary significance of genetic convergence.** *Trends in Genetics* 2010, **26**:400–405.

91. Li Y, Liu Z, Shi P, Zhang J: **The hearing gene Prestin unites echolocating bats and whales.** *Current Biology* 2010, **20**:R55–R56.
92. Liu Y, Cotton J a, Shen B, Han X, Rossiter SJ, Zhang S: **Convergent sequence evolution between echolocating bats and dolphins.** *Current Biology* 2010, **20**:R53–R54.
93. Bernasconi P, Woodworth AR, Rosen BA, Subramanian M, Siehl DL: **A naturally occurring point mutation confers broad range tolerance to herbicides that target acetolactate synthase.** *Journal of Biological Chemistry* 1995, **270**:17381–17385.
94. Wood TE, Burke JM, Rieseberg LH: **Parallel genotypic adaptation: when evolution repeats itself.** *Genetica* 2005, **123**:157–170.
95. Hirano HY, Eiguchi M, Sano Y: **A single base change altered the regulation of the Waxy gene at the posttranscriptional level during the domestication of rice.** *Molecular Biology and Evolution* 1998, **15**:157–170.
96. Olsen KM, Purugganan MD: **Molecular evidence on the origin and evolution of glutinous rice.** *Genetics* 2002, **162**:941–950.
97. Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD: **Selection under domestication: evidence for a sweep in the rice waxy genomic region.** *Genetics* 2006, **173**:975–83.
98. Fan L, Quan L, Leng X, Guo X, Hu W, Ruan S, Ma H, Zeng M: **Molecular evidence for post-domestication selection in the Waxy gene of Chinese waxy maize.** *Molecular Breeding* 2008, **22**:329–338.
99. Fukunaga K, Kawase M, Kato K: **Structural variation in the Waxy gene and differentiation in foxtail millet [*Setaria italica* (L.) P. Beauv.]: implications for multiple origins of the waxy phenotype.** *Molecular Genetics and Genomics* 2002, **268**:214–222.
100. Li Q, Li L, Yang X, Warburton ML, Bai G, Dai J, Li J, Yan J: **Relationship, evolutionary fate and function of two maize co-orthologs of rice GW2 associated with kernel size and weight.** *BMC Plant Biology* 2010, **10**:143.

TABLES

Table 2.1. Average trait values of mapping parents.

Trait	<i>Carthamus palaestinus</i> (progenitor)	<i>Carthamus tinctorius</i> (cultivated safflower)
Rooting rate (cm day ⁻¹)	3.24	2.41
Average leaf size (cm ²)	7.352	12.3
^x Average leaf roundness	0.559	0.412
Spininess (^y spine index/leaf perimeter)	0	19.031
Days to flower	33.29	31
Primary capitulum height (mm)	18.97	21.13
Primary disc diameter (mm)	16.94	16.76
Number of heads	8.78	8.63
Flower color (Lab color space a* units)	5.47	16.88
Stem height (cm)	32.94	31.36
Number of internodes	19.11	12.33
Internode length (cm)	1.74	2.56
Lowest branch height (percent up stem)	72	47
Number of selfed seed	12.44	68.59
Achene weight (mg)	33.8	31.2
Achene length (mm)	6.23	6.63
Achene width (mm)	3.63	3.38
Seed viability (percent)	75	84.9
Seed dormancy (average number of days until germination)	10.95	4.31
Seed oil (percent)	21.35	26.29
Palmitic acid (percent)	6.97	6.78
Stearic acid (percent)	2.79	2.57
Oleic acid (percent)	26.69	12.72
Linoleic acid (percent)	63.55	77.93

Bold indicates trait values that are significantly different from one another (t-test, $p < 0.05$)

^xAverage leaf roundness = $4 \times [\text{leaf area}/(\pi \times (\text{major leaf axis})^2)]$, where values closer to 1 represent circular shapes and values closer to 0 represent oblong shapes

^yspine index = number of spines x length of longest leaf spine (in mm)

Table 2.2. Quantitative trait locus (QTL) positions, modes of gene action, and magnitudes of effect for 19 out of the 24 traits studied.

Trait	Linkage group	Position ^a	Nearest marker	1-LOD interval ^b	Additive effect ^{c, d}	Dominance ratio	PVE
Average leaf size	B	52.1	B353	32.6-60.1	0.44	-3.69	8.7
	H	8.6	H113	0.9-12.5	1.26	0.69	9.9
Average leaf roundness	D	10	D378	4.6-21.3	<u>0.02</u>	-0.05	10.4
	G	56.3	G154	50.3-62.3	-0.02	-0.36	13.1
	H	4.3	H40	0-8.6	-0.02	-0.37	7.7
	L	105.3	L116	99.7-105.3	-0.02	0.82	4.5
Spininess	E	48.2	E190	40.1-53.9	1.74	0.13	4.5
	H	5.4-6.1	H327	0.9-16.5	2.93	-0.21	14.4
	L	105.3	L116	103.3-105.3	5.92	0.65	32.7
Days to flower	D	25.3	D271	15.3-35.3	-0.73	0.06	11.9
	H	8.8	H113	6.7-18.5	<u>0.42</u>	1.25	5.6
	I	35.5	I253	17.9-49.5	-0.49	-0.58	6.4
Primary capitulum height	A	24	A69	16.0-33.0	<u>-0.67</u>	0	8.7
	A	62.2	A117	58.8-75.0	0.39	-0.8	4.5
	D	0.01	D129	0.0-10.0	<u>-0.61</u>	-0.23	8.7
	D	68.3	D275	49.7-81.6	0.43	0.11	4.5
	H	2.1	H312	0.9-3	0.66	0.57	9.5
	I	41.5	I276	29.5-53.5	<u>-0.64</u>	-0.14	9.9
	L	101.7	L333	97.7-105.3	0.66	1.08	6.1
Primary disc diameter	A	66.8	A199	64.2-73.0	-0.60	-0.46	9.2
	H	12.5	H113	6.7-18.5	<u>0.76</u>	0.25	12.3
	I	35.5	I253	22.5-47.5	-0.60	0.18	9.6
	L	101.7	L333	95.7-105.3	<u>0.72</u>	0.15	8.2
Number of heads	H	3*	H76	0.0-18.5	-0.56	-1.4	4.8
Flower color	D	1.3	D234	0.0-2.6	3.70	-0.79	63.4
			E201,				
Stem height	E	37.2	E359	22.0-44.8	2.18	-0.01	6.9
	H	6.7	H130	0.9-10.5	2.31	0.42	7.8
	I	53.2	I276	37.5-57.2	-1.88	0.52	6.3
Number of internodes	C	47.5*	C200	43.1-50.3	-0.63	-0.44	4.4
	L	105.3	L116	101.7-105.3	-1.46	0.07	15.9

Internode length	A	41.8	A245	26.0-49.2	<u>-0.14</u>	0.43	7.6
	E	43.8	E354	40.1-48.2	0.13	0.32	4.6
			L219,				
	L	40.5*	L339	22.5-49	0.10	0.86	4.2
Lowest branch height	L	105.3	L116	99.7-105.3	0.17	0.1	6.7
	G	44.9	G100	36.9-60.3	-0.06	-0.22	5.9
Number of selfed seed	C	42.6*	C278	33.0-44.2	<u>-2.17</u>	7.26	4.2
	H	7.3	H255	0.0-18.5	15.89	-0.14	7.6
	I	13.9	I175	5.5-22.5	13.71	0.29	6.9
Achene weight	C	42.6	C278	39.3-43.5	-0.33	-21.68	11
	H	7.8*	H231	2.1-18.5	<i>1.71</i>	2.51	4.4
	I	0.0	I111	0-3.5	-4.96	0.73	13.1
	K	37.6	K35	34.8-47.6	4.67	0.06	8.2
Achene length	C	41.3	C120	40.9-42.7	0.08	5.004	10.4
	D	62.4*	D213	60.4-68.3	0.15	0.813	5.2
	I	17.9	I223	5.9-21.9	<u>-0.28</u>	0.155	12.0
	K	37.6	K35	36.5-45.6	0.29	-0.349	8.2
Achene width	C	42.6	C278	33.0-43.5	<u>0.09</u>	2.69	8.6
	I	0.0	I111	0.0-1.5	-0.21	0.8	15.3
	J	11.1	J232	0.0-11.1	<u>0.13</u>	0.37	5.1
	K	37.6	K35	36.5-45.6	<u>0.18</u>	-0.11	6.8
Seed dormancy	E	48.6	E190	43.8-53.9	-0.47	0.4	9.0
Seed oil	I	3.9*	I203	0.0-11.9	1.61	0.81	6.4
	I	65.4	I92	55.2-71.4	1.67	1.04	10.6
	J	12.2	J232	11.1-14.2	<u>-1.73</u>	0.21	7.2
	L	75.6	L221	67.6-83.6	2.65	1.35	23.2
Palmitic acid	E	46.4	E140	45.8-53.9	0.27	0.6	7.5
Oleic acid	C	50.3	C98	31.0-58.1	-1.33	-0.34	6.4
	G	29.8*	G26	17.4-37.0	<i>1.28</i>	0.71	6.3
	H	5.4-6.1	H327	2.1-18.5	-1.55	-0.66	11.0
Linoleic acid	G	31.8	G110	15.4-32.9	-1.69	0.67	8.6
	H	5.4-6.1	H327	0.9-18.5	<i>1.47</i>	-0.78	8.7

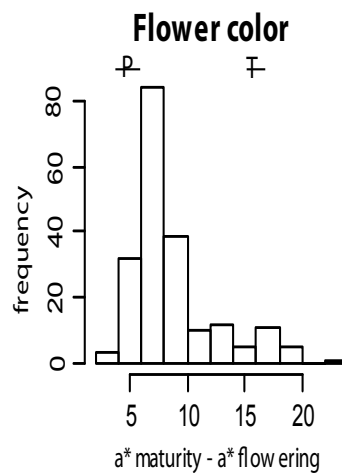
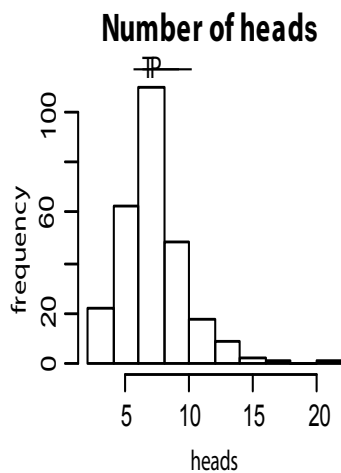
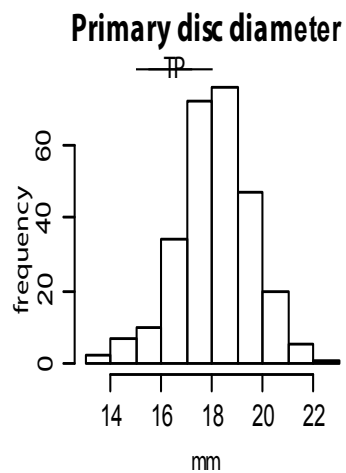
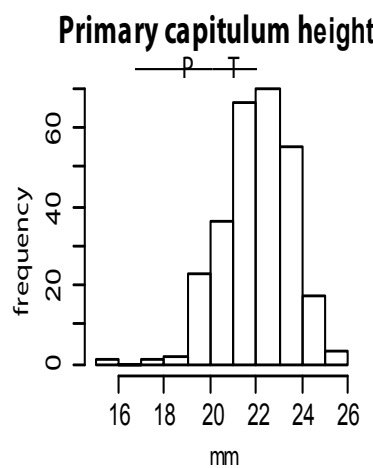
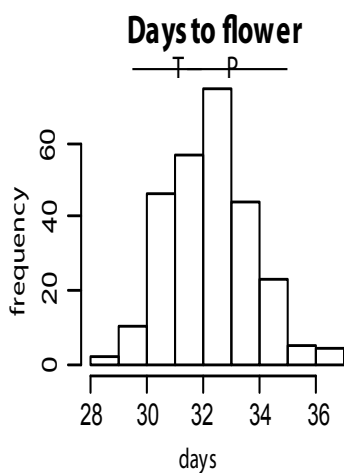
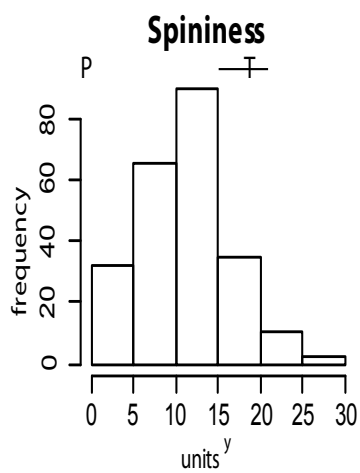
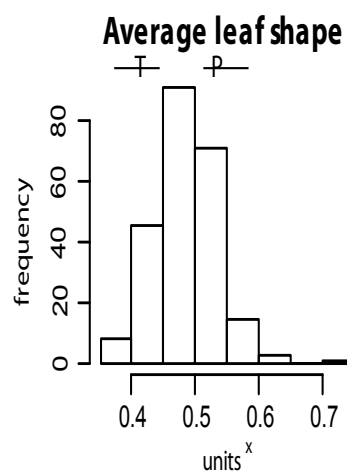
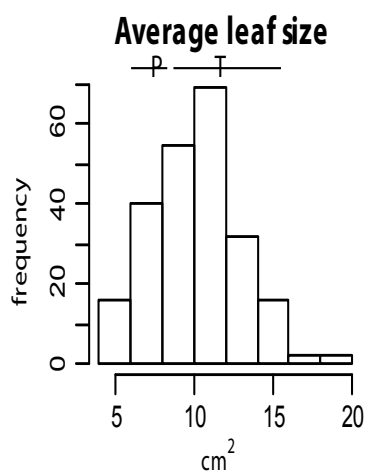
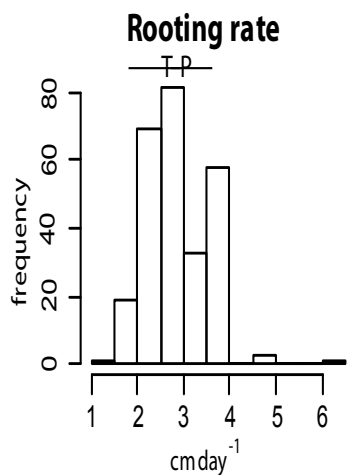
^a Absolute position from the top of the linkage group, in cM.

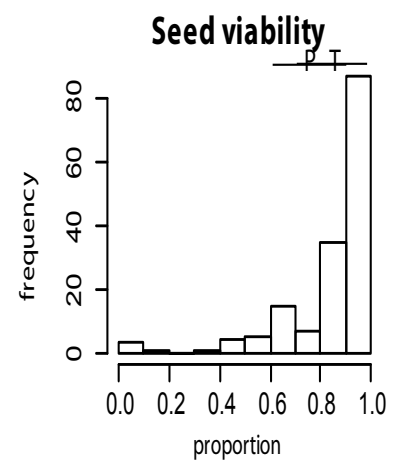
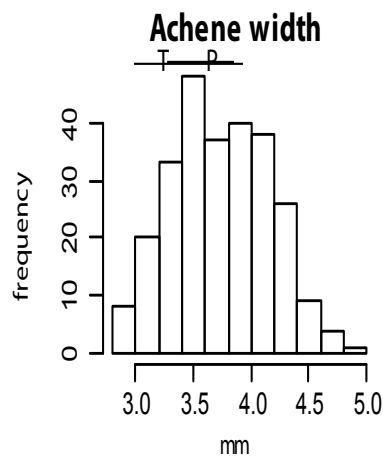
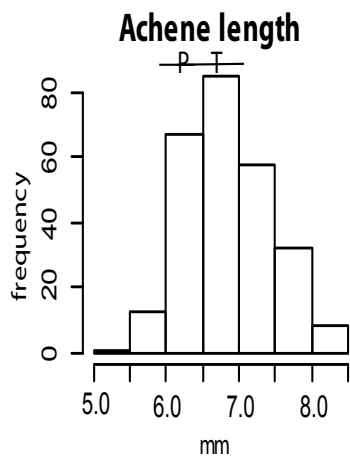
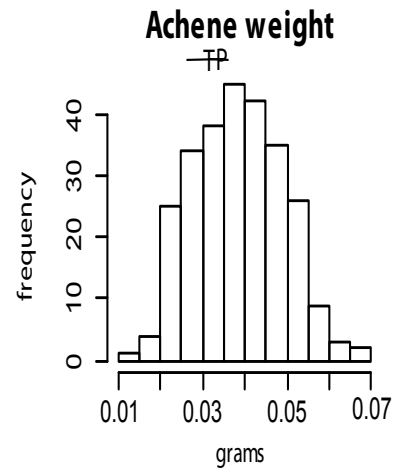
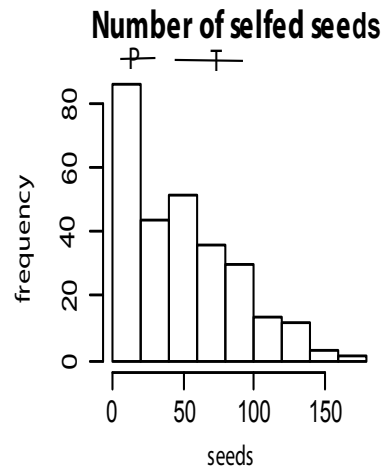
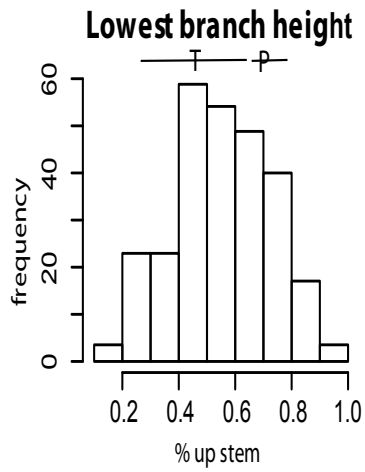
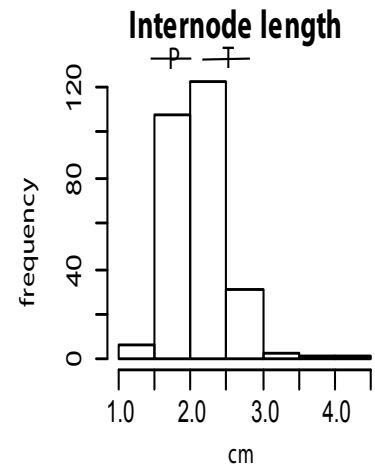
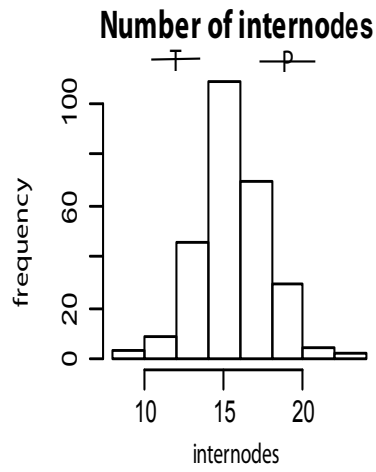
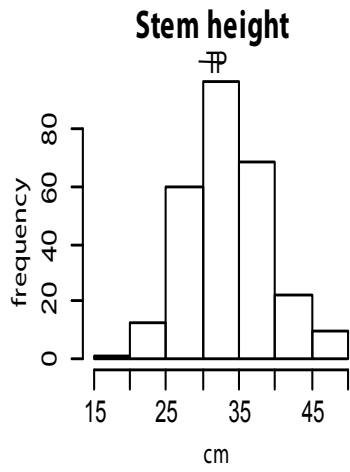
^b Region flanking the QTL peak within a one LOD score decline of the peak.

^c Refers to the effect of the cultivated safflower allele.

^d Underlined values indicate QTL in the “wrong” direction (see text for details) while italicized values describe cases in which directionality cannot be determined due to similarity in average parent trait values.

*Describes lower confidence QTL, identified at $\alpha = 0.1$





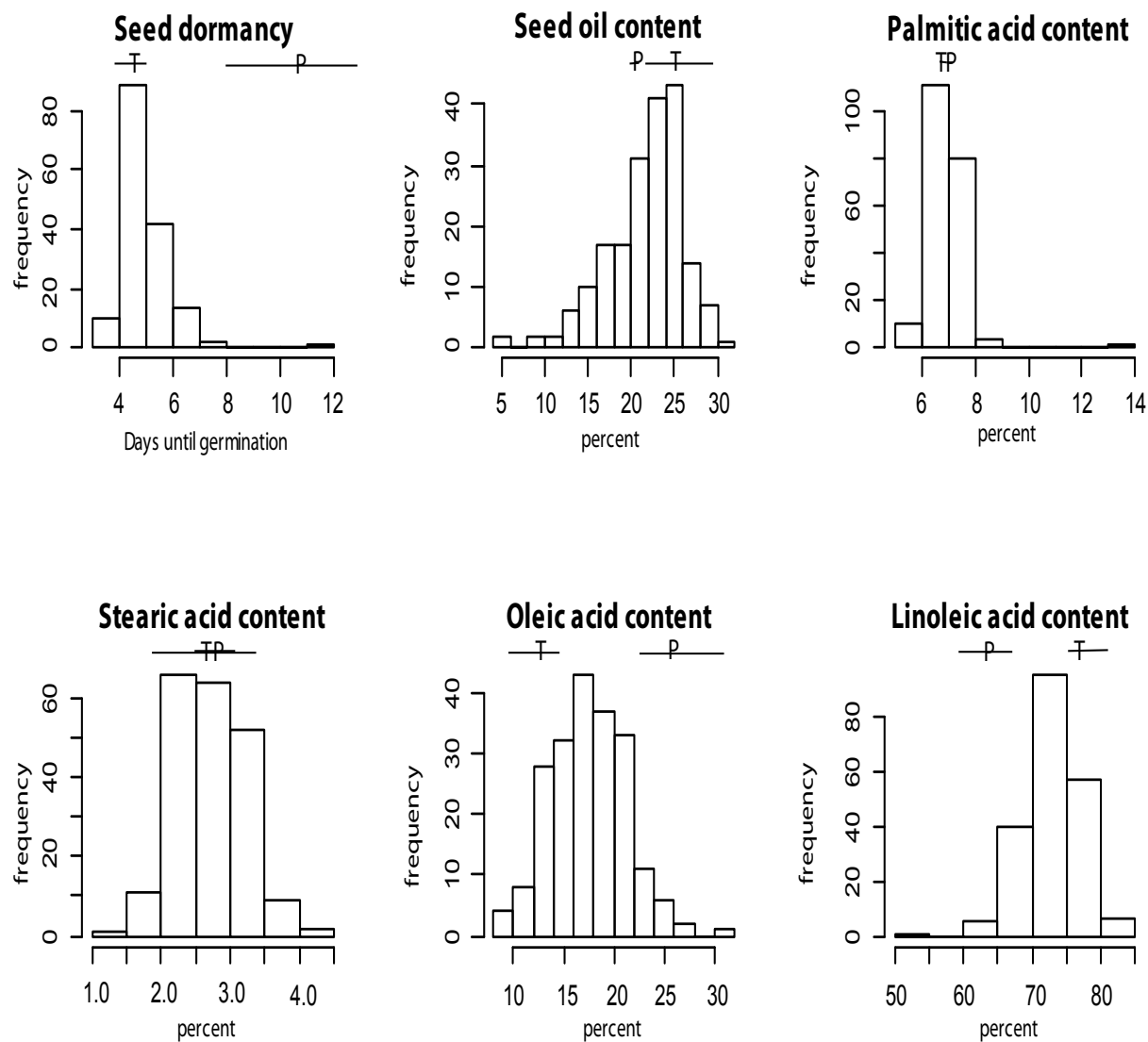
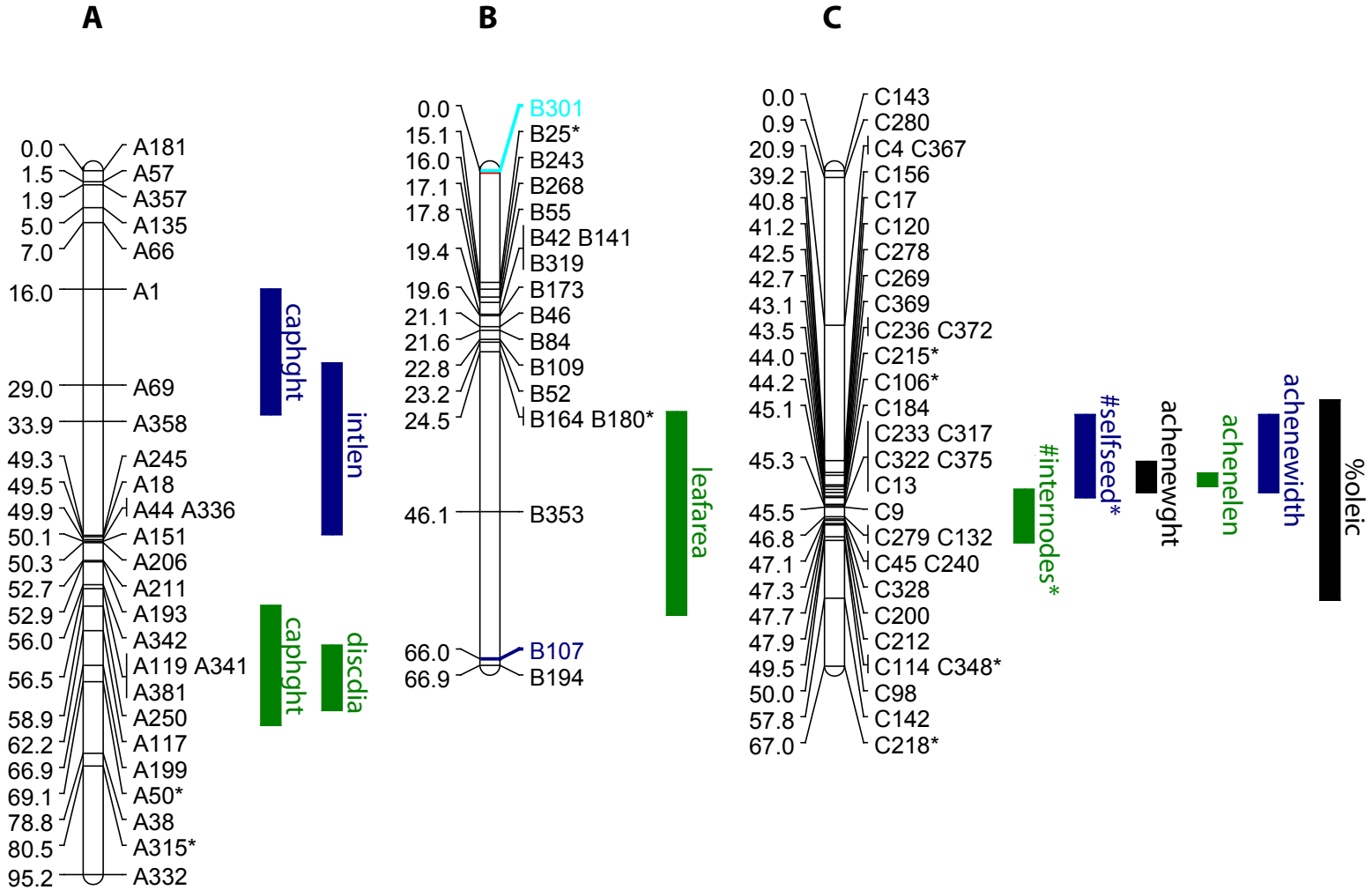
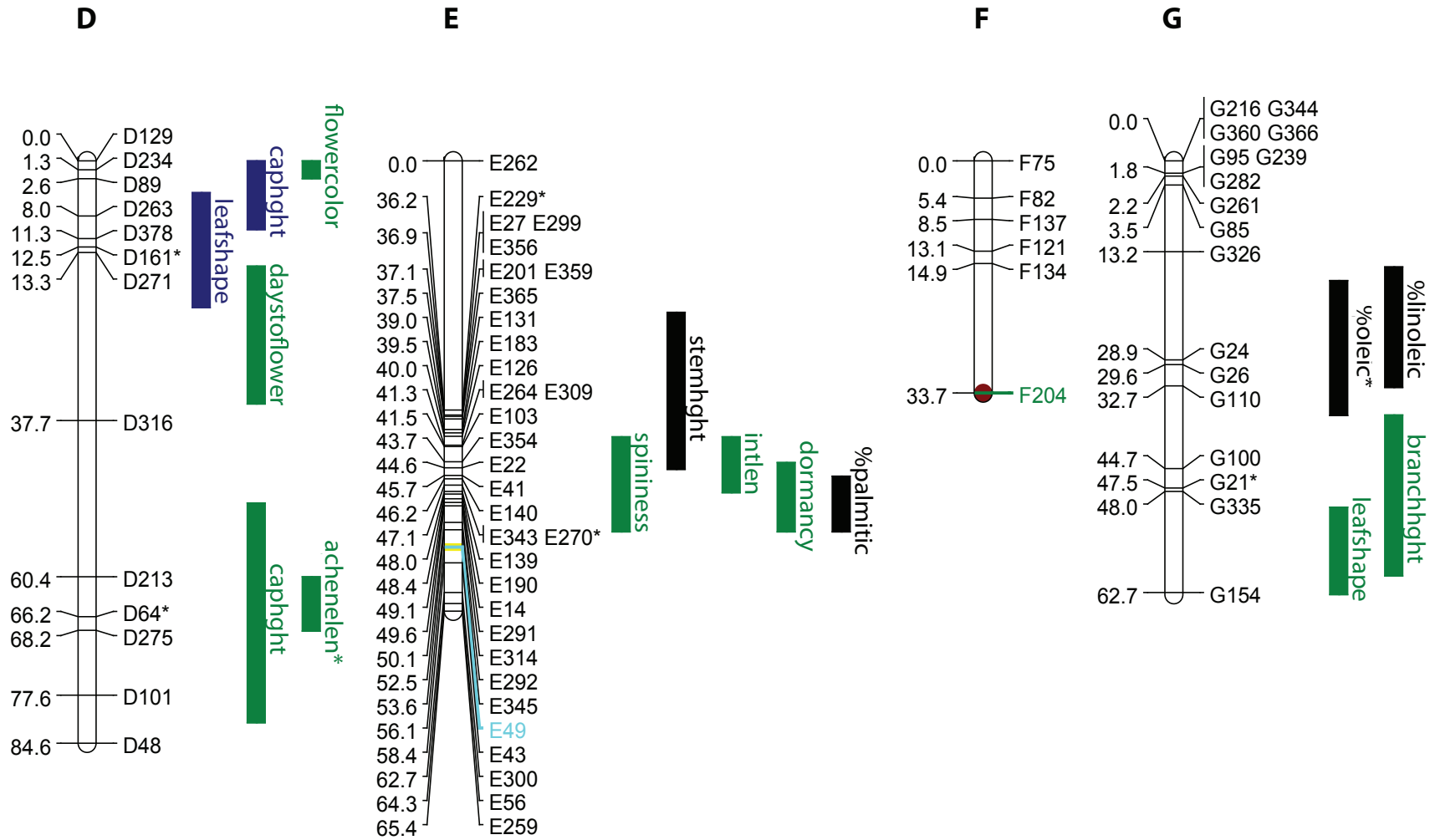
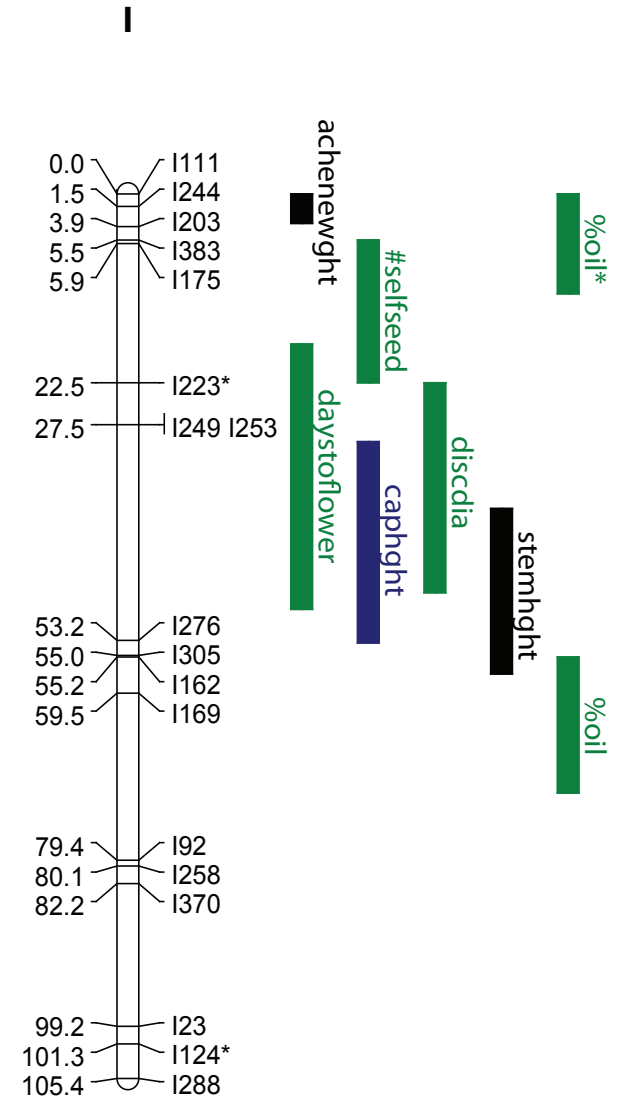
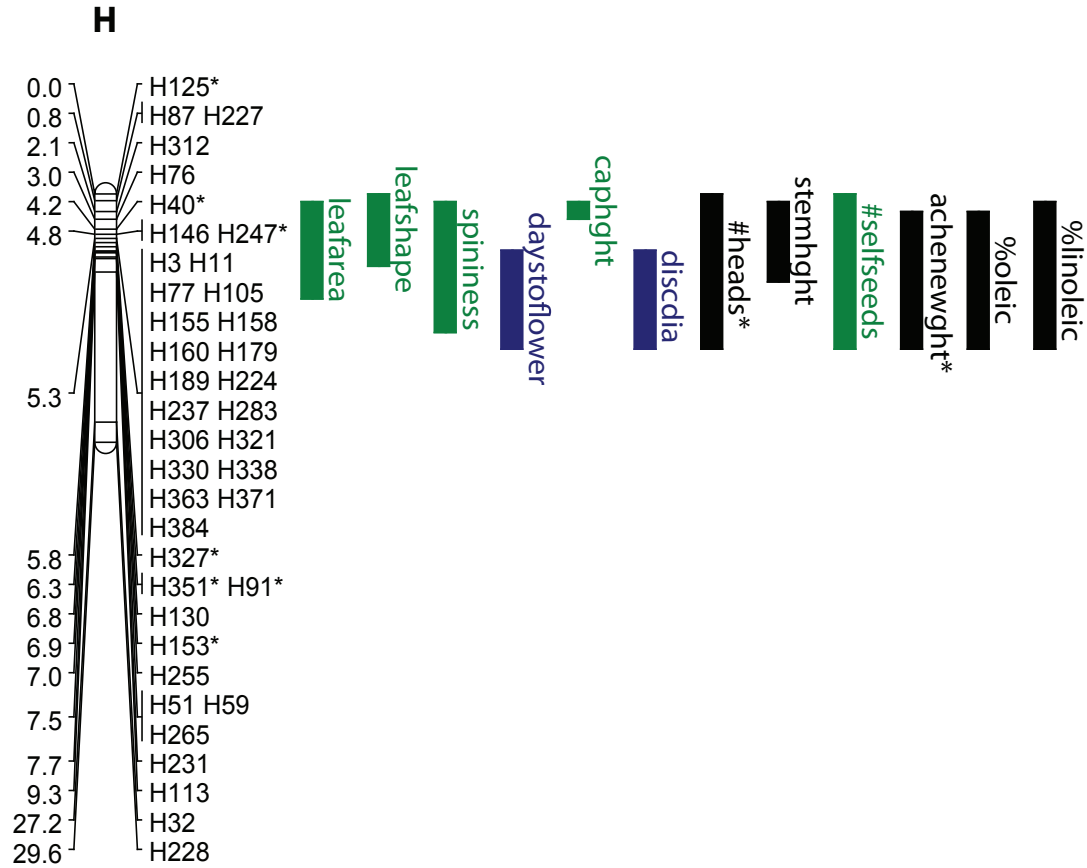


Figure 2.1. Trait distributions of the F₂ mapping population and mapping parent representatives. Means of the mapping parent representatives are represented by a T (*C. tinctorius*) and P (*C. palaestinus*) and solid lines represent one standard deviation around the means

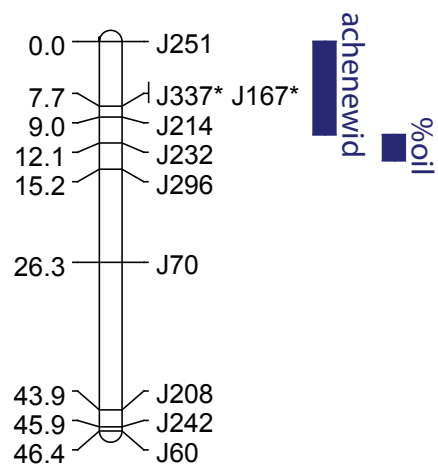
Figure 2.2. Genetic map of the safflower genome and corresponding quantitative trait locus (QTL) positions. Marker names are listed on the right and positions (in cM) are listed on the left of each linkage group. Markers with an asterisk represent SNPs mapped as dominant markers. Bars represent 1-LOD QTL intervals and traits with an asterisk denote “low confidence” QTL ($\alpha=0.10$). Green bars show locations of QTL in which the cultivar allele exhibits phenotypic effects in the expected direction while dark blue bars represent QTL where the cultivar allele confers a wild-like phenotype. Black bars represent QTL for traits in which the parents did not exhibit significant differences. Shaded regions along each linkage group represent regions exhibiting segregation distortion, with the following colors indicating different significance levels: yellow $\alpha < 0.05$, brown $\alpha < 0.01$, and red $\alpha < 0.001$. Marker names are shaded the same colors as QTL bars to indicate the directionality of the distortion, where light blue indicates that there is a heterozygote excess.



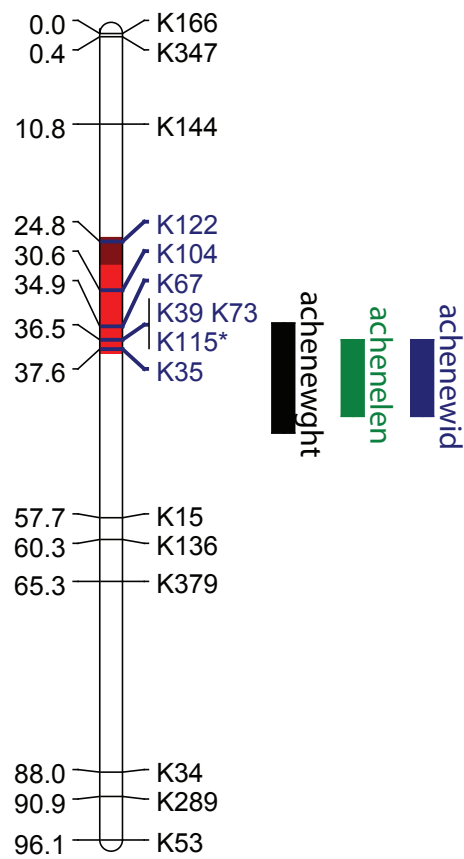




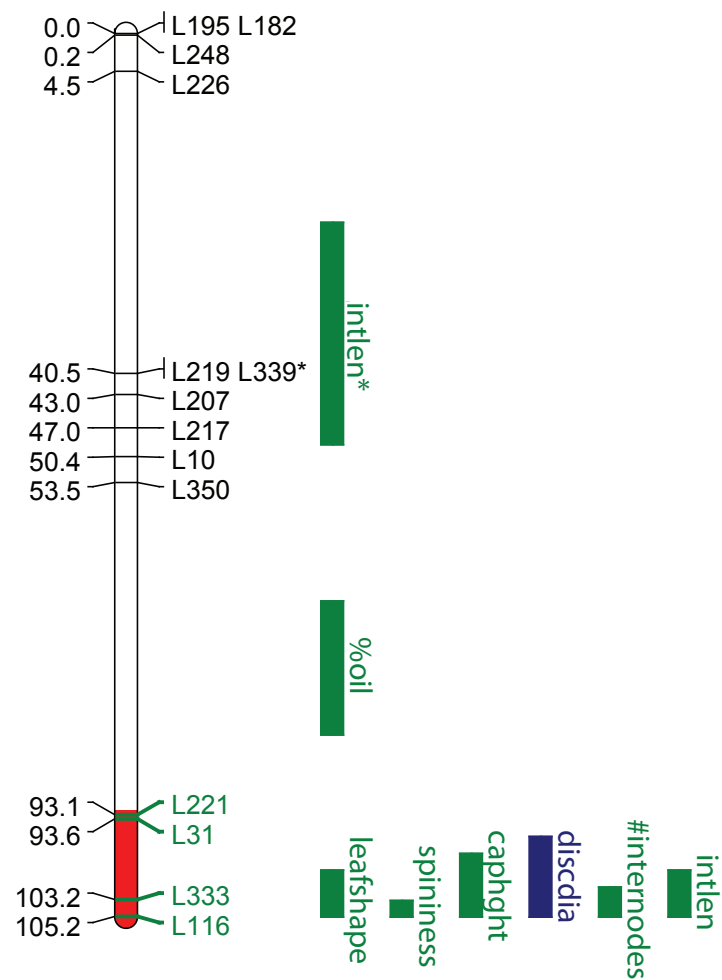
J

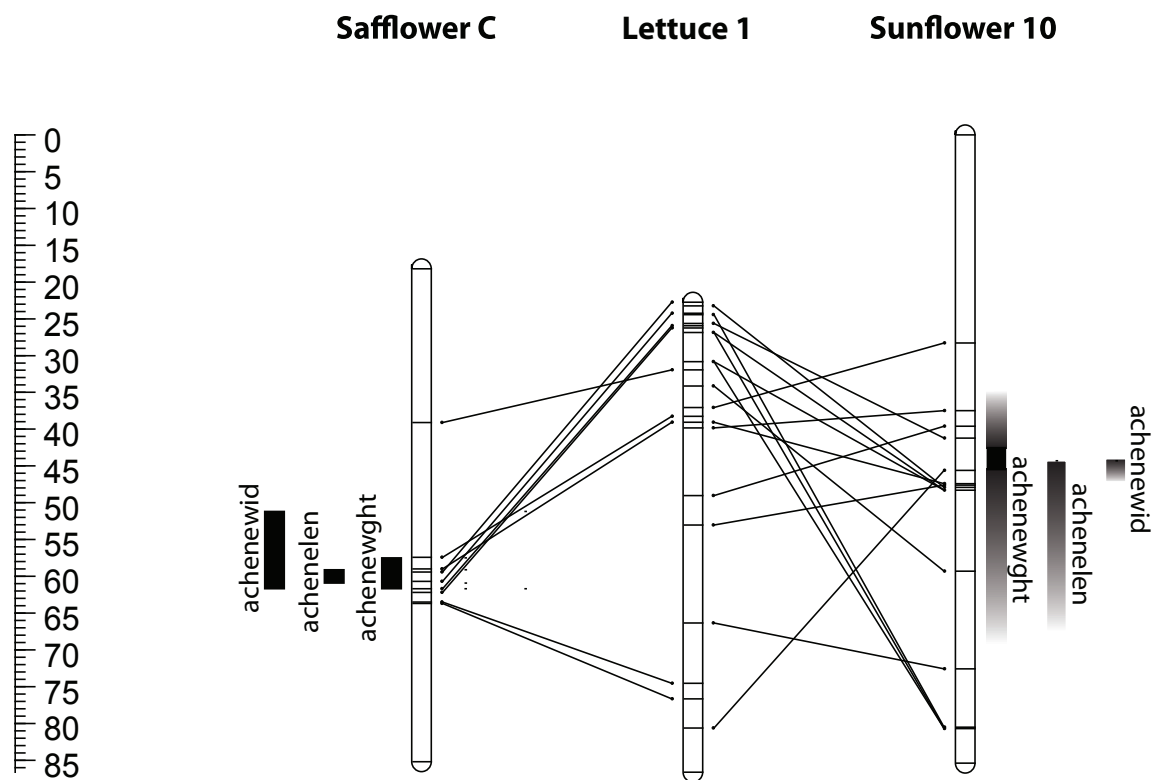
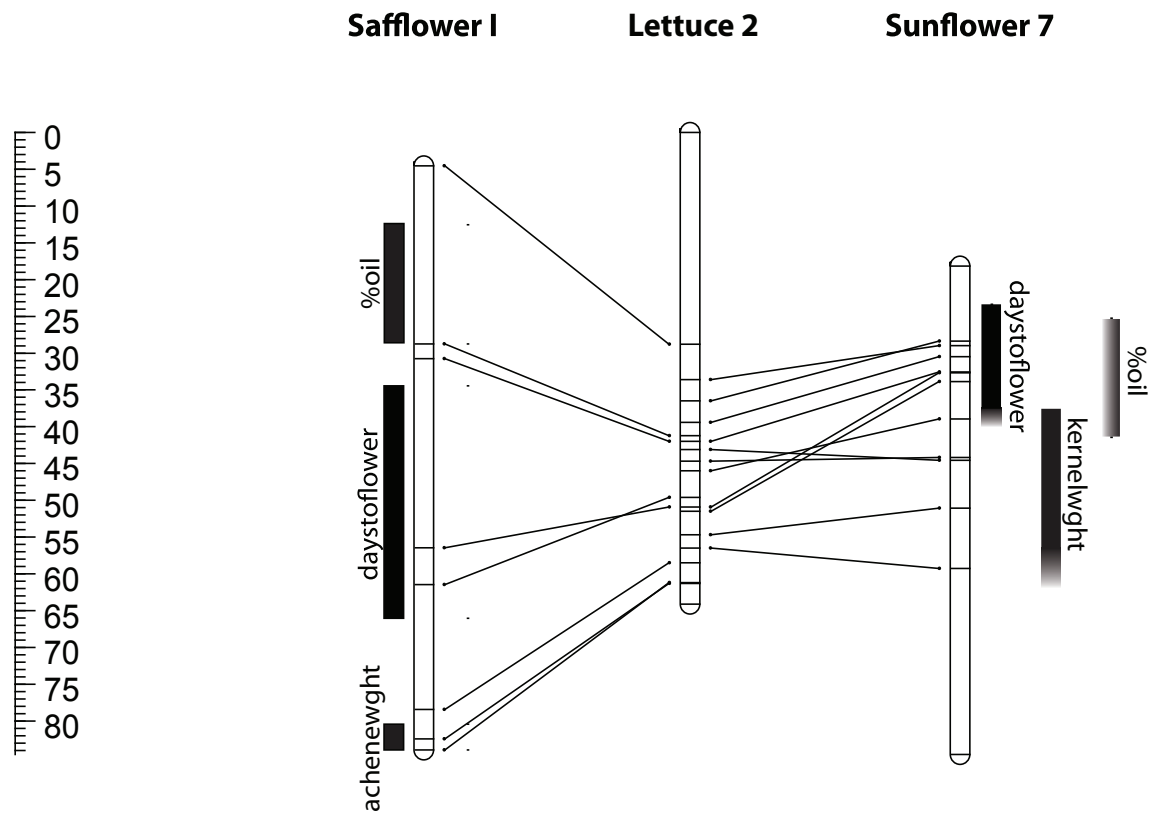


K



L





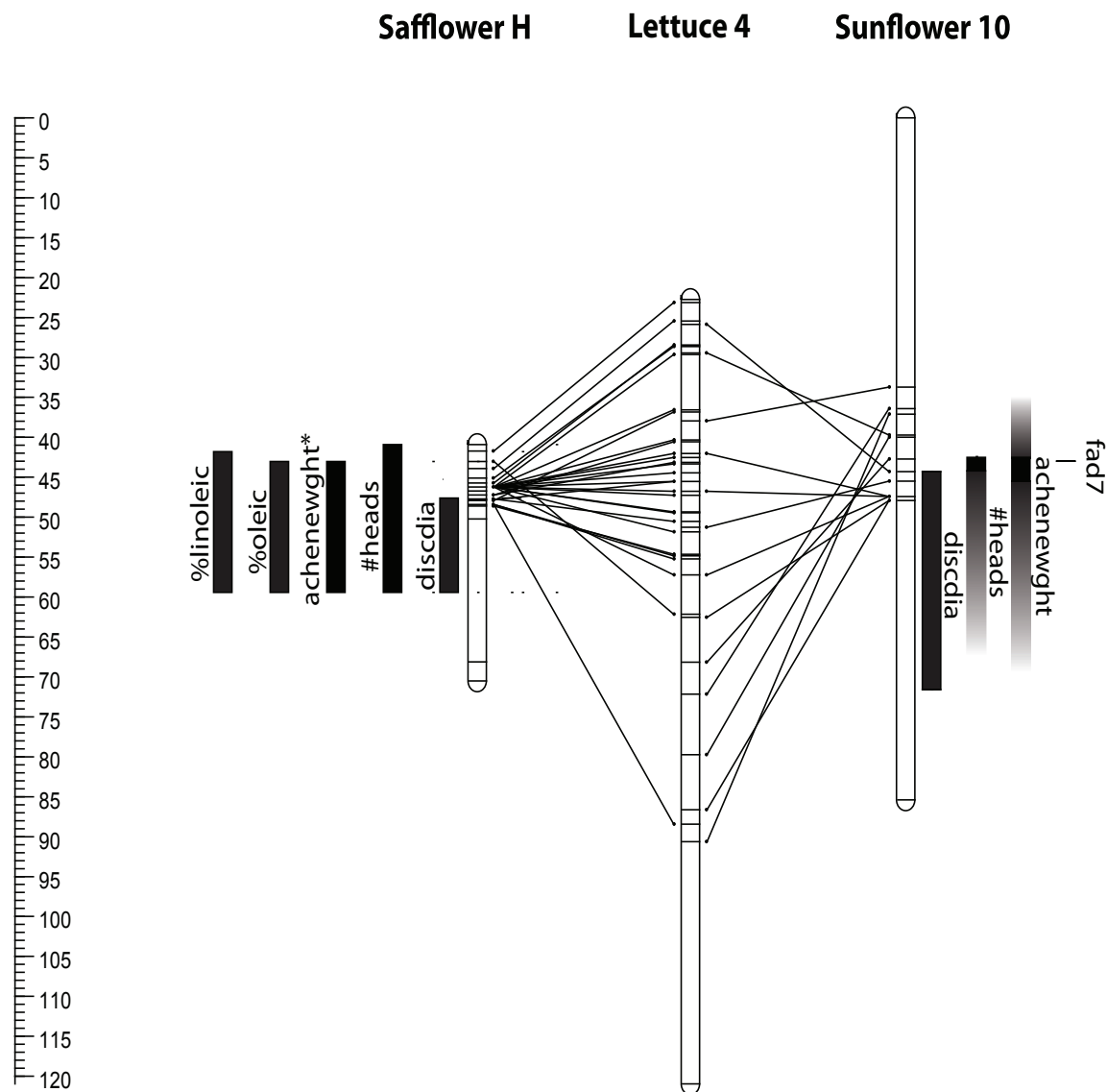


Figure 2.3. Comparative mapping of the safflower, lettuce, and sunflower genomes. Lines connect homologous loci between genomes. Black bars indicate quantitative trait loci (QTL) with exact 1-LOD positions known, while grayed gradient sections of bars represent estimated positions of QTL (based on relative lengths of the Bowers et al. [43] consensus map and the maps in which the sunflower QTL were originally published). Traits with an asterisk denote “low confidence” safflower QTL significant at $\alpha = 0.1$.

CHAPTER III

ANALYSIS OF GENETIC DIVERSITY IN *CARTHAMUS TINCTORIUS* L. (SAFFLOWER),
AN UNDERUTILIZED CROP²

² Pearl, S.A., and J.M. Burke. To be submitted to American Journal of Botany.

ABSTRACT

- *Premise of the study:* Underutilized crops are valuable resources for meeting the increasing food demands of the world's growing population. Safflower, an oilseed crop, is an example of one such underutilized crop that is capable of growing in moisture-limited areas. Characterization of the genetic diversity maintained within gene pools of underutilized crops such as safflower is an important first step in their further development.
- *Methods:* A total of 190 safflowers were genotyped using 133 single nucleotide polymorphism (SNP) loci. This collection of safflowers included 134 USDA accessions, 48 breeding lines from two private North American safflower breeding companies, and 8 wild safflower individuals. Using the multilocus genotypes of each of these individuals, we assessed genetic diversity within and among these collections of safflower.
- *Key results:* Though just a modest reduction in heterozygosity was observed in the commercial breeding lines (relative to the other safflower groupings), there was a significant decrease in allelic richness that accompanied the safflower domestication bottleneck. Our results suggest that this may be due to the fact that most safflower breeding lines appear to have originated from a single pool of diversity within Old World safflower germplasm.
- *Conclusions:* These results highlight the opportunity to access novel diversity from closely related, wild safflower species and parts of the safflower germplasm collection. Paired with future analyses of functional diversity, the molecular resources presented here will be useful in the future development of safflower.

INTRODUCTION

Underutilized crops are defined as those domesticated species whose genetic potential has not been fully realized [1]. These non-commodity crops are part of a “larger biodiversity portfolio” that tends to be underused by farmers and consumers for a variety of agronomic, economic, and cultural factors [1]. Given that food security is improved by the availability of a diverse assemblage of crop species, the development and production of underutilized crops has recently gained increased priority [2]. Because these species are often adapted to cultivation on marginal lands, they also offer viable agricultural alternatives in response to climate change and provide farmers with additional options for maximizing land usage [2]. These crops also help satisfy an increasing demand for “natural” and environmentally-friendly products while offering sources of diversified income to farmers and agricultural businesses [3].

The establishment and genetic characterization of germplasm collections is an important first step in securing and leveraging the resource base of underutilized crops [4]. Such germplasm collections often include cultivated materials obtained from throughout the world and may also include closely related, wild species. These collections thus represent a potentially important source of genetic diversity for ongoing plant breeding efforts [5]. Unimproved landraces and wild germplasm may be particularly valuable sources of novel alleles for the adaptation of crop plants to environmental challenges [6]. Unfortunately, little is often known about the genetic diversity contained within such collections and their genetic potential often goes untapped.

Carthamus tinctorius L. (safflower; Asteraceae; $2n = 2x = 24$; [7]) was domesticated approximately 4,500 years ago in the Fertile Crescent region from its putative wild progenitor, *Carthamus palestinus* Eig. [8–10]. Safflower was originally cultivated for the deep red pigments

(carthamine) in its florets, which were used as a source of dye for various cultural purposes.

Floral extracts have also been used as a food additive and are valued for their supposed medicinal properties [11]. Following its domestication, safflower cultivation spread throughout the Middle East, northern Africa, India, and the Far East. In the late 1890s, safflower was introduced to North America where commercial production commenced in the 1950s.

Today, safflower is grown for its seed oils that are rich in unsaturated fatty acids and as a source of seeds for use in birdseed mixes. Its flowers are also occasionally sold in ornamental bouquets. However, safflower remains something of a niche crop with limited production in North America (<http://faostat.fao.org/>) and much of its production elsewhere is being done in the context of smallholder farms. In the mid-1990s, safflower was identified by the International Plant Genetic Resources Institute (IPGRI) and the German Agency for Technical Cooperation (GTZ) as one of 25 underutilized crops that should be the focus of further development [3, 12]. This interest in safflower was driven by its local and regional importance (e.g., both economically and as a staple food in underdeveloped countries such as India and Ethiopia), potential for socioeconomic and agricultural development throughout the world, adaptation to areas in which surface moisture is limited, and the danger of genetic erosion within the crop gene pool [3]. Breeding efforts have, however, been hampered by a lack of access to molecular tools that could facilitate more rapid improvement. In recent years, however, this situation has begun to change (e.g., [13–15] and Pearl et al, in revision).

Herein, we describe the use of a large collection of single nucleotide polymorphisms (SNPs) to characterize patterns and levels of nucleotide diversity across a broad cross-section of the safflower gene pool. This includes a representative, worldwide sampling of diversity from the USDA germplasm collection, lines from the major two private North American commercial

safflower breeding programs, plus a set of wild safflower individuals. Using the resulting data, we explore the likely origins of the modern breeding materials and consider the utility of the available germplasm resources, including wild species, as possible sources of novel genetic diversity for the advancement of safflower breeding programs.

MATERIALS AND METHODS

Plant materials and genotyping. The focus of this study was a broad sampling of *Carthamus* germplasm (N = 190 individuals total), including 8 wild (3 *C. palaestinus* and 5 *C. persicus* from various sources; Table 3.1) and 182 cultivated safflower individuals. The latter included representatives of 96 geographically widespread Old World accessions and 38 New World (plus Australian) accessions obtained from the USDA (Table B1), as well as 48 lines donated by the two primary safflower breeding companies in North America (CalOils and Safflower Technologies International; referred to hereafter as CO and STI). There is a lack of consensus in the literature as to whether *C. persicus* and *C. palaestinus* are synonymous [16] or separate species [17], and there has even been speculation that *C. palaestinus* is a hybrid between *C. tinctorius* and *C. persicus* [16, 18]. However, authorities on Mediterranean floral taxonomy regard *C. palaestinus* as an invalid designation of *C. persicus* (<http://emplantbase.org>); we thus treat *C. persicus* and *C. palaestinus* as synonymous for the remainder of this paper and generally refer to them as “wild safflower.”

Because cultivated safflower is self-compatible and previous studies have shown that individual accessions are genetically quite uniform [15], just a single representative of each accession or breeding line was used in our study. Of the 134 Old World and New World accessions, 81 are part of the USDA safflower core collection (<http://www.ars-grin.gov/npgs/>), and we included an additional 10 historically important New World accessions that were

developed during the latter half of the 20th century (Table B1). Finally, the materials donated by safflower breeders included a total of 25 commercial varieties (including six dual use oil/birdseed cultivars with the balance being oil lines), 17 elite breeding lines, and 6 lines from a “germplasm conversion” program that has involved introgression from wild safflower.

Seeds were planted in the University of Georgia greenhouses, leaf tissue was collected from seedlings, and DNA was extracted using Qiagen (Valencia, CA) DNeasy Plant Mini Kits following the manufacturer’s protocol. Single nucleotide polymorphisms (SNPs) were then genotyped using the Illumina (San Diego, CA) GoldenGate Assay described by Pearl et al., (in revision) on an Illumina Bead Express at the Georgia Genomics Facility. Finally, allele calls were obtained using the Illumina GenomeStudio software (ver. 2011.1).

Population genetic and statistical analyses

For each of our groupings (wild, Old World, New World, CO, and STI), we estimated expected heterozygosity [19], observed heterozygosity, and the percent polymorphic loci using GenAlEx ver. 6.5b2 [20, 21]. To compare the number of private alleles and estimates of allelic diversity among each of our unequally sized study groups, we used rarefaction [22–24] as implemented in HP-RARE [25]. For each of these statistics, we estimated the significance of the differences among groups using Tukey’s post-hoc test, where group and locus were used as the model effects [26]. Additionally, we calculated these statistics for the pooled sets of Old and New World samples in order to obtain global estimates of diversity in wild vs. cultivated safflower.

To estimate intrachromosomal pairwise linkage disequilibrium (LD) among markers used in this study, a matrix of the squared allele frequency correlations (r^2) was generated and these values were plotted as a function of distance (in cM) using the R programming language (R Core

Team, 2013; <http://www.R-project.org/>). We then summarized the r^2 values using the “locpoly” and “dpill” functions in the R package “KernSmooth” ([28]; <http://CRAN.R-project.org/package=KernSmooth>).

Genetic structure among our five groups was assessed in GenAlEx via analysis of molecular variation (AMOVA; [29]) which hierarchically partitioned genetic variation, estimated F_{ST} [30], and determined significance based on 999 permutations of the data. Additionally, population structure among the five safflower groupings was examined using the Bayesian, model-based structuring algorithm STRUCTURE ver. 2.3 [31]. Initially, STRUCTURE was used to assign cluster membership of each sample without using geographic priors. To determine the most likely K (number of clusters), we followed the methods detailed by [32]. STRUCTURE was used to perform a total of 5 runs (each with 50,000 replicates following a 10,000 replicate burn-in) for each K from 1 through 12. For the most likely value of K , the proportion of membership of each individual in each cluster was averaged across runs. Additionally, STRUCTURE results were depicted geographically in maps drawn using the R package “maptools” ([33]; <http://CRAN.R-project.org/package=maptools>).

STRUCTURE was also used to investigate the origins of New World accessions and North American breeders’ lines. To do this, the STRUCTURE analyses were repeated as detailed above, but with only the Old World accessions as inputs. After determining the most likely number of clusters (K) within the Old World accessions, all individuals with greater than 80% membership in a single cluster were assigned to that cluster and the results were used to inform a third analysis [34]. This involved invoking the USEPOPINFO flag such that the aforementioned Old World individuals were used as a “training set” to assign the New World accessions and

breeders' lines to likely Old World clusters of origin. This same procedure was also used to investigate which Old World cluster corresponded most closely to the wild samples.

Genetic relationships among populations were also visualized using a principal coordinate analysis (PCoA) in GenAlEx. This involved using the multi-locus genotypes of all 190 individuals to create a standard genetic distance matrix [19]. Principal coordinates were then estimated based on this matrix, and the first two coordinates were plotted in two-dimensional space. Relationships among all individuals were further assessed by constructing a Neighbor-joining tree with POPTREE2 [35] using 500 bootstrap replicates. Trees were visualized using FigTree ver. 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>).

RESULTS

Genetic diversity and linkage disequilibrium. Our analyses focused on 133 clearly interpretable SNPs with known map position (Pearl et al., in revision). Estimates of genetic diversity within our five groups of wild and cultivated safflowers were relatively high. Nei's unbiased heterozygosity (H_e) (which could not exceed 0.55 because all loci were biallelic and the smallest group for which we estimated this contained only six individuals) ranged from 0.126 (ST) to 0.272 (wild) (overall mean $\pm$ S.E. = 0.227 ± 0.008 ; Table 3.2). In contrast, the average observed heterozygosity was much lower, ranging from 0.013 (CO) to 0.048 (STI) (overall mean = 0.036 ± 0.002 ; Table 3.2). Neither the observed nor expected heterozygosity estimates differed significantly when comparing just the wild and pooled cultivated safflower samples (Table 3.3). The mean percentage of polymorphic loci across groups was $70.54\% \pm 5.8\%$ (range = 55.64% [CO] to 81.2% [wild]; Table 3.2).

The private allelic richness (based on rarefaction) was significantly greatest in the wild group, both when comparing among the five groupings and comparing between the wild and

pooled cultivated sampling (Tables 3.2 and 3.3). When considering only the four groups comprising the *C. tinctorius* subset, four loci had alleles private to the New World grouping and another four loci had alleles unique to the Old World grouping. However, these alleles were all present in the wild safflower samples, and we found a total of 23 additional private alleles in the wild safflower grouping. Further, 25 alleles not found in the wilds were private to the pooled cultivated safflowers. Meanwhile, the rarefied allelic richness was significantly greatest in the wild grouping in both the pooled analysis and analysis of the five groupings ($A_g \pm \text{S.E.} = 1.82 \pm 0.034$), and the CO group had the lowest estimate of allelic richness (1.43 ± 0.037 ; Tables 3.2 and 3.3).

Analysis of the subgroups within the CO breeding lines revealed that the germplasm conversion lines were significantly less homozygous than the elite lines and commercial varieties ($F_{2,73} = 17.16$, $P < 0.0001$), but were not significantly more diverse (Table B2). Additionally, H_e , A_g , and private allelic richness were not significantly different among any of these groups (Supplementary table 2). Interestingly, the germplasm conversion lines had the lowest percentage of polymorphic loci (Table B2).

Overall levels of LD were generally low. The average intrachromosomal LD was less than 0.1, with the exceptions of linkage groups F and L. Although the extent of LD varied somewhat across linkage groups (Figure B1), LD diminished to less than 0.1 within 9 cM and sometimes within 1 cM (linkage groups D and J). Note that, for some linkage groups (i.e., B and I), it was not possible to summarize LD via the KernSmooth function due to a paucity of SNPs. Interestingly, we found surprisingly high pairwise LD ($r^2 = 0.365$ and 0.304) between opposite ends of linkage group L, a distance spanning over 100 cM.

Population structure, relationships. Among the five pre-defined groups investigated in this study, F_{ST} as estimated from AMOVA ranged from 0.070 (between New World and Old World; $P = 0.001$) to 0.712 (between wild safflowers and the CO breeding lines; $P = 0.001$; Table B3). Our STRUCTURE analyses of all wild and cultivated safflowers indicated $K = 2$ was the most likely number of clusters (Figure B3a,b), with one cluster largely corresponding to the wild individuals and Old World accessions and the other cluster mostly corresponding to the New World accessions and breeders' lines (Figure B2). Examination of the next most likely result ($K = 9$, Figure B3a,b and Figure 3.1) revealed a much more nuanced picture: the wild individuals formed their own cluster (cluster 1), the Old World accessions grouped into five different clusters (clusters 2 through 6), one of which jointly grouped with several New World accessions (cluster 6). Each of these clusters was generally characterized by a predominant geographic region: cluster 2 = Israel, Jordan, and Ethiopia; cluster 3 = Europe; cluster 4 = Iran, Afghanistan, and Turkey; cluster 5 = Far East; and cluster 6 = Egypt and Sudan (Figures 3.1, 3.2a). Two additional clusters included one set of New World accessions plus the majority of the North American breeders' lines (cluster 7) and a separate cluster composed of a subset of New World accessions (cluster 8). Finally, a subset of the STI breeding lines formed their own cluster (cluster 9). An analysis of wild and cultivated safflowers on a per-cluster basis (and excluding those accessions with less than 50% membership in any one cluster) revealed that genetic diversity was greatest in cluster 4 (corresponding to Iran, Afghanistan, and Turkey; unbiased $H_e = 0.182 \pm 0.015$, rarefied $A_g = 1.42 \pm 0.031$) and lowest in cluster 8 (primarily consisting of a third of the "historically important" US accessions; unbiased $H_e = 0.064 \pm 0.013$, $A_g = 1.14 \pm 0.029$; data not shown).

Our separate STRUCTURE analysis of Old World accessions to identify a “training set” for subsequent population assignment yielded a most likely result of $K = 4$ clusters (Figure B3b,c). These clusters largely corresponded to the four Old World clusters described above, with the exception that cluster 6 grouped with cluster 2 (Figure B4). The subsequent population assignment analysis in STRUCTURE assigned each of the New World accessions and breeders’ lines to a mixture of the four Old World clusters, though the greatest proportion of each individual corresponded to the joint Jordan/Israel/Ethiopia-Egypt/Sudan cluster (Figure 3.3). Meanwhile, the Iran/Afghanistan/Turkey showed the highest level of similarity with the wild individuals, followed by the joint Israel/Jordan/Ethiopia-Egypt/Sudan clusters (Figure 3.3).

The results of the PCoA and the Neighbor-joining analyses (Figures 3.4 and B5) were largely congruent with the STRUCTURE results and reflected our F_{ST} estimates, showing that the wild safflowers were largely distinct from cultivated safflowers. Interestingly, the longest branch in the Neighbor-joining tree separated the wild individuals recently collected in Israel from the remaining *C. persicus* (Barcelona) and *C. palaestinus* (USDA) accessions (Figure B5). In both the PCoA and Neighbor-joining analyses, all wild samples clustered nearest to the Iran/Afghanistan/Turkey Old World population, which corresponds to safflower’s putative center of origin. Finally, although Old World accessions from different geographic regions exhibited overlap (Figure 3.4 and Figure B5), geographic structuring was still apparent.

DISCUSSION

An understanding of the amount and distribution of genetic diversity within crop germplasm collections can provide valuable insight into the evolutionary history of the species in question and help to guide future improvement efforts [5]. One caveat of this study was the limited availability of wild safflower samples, and this undersampling of the wild safflower

population may be driving the exceptionally high F_{ST} values observed among wild and all other cultivated safflower groupings. Despite this limited sampling, our analyses revealed an overall reduction in diversity in cultivated vs. wild safflower (Tables 3.2 and 3.3). Further, the low levels of observed heterozygosity were consistent with a history of inbreeding due to both the self-incompatibility of safflower and its wild relatives [36] and the breeding history of cultivated safflower.

Within the CO breeding lines, the germplasm conversion lines were significantly less inbred than the varieties and elite lines though, surprisingly, introgression from the wild has failed to produce the expected infusion of novel molecular diversity (Table B2). Meanwhile, the STI breeding lines were more differentiated, falling predominantly into one of two distinct clusters within the STRUCTURE analysis and Neighbor-joining tree (Figure 3.1, Figure B5). This division largely corresponded to differences in market type – i.e., most of the high oleic lines fell in one grouping, while the linoleic, birdseed, and subset of oleic lines clustered in the other.

Our STRUCTURE analysis of all 190 samples partitioned the data into as many as 9 genetically distinct clusters that largely corresponded with geography and/or breeding history (Figures 3.1 and 3.2). Within our Old World grouping of safflower accessions, STRUCTURE identified four clusters that corresponded to four different geographic regions (Figure B4) that presumably represent somewhat distinct breeding pools. These clusters correspond quite closely with five centers of safflower diversity previously identified by Chapman et al. [37]. Our population assignment analysis investigating the origins of New World accessions and commercial breeding lines grouped these samples primarily with the cluster predominantly comprised of individuals from Israel, Jordan, Ethiopia, Egypt, and Sudan (Figure 3.3). Within

this Old World cluster, a few accessions from China and Sudan likely represent the original source material from which much of the historically important New World and North American breeding lines were derived. In the original STRUCTURE analysis (Figure 3.1), these Chinese and Sudanese accessions were greatly admixed (predominantly red and yellow bars far to the right of the Old World grouping in Figure 3.1), sharing clusters with many of the historically important New World and North American breeding lines. Further supporting this conclusion, our PCoA revealed that the Sudanese and one Chinese accession exhibited a large amount of overlap with North American breeding lines (Figure 3.4).

Our original STRUCTURE analysis also revealed that wild safflowers formed a largely distinct cluster (Figure 3.1), perhaps owing to the substantial number of private alleles that were found within this group. However, our population assignment analysis in STRUCTURE and subsequent PCoA and Neighbor-joining analyses suggested that the wild safflowers shared the greatest similarity with the Iran-Afghanistan-Turkey cluster in the Old World (Figures 3.3 and 3.4; Figure B5). This is consistent with safflower's presumed Near Eastern center of origin [10, 37]. Interestingly, the longest branch of the Neighbor-joining tree (and the only branch with 100% bootstrap support) separated the wild safflowers collected in Israel (obtained from the IGB, Dr. Yuval Sapir, and Dr. Amram Ashri) from the wild safflowers obtained from the Botanical Institute of Barcelona and the USDA (Figure B5), with the former being placed at the distal end of that branch. This may suggest that the samples we obtained from the IGB and the wild samples recently collected in southern Israel have retained more ancestral diversity and perhaps subsequently diverged from other samples maintained by other plant gene banks.

Although our analyses have revealed a rather limited reduction in overall genetic diversity within commercial breeding pools, the elevated private allelic richness in the wild

safflower grouping highlights their potential utility as a source of novel allelic diversity. As such, molecular tools such as those described herein could help to guide the selection of germplasm for pre-breeding efforts. While it is true that molecular diversity may not be an accurate predictor of phenotypic diversity [38], it seems clear that prior efforts to access novel diversity from wild safflower have left behind many of the allelic variants present in the wild. In this light, it is worth noting that a prior study of the genetic basis of domestication-related traits in safflower revealed the presence of numerous QTL with antagonistic effects (i.e., genomic regions in which the wild allele produces a more crop-like phenotype, and vice versa; Pearl et al., in revision). As such, it appears there are indeed agronomically favorable alleles present in wild safflower and that expanded efforts to access this diversity would facilitate the continued improvement of safflower, as it has done for numerous other crops (e.g., tomato, potato, rice, and wheat; reviewed in [39]).

REFERENCES

1. Padulosi S, Hoeschle-Zeledon I: **Underutilized plant species : what are they ?** *Leisa Leusden* 2004:5–6.
2. Mayes S, Massawe FJ, Alderson PG, Roberts J a, Azam-Ali SN, Hermann M: **The potential for underutilized crops to improve security of food production.** *Journal of experimental botany* 2012, **63**:1075–9.
3. Thies E: *Promising and Underutilized Species, Crops and Breeds*. Eschborn, Germany; 2000.
4. Padulosi S, Hodgkin T, Williams JT, Haq N: **Underutilized crops : trends, challenges and opportunities in the 21 st Century.** In *Perspectives on New Crops and New Uses*. Edited by Janick J. Alexandria, VA: ASHS Press; 1999:140–145.
5. Tanksley SD, McCouch SR: **Seed banks and molecular maps: unlocking genetic potential from the wild.** *Science* 1997, **277**:1063–1066.
6. McCouch S, Baute G, Bradeen J, Bramel P, Bretting P, Buckler ES, Burke JM, Charest D, Cloutier S, Cole G, Dempewolf H, Dingkuhn M, Feuillet C, Gepts P, Grattapaglia D, Guarino L, Jackson S, Knapp S, Langridge P, Lawton-Rauh A, Lijua Q, Lusty C, Michael T, Myles S, Naito

- K, Nelson R, Pontarollo R, Richards C, Rieseberg L, Ross-Ibarra J, et al.: **Agriculture: Feeding the future**. *Nature* 2013, **499**:3–5.
7. Patel J, Narayana G: **Chromosome numbers in safflower**. *Current Science* 1935, **4**:412.
8. Chapman MA, Burke JM: **DNA sequence diversity and the origin of cultivated safflower (*Carthamus tinctorius* L.; Asteraceae)**. *BMC Plant Biology* 2007, **7**:60.
9. Knowles P, Ashri A: **Safflower -- *Carthamus tinctorius* (Compositae)**. In *Evolution of Crop Plants*. Edited by Smartt, J. and Simmonds NW. Harlow, UK: Longman; 1995:47–50.
10. Van Zeist W, Rooijen Waterbolk-Van W: **Two interesting floral finds from third millenium BC Tell Hammam et-Turkman, northern Syria**. *Vegetation History and Archaeobotany* 1992, **1**:157–161.
11. Weiss EA: **Safflower**. In *Oilseed crops*. 2nd ed. London: Blackwell Science Ltd; 2000:93–129.
12. Dajue L, Mündel H-H: *Safflower. Carthamus Tinctorius L. Promoting the Conservation and Use of Underutilized and Neglected Crops*. Gatersleben, Germany and Rome, Italy; 1996.
13. Mayerhofer R, Archibald C, Bowles V, Good AG: **Development of molecular markers and linkage maps for the *Carthamus* species *C. tinctorius* and *C. oxyacanthus***. *Genome* 2010, **53**:266–276.
14. Chapman MA, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore RW, Tang S, Knapp SJ, Burke JM: **Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (*Carthamus tinctorius* L.)**. *Theoretical and Applied Genetics* 2009, **120**:85–91.
15. Johnson RC, Kisha TJ, Evans M a.: **Characterizing safflower germplasm with AFLP molecular markers**. *Crop Science* 2007, **47**:1728–1736.
16. Hanelt P: **Monographische übersicht der gattung *Carthamus* L. (Compositae)**. *Feddes Repertorium* 1963, **67**:41–180.
17. Garnatje T, Garcia S, Vilatersana R, Vallès J: **Genome size variation in the genus *Carthamus* (Asteraceae, Cardueae): systematic implications and additive changes during allopolyploidization**. *Annals of Botany* 2006, **97**:461–7.
18. Ashri A, Knowles PF: **Cytogenetics of safflower (*Carthamus* L.) species and their hybrids**. *Agronomy Journal* 1960, **52**:11–17.
19. Nei M: **Estimation of average heterozygosity and genetic distance from a small number of individuals**. *Genetics* 1978, **89**:583–590.

20. Peakall R, Smouse PE: **GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research.** *Molecular Ecology Notes* 2006, **6**:288–295.
21. Peakall R, Smouse PE: **GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research -- an update.** *Bioinformatics* 2012, **28**:2537–2539.
22. Hurlbert SH, Jul N: **The nonconcept of species diversity: a critique and alternative parameters.** *Ecology* 1971, **52**:577–586.
23. Petit R, Mousadik AEL, Pons O: **Identifying populations for conservation on the basis of genetic markers.** *Conservation Biology* 1998, **12**:844–855.
24. Kalinowski ST: **Counting alleles with rarefaction: private alleles and hierarchical sampling designs.** *Conservation Genetics* 2004, **5**:539–543.
25. Kalinowski ST: **HP-Rare 1.0: a computer program for performing rarefaction on measures of allelic richness.** *Molecular Ecology Notes* 2005, **5**:187–189.
26. Sokal R, Rohlf F: *Biometry*. 3rd. ed. New York, New York: W.H. Freeman and Company; 1995.
27. **R: A language and environment for statistical computing** [<http://www.r-project.org>]
28. Wand M: **KernSmooth: Functions for kernel smoothing for Wand & Jones (1995).** 2013.
29. Excoffier L, Smouse PE, Quattro JM: **Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data.** *Genetics* 1992, **131**:479–491.
30. Wright S: **The genetical structure of populations.** *Annals of Eugenics* 1951, **15**:323–354.
31. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**:945–959.
32. Evanno G, Regnaut S, Goudet J: **Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.** *Molecular Ecology* 2005, **14**:2611–2620.
33. Bivand R, Lewin-Koh N: **maptools: Tools for reading and handling spatial objects.** 2013.
34. Hubisz MJ, Falush D, Stephens M, Pritchard JK: **Inferring weak population structure with the assistance of sample group information.** *Molecular Ecology Resources* 2009, **9**:1322–1332.
35. Takezaki N, Nei M, Tamura K: **POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface.** *Molecular Biology and Evolution* 2010, **27**:747–752.

36. Claassen CE: **Natural and controlled crossing in safflower, *Carthamus tinctorius* L.** *Agronomy Journal* 1950, **42**:381–384.
37. Chapman MA, Hvala J, Strever J, Burke JM: **Population genetic analysis of safflower (*Carthamus tinctorius*; Asteraceae) reveals a Near Eastern origin and five centers of diversity.** *American Journal of Botany* 2010, **97**:831–840.
38. Reed DH, Frankham R: **How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis.** *Evolution* 2001, **55**:1095–1103.
39. Hajjar R, Hodgkin T: **The use of wild relatives in crop improvement: a survey of developments over the last 20 years.** *Euphytica* 2007, **156**:1–13.

TABLES

Table 3.1. Collection information of wild safflower samples used in this study.

Species	Identifier	No. individuals sampled in this study	Collection site	Source
<i>C. palaestinus</i>	PI 235663	2	N/A	USDA
<i>C. palaestinus</i>	Ashri	1	Revivim, Israel	Amram Ashri
<i>C. persicus</i>	S-2358	1	Elazig, Turkey	Botanical Institute of Barcelona
<i>C. persicus</i>	23666	2	Arava Valley, Israel	Israel Plant Gene Bank
<i>C. persicus</i>	Sapir	2	Negev Desert, Israel	Yuval Sapir

Table 3.2. Genetic diversity statistics of wild, cultivated (Old World and New World), and commercial (CO and STI) safflower groupings.

Population	<i>N</i>	%P	<i>A_g</i> (± S.E.)	<i>PAL</i>	<i>H_o</i>	<i>H_e</i>
Wild	8	81.2	1.82 ^a (0.034)	0.25 ^a	0.040 ^a	0.272 ^a
Old World	96	79.7	1.68 ^b (0.037)	0 ^b	0.042 ^a	0.266 ^a
New World	38	78.95	1.67 ^{bc} (0.035)	0 ^b	0.038 ^a	0.247 ^a
CO	34	55.64	1.43 ^d (0.037)	0 ^b	0.013 ^b	0.126 ^b
STI	14	57.14	1.55 ^c (0.042)	0 ^b	0.048 ^a	0.223 ^a

N number of plants sampled

%P percent polymorphic loci

A_g allelic richness (based on the rarefaction method) averaged across all loci

PAL private allelic richness (based on the rarefaction method) averaged across all loci

H_o observed heterozygosity averaged across all loci

H_e expected heterozygosity averaged across all loci

Superscript letters indicate differences in significance levels ($P < 0.001$)

Table 3.3. Genetic diversity statistics comparing wilds with USDA cultivated safflower accessions (Old World and New World combined).

Population	<i>N</i>	%P	<i>A_g</i> (± S.E.)	<i>PAL</i>	<i>H_o</i>	<i>H_e</i>
Wild	8	82.71	1.82* (0.034)	0.31*	0.040	0.272
USDA cultivated (Old World + New World)	134	81.20	1.69* (0.036)	0.18*	0.041	0.270

N number of plants sampled

%P percent polymorphic loci

A_g allelic richness (based on the rarefaction method) averaged across all loci

PAL private allelic richness (based on the rarefaction method) averaged across all loci

H_o observed heterozygosity averaged across all loci

H_e expected heterozygosity averaged across all loci

* indicates categories are significantly different ($P < 0.001$)

FIGURES

Figure 3.1. STRUCTURE plots of 190 *Carthamus* individuals. Black bars are used to separate pre-defined groupings of *Carthamus* (labels found at top of graphs). Each vertical bar represents a single individual, and the proportion of membership to each population is indicated on the y-axis. Here, $K = 9$ clusters, with geographic origins of the majority of each cluster indicated along the bottom of the graph; blue = cluster 1, orange = cluster 2, green = cluster 3, light blue = cluster 4, pink = cluster 5, light green = cluster 6, red = cluster 7, yellow = cluster 8, purple = cluster 9.

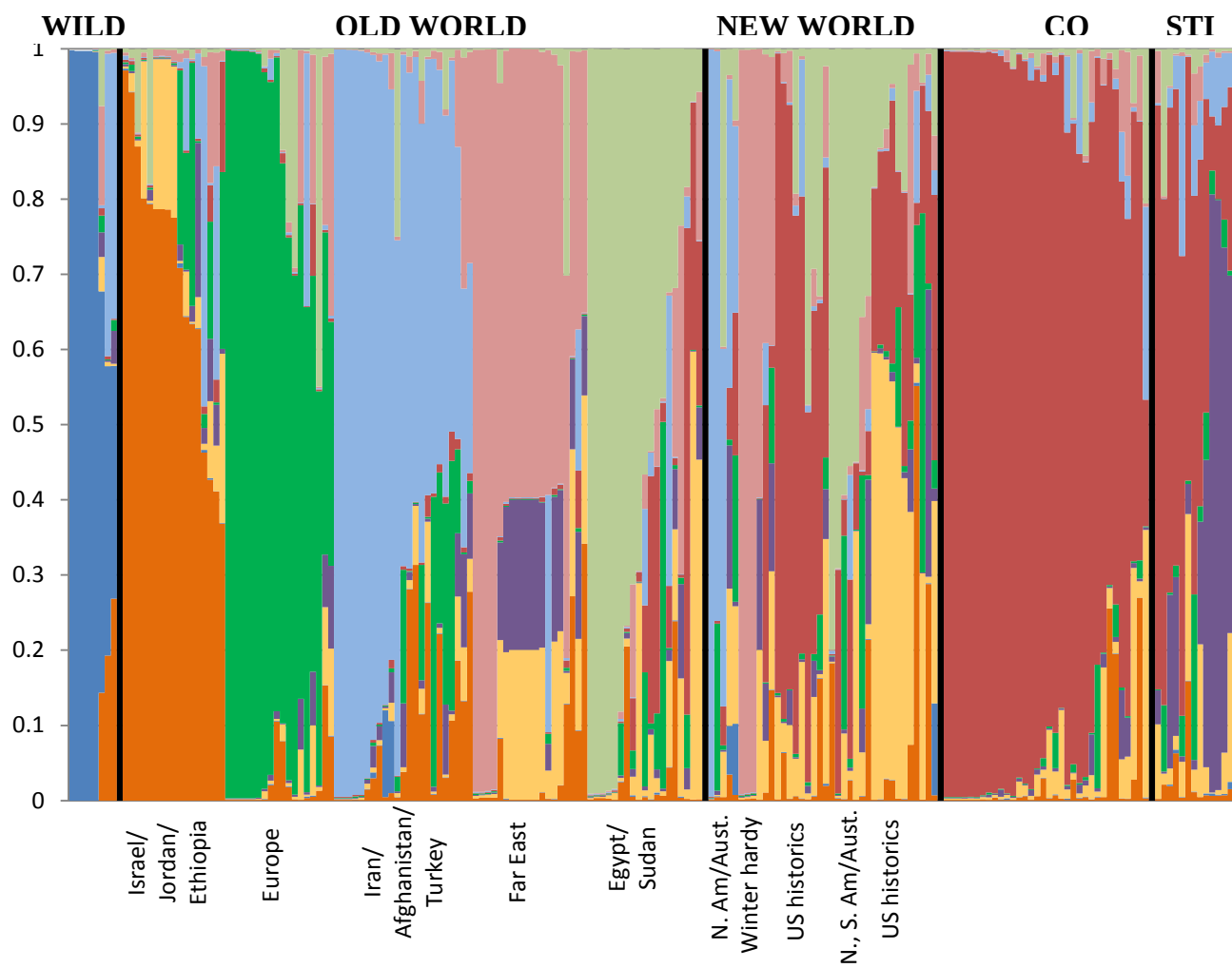
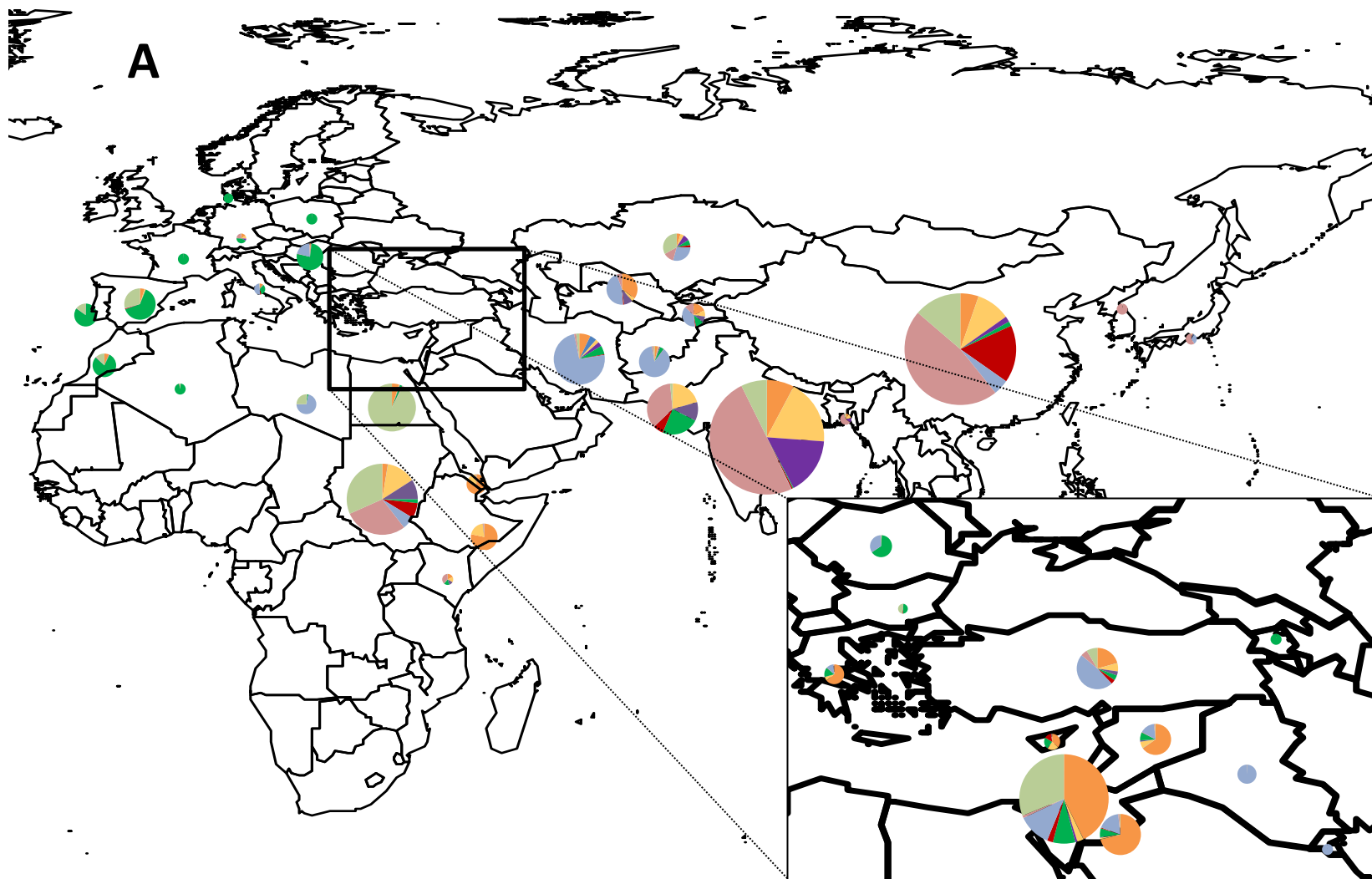
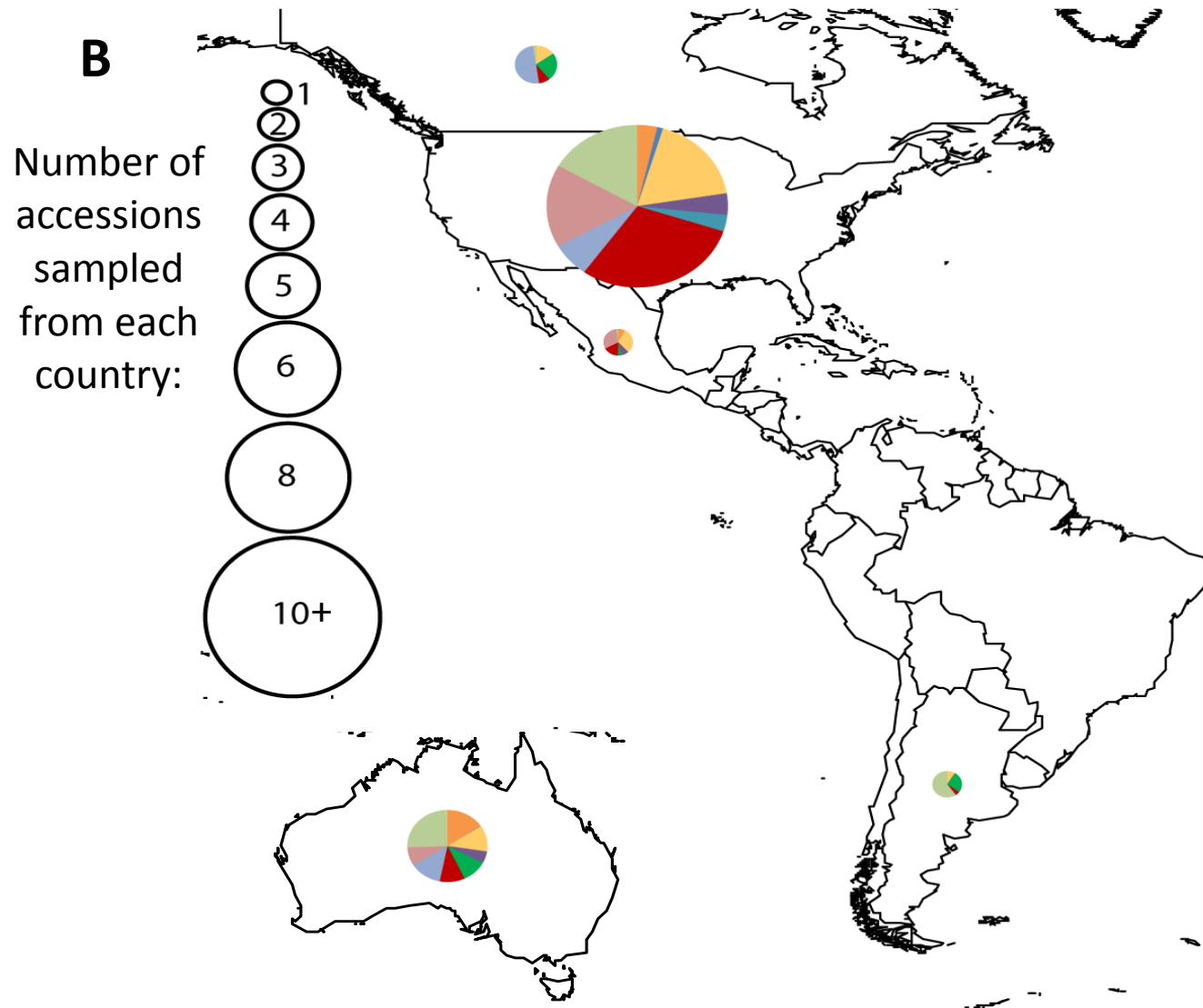


Figure 3.2. Map of sampling locations a) Old World and b) New World USDA accessions and corresponding assignment to each the nine clusters depicted in Figure 3.1. Pie charts are placed on the map to represent samples collected from each country, sizes of the pie charts represent the number of samples collected in each country (as indicated in the legend), and colors are consistent with Figure 3.1.





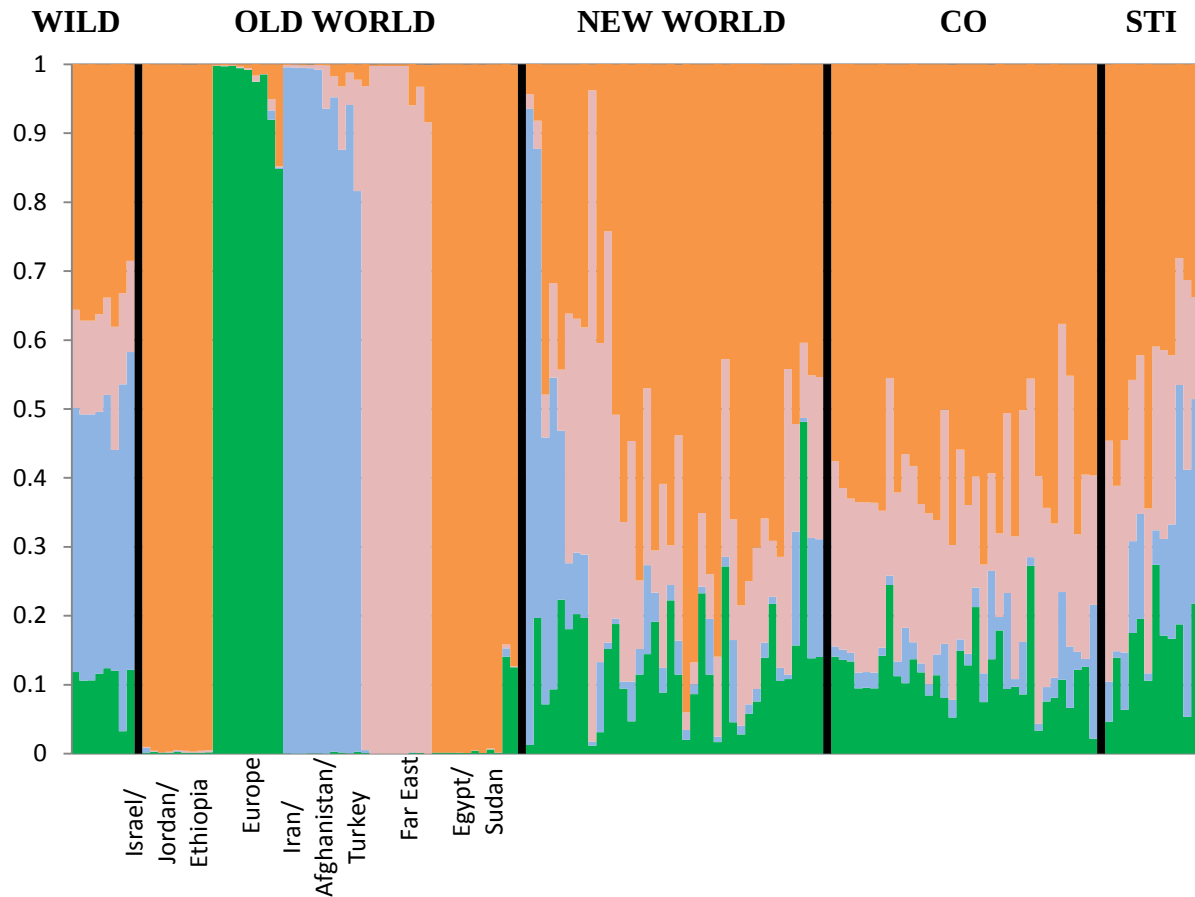


Figure 3.3. STRUCTURE plot of 142 *Carthamus* individuals, in which wild, New World, CO, and STI accessions are assigned to one of four Old World populations. Black bars are used to separate pre-defined groupings of *Carthamus* (labels found at top of graphs). Each vertical bar represents a single individual, and the proportion of membership to each population is indicated on the y-axis. Old World accessions with less than 50% membership to any of these four populations were excluded from the analysis. Colors correspond to Figure 3.1, in which orange = cluster 2, green = cluster 3, light blue = cluster 4, and pink = cluster 5.

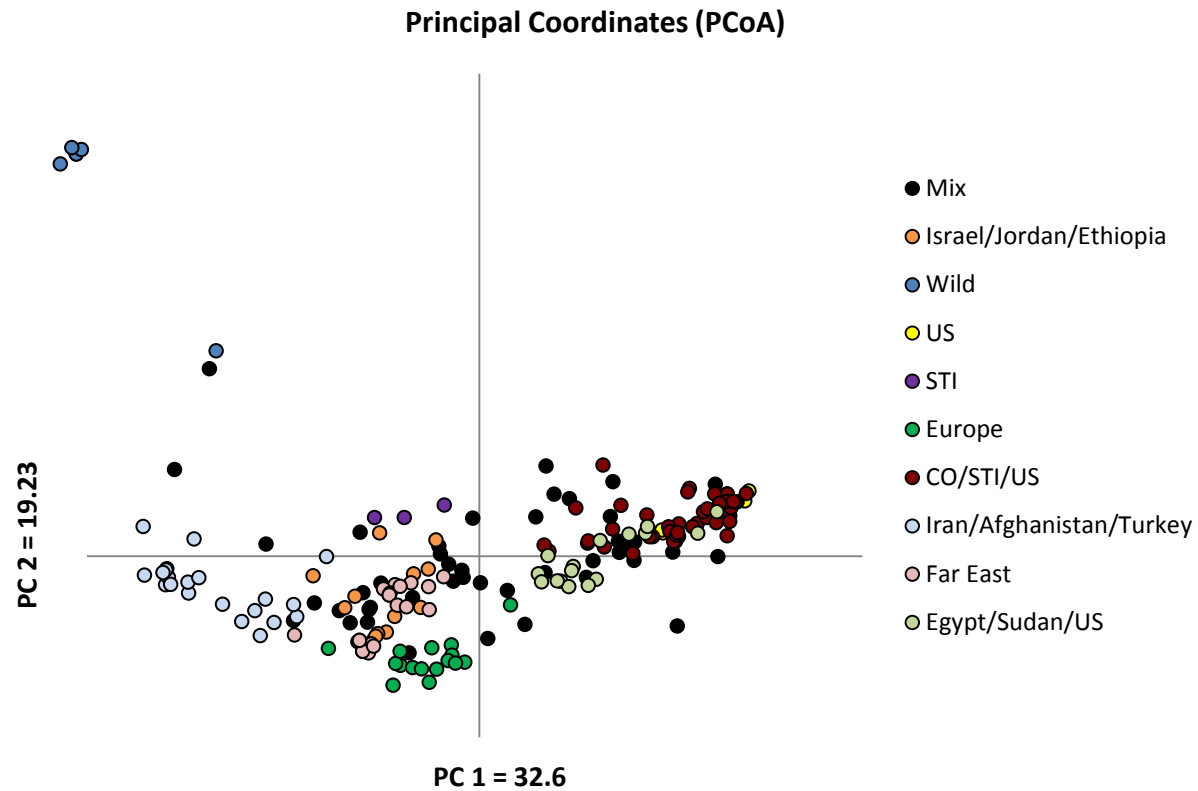


Figure 3.4. Principal coordinate analysis (PCoA) of 190 *Carthamus* individuals. Color-coding corresponds with the nine clusters depicted in Figure 3.1. Black circles represent individuals that have less than 50% membership to any given cluster and therefore a greater proportion of their genotypes are mixes.

CHAPTER IV

MOLECULAR EVOLUTIONARY ANALYSES OF SAFFLOWER AND ITS WEEDY
RELATIVES IN THE CARDUEAE (ASTERACEAE)³

³ Pearl, S.A., M.R. McKain, and J.M. Burke. To be submitted to Journal of Molecular Evolution.

ABSTRACT

This work investigated patterns of selection among crops and weeds within the Cardueae (thistle) tribe within the Asteraceae. We estimated rates of nonsynonymous vs. synonymous substitution among homologous genes within 1645 orthogroups (i.e., gene families) encompassing these five species. Of these 1645 orthogroups, we detected evidence of positive selection within 41. No single lineage resulted from a greater number of positively selected genes than the others, and in four cases, diversifying selection drove evolution in the same homologous gene in all five lineages. Genes with similar functions that were selected in both crop and weed lineages include metabolic enzymes, RING zinc finger domains (implicated in stress response), and a variety involved in protein synthesis. Nearly all of the genes within the same orthogroup mapped to the same part of the safflower genome with high percent identities (>90%), validating the quality of our assemblies and orthogrouping. Further, 12 genes containing positively selected sites colocalized with previously identified domestication quantitative trait loci (QTL), one of which contained a domain found in lipid trafficking proteins that colocalized with oleic and linoleic acid QTL. Taken together, our results show that, similar to the manner in which crops and weeds both have shared and have unique characteristics, they also share and have their own sets of unique, positively selected genes that have shaped their separate evolutionary trajectories.

INTRODUCTION

Domestication is arguably the most important innovation of human societies in the past 13,000 years [1]. It is thus a wonder that so few species have been domesticated. In fact, out of the estimated 300,000 plant species that presently exist [2], fewer than 1% have been domesticated by humans [1, 3]. Given the central importance of agriculture to human survival, why have so few plant species been domesticated? Several lines of evidence suggest that this paucity is not due to a lack of human effort to cultivate wild species. For example, the rapid rise and spread of domesticates (in some cases with multiple origins of domestication; e.g., rice [4], common bean [5, 6], and barley [7]), identification of many species that were regularly harvested but never domesticated (e.g., goosefoot, sumpweed, and knotweed [3, 8, 9]), and failure of modern breeders to significantly add to the list of domesticated species all suggest that other intrinsic factors have prevented more widespread domestication [1].

Life history traits are among the factors that may predispose a given species to domestication. Annual species are favorable candidates for domestication due to their typically rapid growth, short generation times, and relatively high investment in reproduction. Selfing species also tend to be favorable for domestication, particularly in the context of fruit and seed crops, as self-fertilization provides reproductive assurance and promotes high yields [3, 10]. Coincidentally, many of these traits also predispose species to the evolution of weediness. Like crops, weeds thrive in disturbed habitats, but their success depends less on human intervention than does that of crops [11]. Furthermore, weeds are better-adapted to stress and environmental heterogeneity than are crops [12–14]. A unique class of weeds is the agricultural weeds, which are neither fully wild nor domesticated, but have evolved in parallel to crop domestication. Agricultural weeds often mimic certain crop genotypes while exhibiting the seed dispersal and dormancy characteristics present within wild species [15]. The crops and weeds around us are

the products of past selection on distinct subsets of genes present within their progenitor species. The evolution of both types of species has depended on a variety of factors, including effective population size (e.g., [16, 17]), availability of genetic variation (e.g., [18–21]), and genetic architecture of traits comprising the weed and domestication syndromes [1, 3, 10, 22].

The Asteraceae, which is one of the largest and most ecologically successful flowering plant families, includes over 40 economically important crops as well as numerous weedy species [23]. Within the Asteraceae, the Cardueae (thistles) is one of the largest tribes, including over 2,400 species across 73 genera on all continents except Antarctica [24]. Many Cardueae species are problematic weeds, including species of *Cirsium*, *Centaurea*, and *Carthamus* (Figure 4.1). *Cirsium arvense* L. (Canada thistle) is a dioecious, outcrossing, invasive perennial and is the noxious weed species most commonly listed by farmers in the US and Canada [25]. It spreads both sexually and vegetatively via a deep, rhizomatous root system that can only be removed mechanically [26]. Canada thistle outcompetes crops [27] and reduces native species diversity [28]. *Centaurea solstitialis* L. (yellow starthistle) is a self-incompatible, annual herb native to eastern Europe and the Caucasus [29] that has spread across millions of acres in California. It has displaced native plants, reduced forage quality and yield [30], and caused chewing disease in horses [31]. *Carthamus oxyacanthus* Bieb. (jeweled distaff thistle) is an annual weed native to the Middle East that exhibits varying levels of self-incompatibility [32, 33]. Although it is reportedly very rare in the US [34], it is nonetheless a federally listed noxious weed due to the problems it poses as an agricultural weed in Pakistan and northwest India [32].

Carthamus oxyacanthus has reportedly hybridized with both *Carthamus tinctorius* L. (safflower), and *Carthamus palaestinus* Eig. (wild safflower) [35]. Safflower was domesticated in the Fertile Crescent region approximately 4,000 years ago and has since been cultivated on all

continents except Antarctica [32]. It was originally domesticated for its floral pigments (carthamine) and is now largely grown as an oilseed crop, for use in birdseed mixes, and as an ornamental species [32]. Safflower is the only domesticated species in the genus *Carthamus*. Its progenitor, wild safflower, is restricted to southern Israel [36] and does not pose a threat as an agricultural weed.

In the present study, we investigated patterns of molecular evolution in genes from each of these five species. Specifically, we compared nonsynonymous vs. synonymous substitution rates (dN/dS) among homologous sequences from these five species. We focused primarily on the identification of instances of positive selection, with a significant excess of nonsynonymous substitutions being taken as evidence of such selection [37, 38]. Of particular interest was the nature of the genes selected in the various crop and weed lineages and the extent to which these types of genes overlapped across lineages and colocalized with previously identified domestication QTL (Pearl et al., in revision). Given that crops and weeds are both thought to have evolved in response to strong, directional selection, methods aimed at detecting recent bursts of strong selection are ideal for the identification of genes that played an important role in the evolution of both crop and weed species.

METHODS

Plant samples, transcriptome sequencing, and assembly. Plant materials, methods of transcriptome sequencing, and assembly for the majority of the samples used in this study were previously described elsewhere ([39]; Pearl et al., in revision; Table 4.1). The *C. tinctorius* assembly used in this study is a combined assembly of existing Sanger expressed-sequence tag (EST) reads obtained from NCBI dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>; [40]; Table 4.1) and Illumina MiSeq transcriptome sequencing data. To generate the latter, RNA was extracted

separately from bracts, florets, and developing ovules of a single AC Sunset safflower individual using a combined trizol (Invitrogen, Carlsbad, CA) RNeasy mini column (Qiagen, Valencia, CA) method. RNA from each tissue type was separately barcoded during Illumina TruSeq RNA (San Diego, CA) sample preparation and libraries were subsequently sequenced on an Illumina MiSeq with 150 bp paired-end reads. Prior to assembly, barcodes were trimmed using cutadapt v1.2.1 [41] and reads were quality trimmed with a custom script that invoked a sliding window approach to eliminate low quality bases. Cleaned reads were assembled using Trinity v. r2013-02-25 [42]. Contigs with low support (i.e., lowly expressed transcripts and putative sequencing artifacts) were filtered out via RSEM abundance estimation in Trinity and subsequent FPKM filtering using custom scripts. The resulting assembly was combined into a meta-assembly with the Sanger reads using CAP3 [43]. Because Trinity produces multiple isoforms of a single gene and our subsequent selection analyses do not consider alternatively spliced variants, we only retained the longest isoform of each gene.

Orthogroup clustering, sequence alignment, and phylogenetic reconstruction.

Reading frames for each set of unigenes were determined by comparing reads to homologous genes in the 22 plant genome gene family scaffold (dePamphilis et al. in prep; Table C1) via BLASTX (<http://www.ncbi.nlm.nih.gov/BLAST/>), and GENEWISE [44] was subsequently used to translate BLAST hits. Because no Asteraceae genomes are included in the 22 plant genome circumscriptions, this process was repeated using the predicted protein set of the lettuce genome (v4; <https://lgr.genomecenter.ucdavis.edu/>) in order to obtain longer and/or lineage-specific translations. Finally, a maximally-encompassing set of translated unigenes were combined for each of the five Cardueae transcriptomes, and only sequences greater than 100 amino acids in length were retained.

Best BLASTP hits of each species' set of proteins queried against the 22 genome set were used to guide the clustering of unigenes into orthogroups (i.e., clusters of homologous genes derived from a single gene; [45]) via a custom script. For each orthogroup, amino acids were aligned using MUSCLE v3.8.31 [46, 47] and nucleotide sequences were aligned to the amino acid sequences using Pal2nal v13 [48]. Custom scripts were used to eliminate positions at which >10% of the sequences had a gap, merge sequences within species that were $\geq 95\%$ similar, and remove duplicate sequences in each alignment.

Phylogenetic trees were constructed using 500 bootstraps and the GTR+gamma model of nucleotide substitution in RAxML [49]. For the purposes of our analyses, no outgroup was specified. Branch lengths were multiplied by three in order to convert them from the number of nucleotide substitutions per nucleotide site to the number of nucleotide substitutions per codon site, as required for the subsequent selection analyses. Given that long branches may be indicators of erroneously grouped genes and can result in unreliable estimates of dN/dS, branches greater than 7 times the median branch length in each tree were eliminated [50]. Orthogroups with more than four remaining sequences were re-run through RAxML and branch lengths were multiplied by three to create a final set of input trees for the subsequent molecular evolution analyses.

Tests for selection. Two approaches were used to estimate values of dN and dS within orthogroups. The first approach used PAML v4.5 [51, 52], which estimates the instantaneous rate of change between two codons ($\omega = \text{dN/dS}$) via maximum likelihood. PAML describes the substitution at each site via a continuous-time Markov chain with 61 possible states and assumes independence of the substitution at each site [53–55]. Specifically, the sites model in CODEML [56] was invoked to estimate dN and dS at each codon averaged across all branches in each tree

using M0 (one-ratio), M1a (nearly neutral), M2a (positive selection), and M3 (discrete). In each of these models, equilibrium codon frequencies were estimated from the average nucleotide frequencies at the three codon positions (F3X4), and transition/transversion ratios were estimated by iteration of the data. Twice the difference in log-likelihood values of the M0:M3 (4 degrees of freedom) and M1a:M2a (2 degrees of freedom) comparisons were assessed for statistical significance using the χ^2 distribution, where the degrees of freedom equaled the difference in the number of parameters used in each test [52]. These two tests were used to determine whether one versus multiple ω ratios best fit the data and whether any sites exhibited positive selection, respectively. To statistically minimize the false discovery rate (FDR), we compared *P*-values to critical values calculated based on $\alpha = 0.05$ so that fewer than 5% of genes identified to be significant were false positives [57]. The Bayes empirical Bayes procedure implemented in codeml [58, 59] was used to calculate the posterior probability that each site belonged to each site class.

Our second approach used the program fitmodel [60] to assess the potential for variable ω values over branches and sites throughout each orthogroup. Like PAML, fitmodel uses a codon-based model of DNA substitution that involves the maximum likelihood estimation of branch lengths, transition/transversion ratios, equilibrium frequencies of selection classes, and nonsynonymous/synonymous substitution ratios. However, fitmodel models the variability of selection patterns across sites and throughout time using a stochastic process and is more powerful than PAML when there is no *a priori* hypothesis regarding which branch selection may have acted upon [61]. Specifically, we compared the results of model MX+S1 with no selection (ω_1 and $\omega_2 \leq 1$, $\omega_3 = 1$) and selection (ω_1 and $\omega_2 \leq 1$, $\omega_3 > 1$) to determine the most likely ω

ratio describing each site on each branch of the tree, and identified sites with high posterior probabilities ($P > 0.90$) as candidates targets of positive selection [62].

Protein function. Pfam-scan ([ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools](http://ftp.sanger.ac.uk/pub/databases/Pfam/Tools)) was used to search positively selected genes against a Pfam-A v27 [63] library of hidden Markov models (HMMs) constructed by HMMER v3.0 [64], and hits with an e-value better than 1×10^{-6} were retained. Pfam accession numbers were cross-referenced with InterPro [65, 66] via the Pfam database [63], and mapping between Pfam and Gene Ontology was conducted using InterPro and Amigo [67].

Mapping of positively selected genes. Sequences of positively selected genes were mapped to an ultra-dense *C. palaestinus* x *C. tinctorius* genetic map (Bowers et al. unpublished) using BLASTN with an e-value cutoff of 1×10^{-6} . This map was constructed using the same *C. palaestinus* individual and *C. tinctorius* inbred line as those used in this study, and an estimated 60% of the safflower genome is represented in this map. Shared markers between the Bowers et al. map and those used in a Pearl et al. (in revision) domestication QTL mapping study enabled an investigation into whether positively selected genes mapped to regions containing QTL.

RESULTS

Sequence analysis. A total of 38,619,830 base pairs were assembled into 92,351 unigenes using the *C. tinctorius* MiSeq data (Table 4.2). This assembly resulted in a nearly 1000-fold increase in the number of assembled base pairs and a 5-fold increase in the number of assembled unigenes compared to the original Sanger assembly for safflower. Compared to the other transcriptomes, the *C. tinctorius* combined Sanger/MiSeq assembly had the greatest N50 (1368, compared to the others that ranged from 776 to 880; Table 4.2). The total number of translated unigenes in each of the five transcriptome assemblies ranged from 13,706 to 51,173

(average $\pm$ S.E. = 28,382 $\pm$ 6600; Table 4.2). With the exception of the *C. tinctorius* transcriptome, increases in the number of translated unigenes corresponded with increases in the total number of assembled base pairs in each transcriptome (Table 4.2).

The translated unigenes of each of the 5 species clustered into 11,382 total orthogroups. These orthogroups contained sequences of anywhere from two to all five species, with the majority having all five species represented (5,497). The total number of orthogroups was reduced to 9,878 after filtering out gaps and short sequences from alignments, merging identical unigenes, and eliminating orthogroups containing less than three remaining gene sequences. This was further reduced to 5,512 after tree building (which requires multiple sequence alignments containing at least 4 sequences), and finally reduced to 1,644 orthogroups after branch pruning and retaining trees with 5 or more total sequences. Of these 1,644 orthogroups, 638 included all 5 species, and *C. tinctorius* was the species represented in the greatest number of orthogroups (1,561; Figure C1).

Positively selected genes. Codeml and fitmodel identified 72 and 413 orthogroups containing putative positively selected sites, respectively, and 67 of these orthogroups were identified by both programs. After closer inspection of the orthogroups, however, it became apparent that several orthogroups contained what appeared to be non-homologous sequences that had been clustered together simply due to a shared, conserved domain. To remedy this, an all-by-all BLASTN search was performed within each of the 418 putative orthogroups undergoing positive selection. Orthogroups containing at least one sequence with a BLASTN alignment length of less than 90% of the full length of any other sequence within that orthogroup were discarded. Subsequently, 21 and 85 putative positively selected orthogroups identified by codeml and fitmodel remained. Further, because gaps and other alignment errors have been known to

generate false positives in dN/dS analyses [68], each putatively selected site within each orthogroup was manually inspected in MEGA v4 [69], and any site occurring in a potentially misaligned segment (e.g., positions with offset proteins or a string of mismatching proteins within a single sequence) was discarded.

After manual filtering, a total of 3 and 41 positively selected orthogroups identified by codeml and fitmodel remained (Tables 4.3, C2, C3). Because the sites model in codeml only detects selection occurring at the same site across an entire phylogeny, it was not surprising to find that all three orthogroups identified by codeml were also identified by fitmodel as undergoing diversifying selection (i.e., all species within an orthogroup were experiencing positive selection at the same site). In each of these cases, two to three sites were targeted by positive selection (Table 4.3; Figure 4.2). Fitmodel also identified an additional orthogroup undergoing diversifying selection that was not significant in codeml following FDR-correction ($P = 0.003$, critical value = 0.002).

Among the 41 positively selected orthogroups identified by fitmodel, each contained 1 to 4 positively selected sites (mean $\pm$ S.E. = 1.34 ± 0.11). Each of these orthogroups included 5 to 13 sequences (5.59 ± 0.22 , containing up to four sequences from a single species), and the majority (27) included all 5 species (Table C2). The remaining 13 orthogroups had 3 or 4 species represented. The total tree length (number of substitutions per codon site) in the positively selected orthogroups ranged from 0.097 to 0.997 (mean = 0.487 ± 0.013), and the average branch length of each orthogroup ranged from 0.021 to 0.125 (overall mean = 0.059 ± 0.004 ; Table C2). Of the orthogroups containing all five species, 23 exhibited positive selection on one or more of the branches leading to one or more of the *Carthamus* species under study, and 18 of these 23 cases only exhibited selection in *Carthamus* (i.e., positive selection was not identified to be

occurring on either *Cirsium* or *Centaurea*). When combining *Cirsium* and *Centaurea* into a single outgroup category and comparing positive selection trends with each of the three *Carthamus* species, each of the categories contained three to four positively selected genes that were unique to each category (Figure 4.3). Further, there were three genes selected in all three *Carthamus* species, plus the additional four genes undergoing diversifying selection. In the orthogroups with only 3 or 4 species represented, *C. palaestinus* and *C. oxyacanthus* (but never *C. tinctorius*) were among the missing species (Table C3). In these instances, the majority of the positively selected sites were identified in at least one *Carthamus* lineage.

Functional annotation of positively selected genes. Of the 41 putatively selected orthogroups, 5 had no significant hits to the Pfam database and 4 contained domains of unknown function. Among the remaining 32 orthogroups, multiple groups contained domains involved in metabolic and biosynthetic processes (e.g., pectate lyase, riboflavin synthase, and adenylate kinase; Table 4.3; Figure 4.4), and several others were involved in translation and protein synthesis (Table 4.3). Two additional orthogroups contained domains that are found in seed storage proteins. Of the orthogroups containing sites undergoing diversifying selection, one had no significant Pfam hit, another contained a nucleic acid binding domain, a third contained a domain found in proteins involved in lipid trafficking, metabolism, and/or sensing (Figure 4.4), and the last contained a thioredoxin domain, which is found in a wide range of plant proteins involved in photosynthesis, growth, flowering, or seed germination [70].

Mapping of positively selected genes. Twenty-three of the of the forty-one orthogroups containing positively selected genes mapped to the safflower genetic map, and the remaining eighteen orthogroups mapped to contigs that have undetermined positions within the genetic map. The majority of the sequences within positively selected orthogroups shared BLASTed

with >90% identity, and only 7 sequences mapped with 84-90% identity. These seven sequences were all from *C. arvense* and all were likely to be examples of more divergent paralogs. All sequences from each orthogroup mapped to the same part of the safflower genome, with the exception of one sequence (from set 4311). This sequence was the *C. arvense* sequence that mapped with the lowest % identity of all of the sequences. A total of 12 orthogroups colocalized with previously identified domestication QTL, and 9 of these contained domains of known function.

DISCUSSION

In this study, we have identified a set of over 1600 homologous genes from five Cardueae species and have further identified a subset containing sites that are thought to have experienced variable selection pressures. As expected, the majority of the codon positions within the genes in these orthogroups showed clear evidence of purifying selection. Approximately 1400 of these orthogroups also contained neutrally evolving codon positions. In 41 cases, orthogroups contained a third class of positively selected sites in at least one of the five study species (Table C2). The majority of these orthogroups experienced positive selection on one or more of the *Carthamus* lineages. Within the 41 orthogroups exhibiting positively selected sites, 31 of these contained sites in which selection resulted in the exchange of amino acids with different chemical properties. For example, within the TLC domain (set 8059; Figure 4.2), diversifying selection drove the replacement of amino acids with different charges, polarities, and hydrophobicities among several of the *Carthamus*, *Centaurea*, and *Cirsium* lineages. This suggests that diversifying selection may have caused a more drastic change in the function of the TLC domain-containing protein among these five species in these 31 cases compared to the other

10 orthogroups in which selection has resulted in the replacement of amino acids by others with similar chemical properties.

Several genes identified to have been targeted by positive selection in both the crop and weed species were shared and/or thought to have similar functions across lineages. In addition to the genes experiencing diversifying selection among the five focal species (Table C2), these included GTP-binding elongation factors (set 2262 and 2611), RING zinc finger domains involved in stress response (C3H2C3-type in *C. oxyacanthus* [set 4768] and C3HC4-type in *C. tinctorius* [set 3402]; [71]), and several other enzymes involved in catalyzing metabolic reactions.

There was no set of positively selected genes shared only among the three agricultural weeds. However, ribosomal proteins involved in catalyzing mRNA-directed protein synthesis (sets 967 and 8990) were positively selected in both *C. arvense* and *C. solstitialis*, and a cytochrome oxidase (*COX16*; set 6276) assembly protein experienced positive selection in both *C. arvense* and *C. oxyacanthus*. Pectate lyase (set 1008), an enzyme typically found in germinating seeds, pollen, and ripening fruits and implicated in playing a role in pollen tube emergence [72], was positively selected in just *C. arvense*.

There were four genes that were positively selected in just safflower (Figure 4.3; Table C3), including a cupin superfamily domain (set 7215), which is found in a diverse array of proteins such as seed storage proteins. Also included was a nucleoside diphosphate kinase (set 3943), which is involved in making UTP (for polysaccharide synthesis), CTP (for lipid synthesis), and GTP (for protein elongation, signal transduction, and microtubule polymerization). Interestingly, the positively selected CH3C4-type RING zinc finger (set 3402) in safflower has a homolog that showed evidence of positive selection during sunflower

domestication [73]. In contrast, riboflavin synthase (set 3859) was not found to have experienced positive selection in safflower, but was targeted by selection in safflower's progenitor and close, weedy relative. Riboflavin (vitamin B2) is involved in biosynthetic pathways affecting plant growth and defense [74], and foliar application of riboflavin has been shown to induce resistance to bacterial, fungal, and viral pathogens [75].

Nine of the genes that exhibited positive selection mapped to regions of the safflower genetic map that contained previously identified domestication QTL (Pearl et al., in revision). For example, the CH3C4-type RING zinc finger and TLC domains that exhibited positive selection in *C. tinctorius* colocalized with a large cluster of QTL at the beginning of LG H (data not shown). The QTL found in this cluster were for traits that included leaf size and shape, leaf spininess, number of heads, number of selfed seed, achene weight, and oleic and linoleic acid. Notably, the TLC domain is found in proteins with a diverse array of functions, including lipid trafficking [76], and it has been inferred that TLC domains may be involved in long-chain fatty acid biosynthesis (<http://www.uniprot.org/uniprot/F6HQA9>). Although additional fine-mapping will need to be conducted in order to confirm whether any of these genes truly underlie these domestication QTL, many of them provide promising candidates for future study.

In summary, we have leveraged previously developed transcriptome assemblies (in combination with a newly generated *C. tinctorius* assembly) to investigate patterns of selection in three agricultural weeds, an underutilized crop, and its wild progenitor. Our results suggest that selection on genes involved in metabolism and protein synthesis are important in both crops and weeds, while there were also non-overlapping sets of genes that were targeted by selection during the evolution of each of these separate lineages. Going forward, molecular evolutionary analyses among other sets of crops and weeds, coupled with further characterization of genes

that exhibited positive selection, will provide broader insights into the evolution of crop and weed species.

REFERENCES

1. Diamond J: **Evolution, consequences and future of plant and animal domestication.** *Nature* 2002, **418**:700–707.
2. Harlan JR: *Crops and Man*. 2nd ed. Madison, WI: American Society of Agronomy; 1992.
3. Diamond J: *Guns, Germs, and Steel*. New York, New York: W.W. Norton & Company, Inc.; 1997.
4. Londo JP, Chiang Y-C, Hung K-H, Chiang T-Y, Schaal B a: **Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*.** *Proceedings of the National Academy of Sciences* 2006, **103**:9578–9583.
5. Sonnante G, Stockton T, Nodari RO, Becerra Velásquez VL, Gepts P: **Evolution of genetic diversity during the domestication of common-bean (*Phaseolus vulgaris* L.).** *Theoretical and applied genetics* 1994, **89**:629–635.
6. Gepts P, Osborn T, Rashka K, Bliss F: **Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication.** *Economic Botany* 1986, **40**:451–568.
7. Morrell PL, Clegg MT: **Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent.** *Proceedings of the National Academy of Sciences* 2007, **104**:3289–3294.
8. Delcourt PA, Delcourt HR, Ison CR, Sharp WE, Gremillion KJ: **Prehistoric human use of fire, the eastern agricultural complex, and Appalachian oak-chestnut forests: Paleoecology of Cliff Palace Pond, Kentucky.** *American Antiquity* 1998, **63**:263–278.
9. Gremillion KJ: **Plant husbandry at the Archaic/Woodland transition: Evidence from the Cold Oak shelter, Kentucky.** *Midcontinental Journal of Archaeology* 1993:161–189.
10. Gepts P: **Crop domestication as a long-term selection experiment.** *Plant Breeding Reviews* 2004, **24.2**:1–44.
11. De Wet JMJ, Harlan JR: **Weeds and domesticates: evolution in the man-made habitat.** *Economic Botany* 1975, **29**:99–108.
12. Baker H: **Characteristics and modes of origin of weeds.** In *The genetics of colonizing species*. Edited by Baker H, Stebbins G. New York: Academic Press, Inc.; 1965:147–168.

13. Baker HG: **The evolution of weeds.** *Annual Review of Ecology and Systematics* 1974, **5**:1–24.
14. Sakai AK, Allendorf FW, Holt JS, Lodge M, Molofsky J, With KA, Cabin RJ, Cohen JE, Norman C, Mccauley DE, Neil PO, Parker M, Thompson JN, Weller SG: **The population biology of invasive species.** *Annual Review of Ecology and Systematics* 2001:305–332.
15. Vigueira CC, Olsen KM, Caicedo a L: **The red queen in the corn: agricultural weeds as models of rapid adaptive evolution.** *Heredity* 2013, **110**:303–311.
16. Wright S: **Evolution in Mendelian populations.** *Genetics* 1931, **16**:97–159.
17. Ross-Ibarra J: **Recombination, genetic diversity, and plant domestication.** The University of Georgia; 2006.
18. Darwin C: *On the Origin of Species by Natural Selection.* 1859.
19. Darwin C: *The Variation of Animals and Plants Under Domestication.* 1868.
20. Avise JC, Hamrick J (Eds): *Conservation Genetics.* Springer; 1996.
21. Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits.* 1998.
22. Burger JC, Chapman MA, Burke JM: **Molecular insights into the evolution of crop plants.** *American Journal of Botany* 2008, **95**:113–122.
23. Dempewolf H, Rieseberg LH, Cronk QC: **Crop domestication in the Compositae: a family-wide trait assessment.** *Genetic Resources and Crop Evolution* 2008, **55**:1141–1157.
24. Susanna A, Garcia-jacas N: **Cardueae (Carduoideae).** In *Systematics, evolution, and biogeography of the Compositae.* Edited by Funk VA, Susanna T, Stuessy T, Bayer R. Vienna, Austria: International Association for Plant Taxonomy; 2009:293–313.
25. Skinner K, Smith L, Rice P: **Using noxious weed lists to prioritize targets for developing weed management strategies.** *Weed Science* 2000, **48**:640–644.
26. Evans J: **Canada thistle (Cirsium arvense): a literature review of management practices.** *Natural Areas Journal* 1984, **4**:11–21.
27. Malicki L, Berbeciowa C: **Uptake of more important mineral components by commonfield weeds on loess soil.** *Acta Agrobotanica* 1986, **39**:129–142.
28. Stachon W, Zimdahl RL: **Allelopathic Activity of Canada Thistle (Cirsium arvense) in Colorado.** *Weed Science* 1980, **28**:83–86.

29. Maddox DM, Mayfield A, Poritz NH: **Distribution of yellow starthistle (*Centaurea solstitialis*) and Russian Knapweed (*Centaurea repens*).** *Weed Science* 1985, **33**:315–327.
30. Pitcairn MJ, Schoenig S, Yacoub R, Gendron J: **Yellow starthistle continues its spread in California.** *California Agriculture* 2006, **60**:83–90.
31. Kingsbury J: *Poisonous Plants of the United States and Canada*. Englewood Cliffs, NJ: Prentice-Hall, Inc.; 1964.
32. Dajue L, Mündel H-H: *Safflower. Carthamus Tinctorius L. Promoting the Conservation and Use of Underutilized and Neglected Crops*. Gatersleben, Germany and Rome, Italy; 1996.
33. Chapman MA, Burke JM: **DNA sequence diversity and the origin of cultivated safflower (*Carthamus tinctorius* L.; Asteraceae).** *BMC Plant Biology* 2007, **7**:60.
34. **USDA plants database.** [<http://plants.usda.gov/core/profile?symbol=CAOX6>]
35. Ashri A, Knowles PF: **Cytogenetics of safflower (*Carthamus* L.) species and their hybrids.** *Agronomy Journal* 1960, **52**:11–17.
36. Hanelt P: **Monographische übersicht der gattung *Carthamus* L. (Compositae).** *Feddes Repertorium* 1963, **67**:41–180.
37. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Molecular biology and evolution* 1986, **3**:418–426.
38. Yang Z, Bielawski J: **Statistical methods for detecting molecular adaptation.** *Trends in ecology & evolution* 2000, **15**:496–503.
39. Lai Z, Kane NC, Kozik A, Hodgins KA, Dlugosch KM, Barker MS, Matvienko M, Yu Q, Turner KG, Pearl SA, Bell GDM, Zou Y, Grassa C, Guggisberg A, Adams KL, Anderson J V, Horvath DP, Kesseli R V, Burke JM, Michelmore RW, Rieseberg LH: **Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression.** *American Journal of Botany* 2012, **99**:209–218.
40. Chapman MA, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore RW, Tang S, Knapp SJ, Burke JM: **Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (*Carthamus tinctorius* L.).** *Theoretical and Applied Genetics* 2009, **120**:85–91.
41. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet journal* 2011, **17**.
42. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome**

assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011, **29**:644–652.

43. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Research* 1999, **9**:868–877.

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

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

46. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research* 2004, **32**:1792–1797.

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

48. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Research* 2006, **34**:W609–W612.

49. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688–90.

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

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

52. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Molecular Biology and Evolution* 2007, **24**:1586–1591.

53. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929–936.

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

55. Yang Z: **Maximum-likelihood models for combined analyses of multiple sequence data.** *Journal of molecular evolution* 1996, **42**:587–596.

56. Bielawski JP, Yang Z: **A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution.** *Journal of Molecular Evolution* 2004, **59**:121–132.
57. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proceedings of the National Academy of Sciences* 2003, **100**:9440–0445.
58. Yang Z, Wong WSW, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Molecular Biology and Evolution* 2005, **22**:1107–1118.
59. Zhang J, Nielsen R, Yang Z: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.** *Molecular Biology and Evolution* 2005, **22**:2472–2479.
60. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP: **Modeling the site-specific variation of selection patterns along lineages.** *Proceedings of the National Academy of Sciences* 2004, **101**:12957–12962.
61. Lu A, Guindon S: **Performance of standard and stochastic branch-site models for detecting positive selection amongst coding sequences.** *Molecular Biology and Evolution* 2013.
62. Guindon S, Black M, Rodrigo A: **Control of the false discovery rate applied to the detection of positively selected amino acid sites.** *Molecular biology and evolution* 2006, **23**:919–926.
63. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic Acids Research* 2012, **40**:D290–D301.
64. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755–763.
65. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, Sangrador-Vegas A, Yong S-Y, Mulder N, Hunter S: **Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation.** *Database* 2012, **2012**.
66. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, De Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, et al.: **InterPro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Research* 2012, **40**(Database issue):D306–D312.
67. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**:288–289.

68. Jordan G, Goldman N: **The effects of alignment error and alignment filtering on the sitewise detection of positive selection.** *Molecular Biology and Evolution* 2012, **29**:1125–1139.
69. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Molecular Biology and Evolution* 2007, **24**:1596–1599.
70. Arnér ES, Holmgren A: **Physiological functions of thioredoxin and thioredoxin reductase.** *European Journal of Biochemistry* 2000, **267**:6102–6109.
71. Zeba N, Isbat M, Kwon N-J, Lee MO, Kim SR, Hong CB: **Heat-inducible C3HC4 type RING zinc finger protein gene from *Capsicum annuum* enhances growth of transgenic tobacco.** *Planta* 2009, **229**:861–871.
72. Marin-Rodriguez M, Orchard J, Seymour G: **Pectate lyases, cell wall degradation and fruit softening.** *Journal of Experimental Botany* 2002, **53**:2115–2119.
73. Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM: **A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*).** *The Plant Cell* 2008, **20**:2931–45.
74. Zhang S, Yang X, Sun M, Sun F, Deng S, Dong H: **Riboflavin-induced priming for pathogen defense in *Arabidopsis thaliana*.** *Journal of Integrative Plant Biology* 2009, **51**:167–174.
75. Dong H, Beer S V: **Riboflavin induces disease resistance in plants by activating a novel signal transduction pathway.** *Phytopathology* 2000, **90**:801–811.
76. Winter E, Ponting CP: **TRAM, LAG1 and CLN8: members of a novel family of lipid-sensing domains?** *Trends in Biochemical Sciences* 2002, **27**:381–383.

TABLES

Table 4.1. Collection information and tissue types used in the RNA extraction of each of the five species used in this study.

Species	Collection ID	GenBank accessions or doi ^a	Roots	Leaves	Stalk	Bracts	Buds	Mature flowers	Florets	Ovules
<i>Cirsium arvense</i>	NW-22-1-M	doi: 10.5061/dryad.cm7td/14	X	X			X	X		
<i>Centaurea solstitialis</i>	AR-13-24	doi: 10.5061/dryad.cm7td/4		X						
<i>Carthamus oxyacanthus</i>	PI 407602	doi: 10.5061/dryad.cm7td/15	X	X		X			X	X
<i>Carthamus palaestinus</i>	PI 235663	N/A ^b		X		X			X	X
<i>Carthamus tinctorius</i> ¹	PI 592391	N/A ^b				X			X	X
<i>Carthamus tinctorius</i> ²	PI 592391	EL372565-EL412381, EL511108-EL511145	X	X	X					

*Information on *C. arvense*, *C. solstitialis*, and *C. oxyacanthus* adapted from Lai et al. (2012)

^aTranscriptome assemblies available at Dryad data repository (<http://dx.doi.org>)

^b*C. palaestinus* and *C. tinctorius* assemblies will be made available upon publication.

¹Refers to *C. tinctorius* Sanger assembly

²Refers to *C. tinctorius* MiSeq assembly

Table 4.2 Sequencing information regarding each of the species used in this study.

Species	Technology	No. raw reads	Total sequenced (mbp)	Total assembled length (mbp)	No. unigenes	N50	No. translated sequences
<i>Cirsium arvense</i>	454	3,770,510	1,411	61	88,374	821	51,173
<i>Centaurea solstitialis</i>	454	649,880	274	32	43,503	880	29,608
<i>Carthamus oxyacanthus</i>	454	406,005	125	40	27,255	776	13,706
<i>Carthamus palaestinus</i>	454	564,681	266	27	35,138	853	16,951
<i>Carthamus tinctorius</i>	Sanger	40,879	30	15	19,854	804	N/A
<i>Carthamus tinctorius</i>	Illumina MiSeq	38,619,830	5,792	79	92,351	1462	N/A
<i>Carthamus tinctorius</i>	Sanger/Illumina combined	38,660,709	5,822	79	93,334	1368	30,472

*Information from *C. arvense*, *C. solstitialis*, and *C. oxyacanthus* adapted from Lai et al. (2012).

Table 4.3. Functional annotations of genes exhibiting positive selection.

Set ID	No. positively selected sites	Positively selected branch*	PfamID	Description
2648	1	Co	DUF962	Domain of unknown function
2635	1	Cp, Ct	No hit	
7215	1	Ct	Cupin_2	Found in a diverse array of proteins, including those involved in seed storage
8816	2	Cp, Ct, Ce	Carb_kinase	Carbohydrate kinase domain in protein of unknown function; Molecular function: ATP binding; Cellular component: Mitochondrion
3402	1	Ct	zf-C3HC4_3	Zinc RING finger, involved in abiotic stress response
1008	1	Ci	Pec_lyase_C	Enzyme involved in maceration and rotting of plant tissue; Expressed in pollen tube late in development; Molecular function: Pectate lyase activity
5547	1	Co	No hit	
5925	1	Cp, Ct	Indigoidine_A	Implicated in protection from oxidative stress; Molecular function: Hydrolase activity, acting on hydrogen bonds
6970	1	Ct, Cp	UMPH-1	Dephosphorylates nucleic acids; Molecular function: Magnesium ion binding; Cellular component: cytoplasm
1198	1	Co, Ct	HAD_2	Haloacid dehalogenase-like hydrolase
2039	3	Ce, Ct	Transket_pyr, Transketolase_C	Links glycolytic acid and pentose-phosphate pathways
7143	1	Ct	DUF2930	Domain of unknown function
3407	1	Co	TRAM_LAG1_CLN8	Involved in lipid trafficking, metabolism, or sensing
2076	3	All 5	Thioredoxin	Involved in photosynthesis, growth, flowering, or seed germination; Molecular function: Protein binding; Cellular component: Nucleoplasm; Biological process: Negative regulation of transcription from RNA polymerase II promoter
8235	2	Cp	Pkinase	Protein phosphorylation; Molecular function: Protein kinase activity; Biological process: Protein phosphorylation
12423	1	Ce	No hit	
7594	1	Cp	DUF3119	Domain of unknown function
3859	1	Co, Cp	DMRL_synthase	Riboflavin synthase; Cellular component: Riboflavin synthase complex
3943	1	Ct	NDK	Synthesizes NTPs required for lipid synthesis, polysaccharide synthesis, protein elongation, signal transduction, and microtubule polymerization; Molecular

				function: Nucleoside diphosphate kinase activity; Biological process: UTP biosynthetic process
8059	2	All 5	TRAM_LAG1_CLN8	Involved in lipid trafficking, metabolism, or sensing; Cellular component: Integral to membrane
1780	1	Ci	Epimerase	Molecular function: Catalytic activity; Biological process: Cellular metabolic process
11119	1	Ct	CK2S	Casein kinase 2 substrate
5754	1	All 5	RRM_1	RNA binding domain; Molecular function: Nucleic acid binding
1325	3	Ce, Co, Ct	Ferritin_2	Ferritin-like domain; Molecular function: Structural constituent of the ribosome; Cellular component: Ribosome; Biological process: Translation
967	1	Ce	Ribosomal_L35Ae	Ribosomal protein involved in catalyzing mRNA-directed protein synthesis
4768	1	Co	zf-RING_2	Zinc RING finger; Biological process: Zinc ion binding
591	1	Inner branch only	PGK	Phosphoglycerate kinase; Molecular function: Phosphoglycerate kinase activity; Biological process: Glycolysis
948	4	Ct, Co, Cp	LTP_2	Probably involved in lipid transfer
6276	1	Ci, Co	COX16	Cytochrome oxidase assembly protein; Cellular component: Mitochondrial membrane
3843	2	Co	Bystin	Putative role in early rRNA processing
7305	1	Ce	TFIIS_C	Transcription factor S-II; Molecular function: Nucleic acid binding; Biological function: Transcription, DNA-dependent
4930	1	Inner branch only	No hit	
12570	1	Co	F-box-like	F-box-like family
4311	1	Cp	SUI1	Translation initiation factor SUI1; Molecular function: Translation initiation factor activity; Biological process: Translational initiation
8948	1	All 5	No hit	
2238	1	Co	ADK, ADK_lid	Adenylate kinase; Molecular function: Nucleobase-containing compound kinase activity; Biological process: Nucleobase-containing compound
8058	1	Cp	UDPGP	UTP--glucose-1-phosphate uridylyltransferase; Molecular function: Nucleotidyltransferase activity; Biological process: Metabolic process
5390	1	Co, Cp	IGPS	Involved in biosynthesis of phenylalanine, tyrosine, and tryptophan; Molecular function: Indole-3-glycerol-phosphate synthase activity
2611	2	Ct, Cp, Co	GTP_EFTU	Promotes binding of aminoacyl tRNA to the A site of ribosomes during protein biosynthesis; Molecular function: GTP binding

8990	1	Ci, Ce	Ribosomal_L18e	Catalyzes mRNA-directed protein synthesis
			GTP_EFTU, GTP_EFTU_D2, eIF2_C	Promotes binding of aminoacyl tRNA to the A site of ribosomes during protein biosynthesis; Molecular function: GTP binding
2262	1	Ci		

*Ct = *C. tinctorius*; Cp = *C. palaestinus*; Co = *C. oxyacanthus*; Ce = *C. solstitialis*; Ci = *C. arvense*

FIGURES

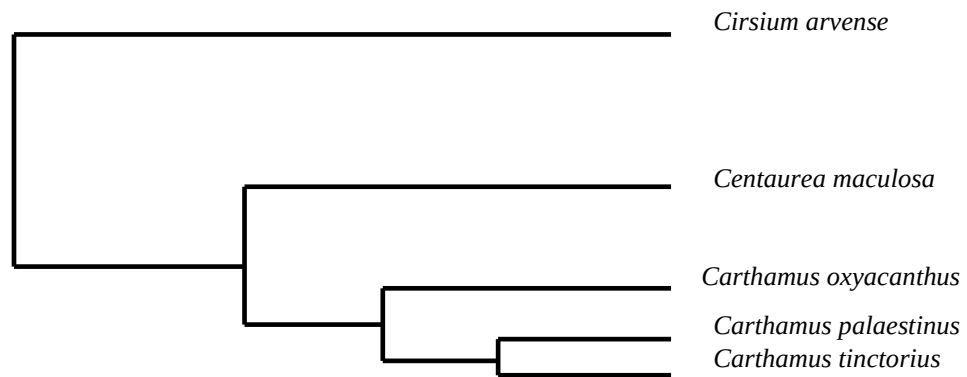


Figure 4.1. Phylogenetic relationships among the species used in this study, divergence times not to scale.

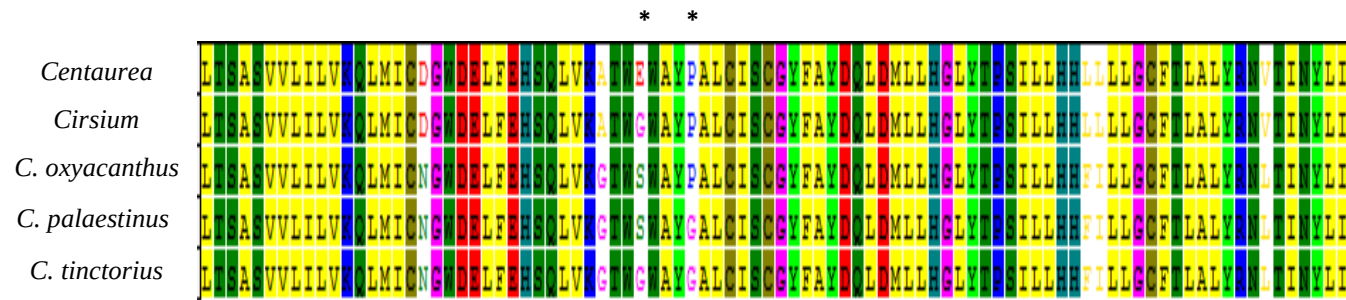


Figure 4.2. Example orthogroup (set 8059) in which diversifying selection is acting upon two codon sites (proteins indicated by an asterisk). This orthogroup contains the TLC domain, which is found in proteins involved in functions that include lipid trafficking, metabolism, or sensing.

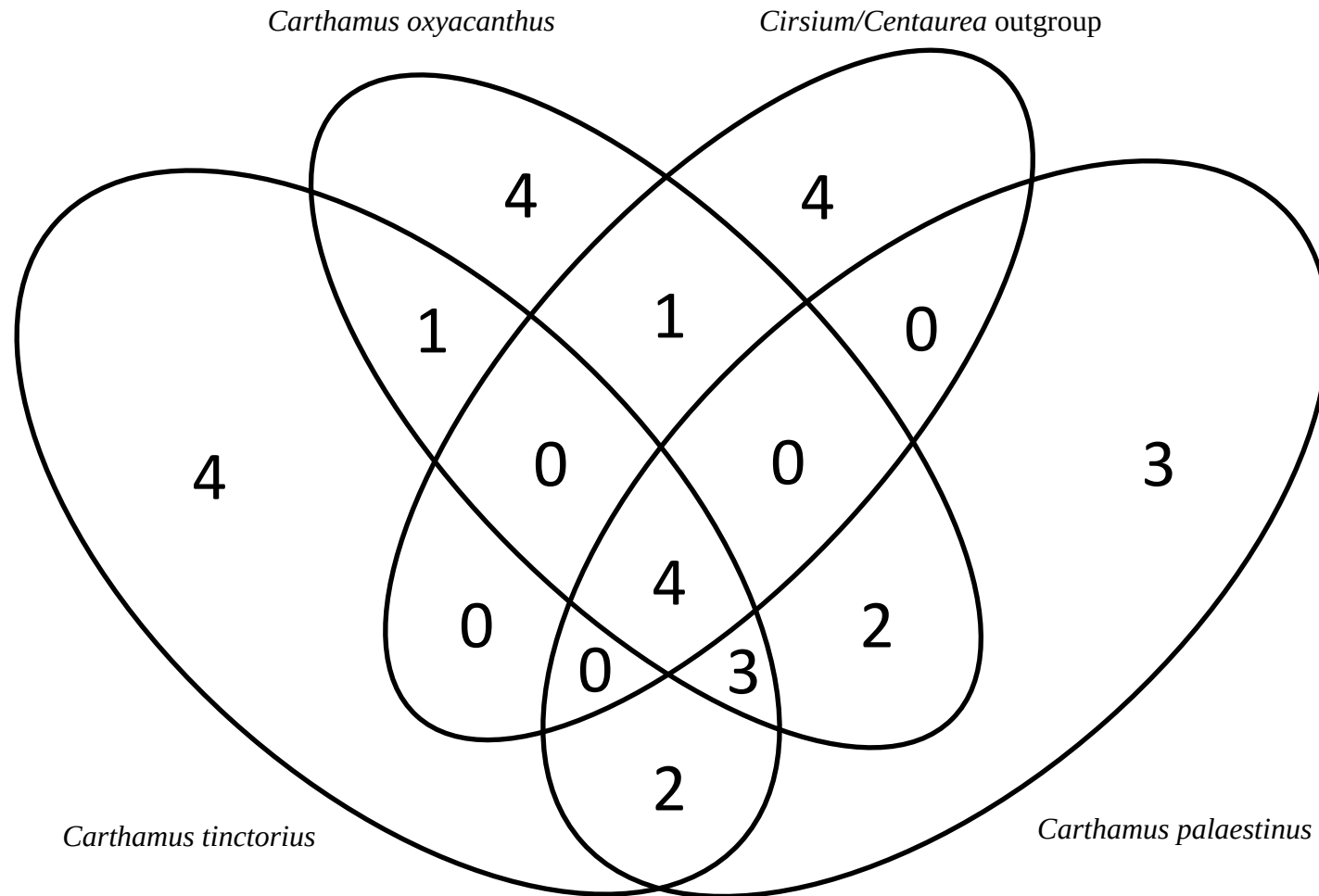


Figure 4.3. Venn diagram depicting positively selected genes that are shared among and unique to *C. tinctorius*, *C. palaestinus*, *C. oxyacanthus*, and the *Centaurea/Cirsium* outgroup. This summary is specific to the orthogroups in which all five species are represented.

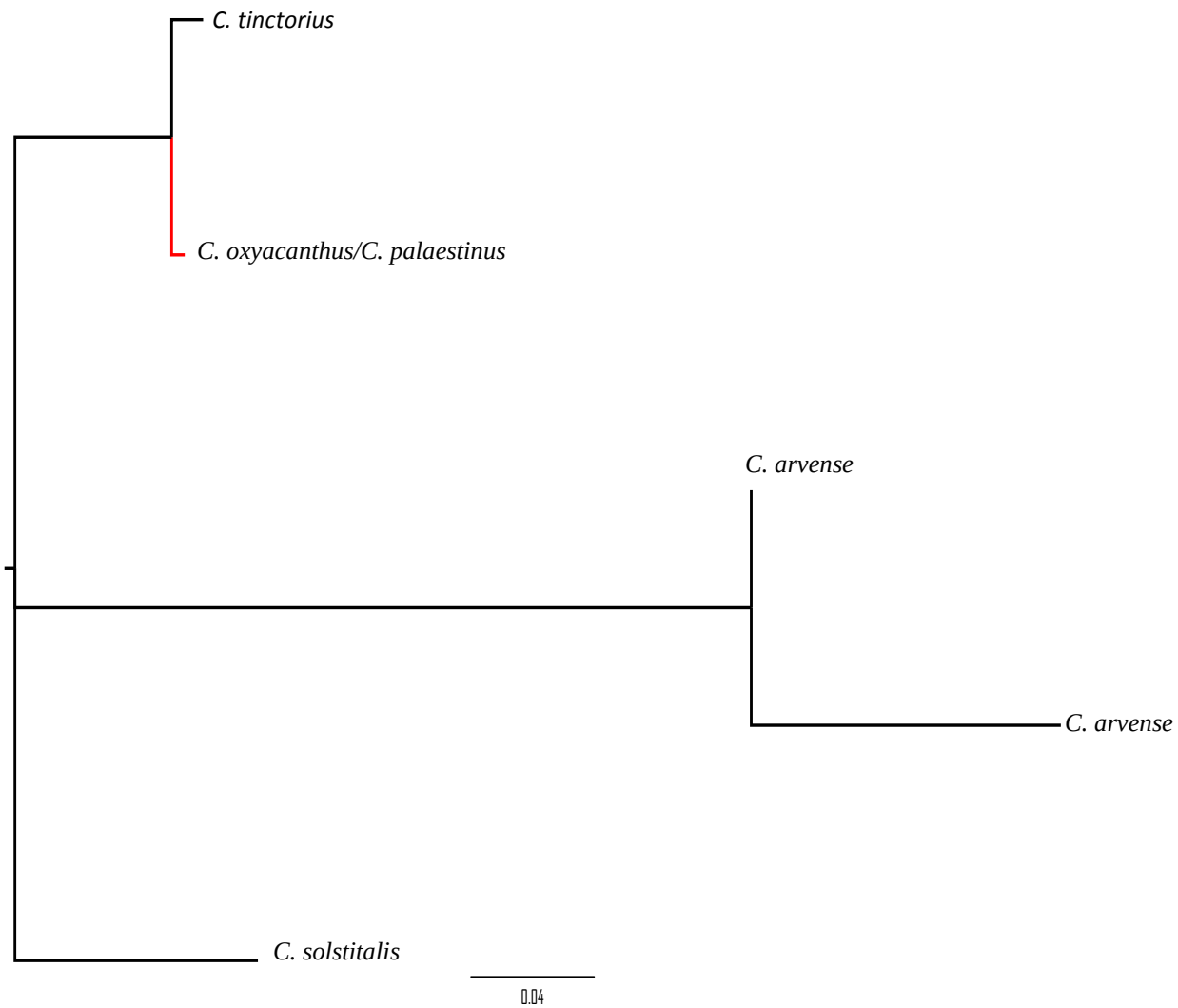


Figure 4.4. Example orthogroup (set 3859) in which positive selection is acting jointly on *C. oxyacanthus* and *C. palaestinus*, which share the same protein sequence. This orthogroup corresponds to the riboflavin synthase protein.

CHAPTER V

CONCLUSIONS

This work assessed the genetics of phenotypic variation in safflower and its relatives. It addressed questions regarding genetic architecture, parallel evolution, comparative genomics, genetic diversity, and selection within the context of domestication. First, QTL mapping was used to investigate the genetic architecture of domestication traits in safflower. This analysis revealed that the genetic architecture of safflower domestication is complex, where most traits are controlled by multiple genes with small to moderate effects on the phenotype. To date, safflower and sunflower are the only two crops in which traditional QTL studies have revealed a complex genetic architecture [1]. This may be due to the fact that safflower and sunflower are morphologically quite similar to their progenitors. It is thought that traits with simple genetic architecture (e.g., controlled by just one or two genes) are major structural traits and presence/absence characters [2]. This is consistent with the simple genetic architecture for traits such as branching and glume architecture in maize [3, 4]. However, more recent population genomic studies have, in fact, revealed that domestication has involved selection on more genes than originally thought [5, 6].

The identification of safflower domestication QTL also provided the opportunity to investigate whether parallel genotypic evolution has occurred during the independent domestication of two Asteraceae oilseed crops, safflower and sunflower. Through comparative QTL mapping, we identified 15 colocalizing QTL for 10 different domestication traits, including oil quantity and quality, fruit size, and flowering time. While additional work, including fine-

mapping, positional cloning, and functional analyses, will be required to establish with certainty that the same underlying genes are responsible for these instances of colocalization, our findings have been consistent with parallel trait evolution having been driven by parallel genotypic changes. This work provides insights into the predictability of evolution, and comparative genomic analyses such as these have the potential to aid in the improvement of other crops about which less is known.

The molecular markers developed for addressing questions in the first part of this dissertation were also used in an investigation of genetic diversity in a broad collection of safflowers. Such investigations of genetic diversity are an important first step in the furthered development of underutilized crop species. This work provided insights into the domestication history of safflower and assessed the genetic variability within commercial breeding pools, revealing a decrease in allelic variation that has accompanied the domestication bottleneck and showing that past germplasm conversion efforts have provided a marginal benefit. The results of this study also indicated that the majority of safflower breeding material originated from just two of the five Old World safflower centers of diversity. Therefore, it was possible to identify sources for the introgression of novel allelic variation that would result in increased gains during future commercial safflower improvement. The introgression of novel genetic diversity from crop wild relatives is an established practice that has proved its utility in other crops [7].

The final portion of this dissertation focused on comparing patterns of positive selection among crop and weed species within the Cardueae tribe. This work leveraged transcriptome resources that were generated for the development of markers used in the first part of this dissertation, as well as other transcriptomes that were developed by the Compositae Genome Project. Specifically, rates of nonsynonymous vs. synonymous substitutions were estimated

among homologous sets of genes from *Cirsium arvense*, *Centaurea solstitialis*, *Carthamus oxyacanthus*, *Carthamus tinctorius*, and *Carthamus palaestinus*. This study revealed that no single lineage resulted from a greater number of positively selected genes than the others, and in four cases, diversifying selection drove evolution in the same homologous gene in all five lineages. Among the genes that were putatively selected in both our crop and weed sample included metabolic enzymes, RING zinc finger domains (implicated in stress response), and a variety of proteins involved in protein synthesis. Selection on ribosomal proteins and cytochrome oxidase were unique to two separate weed lineages. Among the positively selected genes identified in just the crop included a diphosphate kinase. These enzymes synthesize NTPs, including the CTPs required for lipid synthesis and of essential importance in an oilseed crop.

This final study also leveraged an ultra-dense genetic map that has been developed from the same mapping population generated in the QTL analysis (Bowers et al. unpublished). The majority of the positively selected genes identified in this last study mapped with a high percent similarity to the safflower genome, and we were therefore able to identify cases in which these genes colocalized with the safflower domestication QTL identified in the first study. Only nine genes with domains of known function colocalized with these domestication QTL. In one case, a gene potentially involved in lipid trafficking and long-chain fatty acid synthesis colocalized with oil quality QTL.

This dissertation has contributed to a body of knowledge on the genetics of adaptation, using domestication as a model. Taken together, it has added to a growing understanding that the genetic architecture of domestication traits are relatively complex, and it has also provided the first investigation within the Asteraceae of parallel domestication. During the course of this dissertation, a breadth of genetic resources have been developed for safflower, including genetic

maps, molecular markers, and transcriptome sequences (Pearl et al. in revision, Bowers et al. unpublished, [8]). Further, connections have been made directly with the North American safflower breeding community, and one company has since begun to utilize the markers developed here. Future work can now begin to target candidate domestication genes and identify desirable, functional variation in wild safflower species. This work will benefit from the continued development and maintenance of public-private partnerships among producers, breeders, and basic science researchers.

REFERENCES

1. Burke JM, Tang S, Knapp SJ, Rieseberg LH: **Genetic analysis of sunflower domestication.** *Genetics* 2002, **161**:1257–1267.
2. Gottlieb LB: **Genetics and morphological evolution in plants.** *The American Naturalist* 1984, **123**:681–709.
3. Doebley J, Stec A: **Genetic analysis of the morphological differences between maize and teosinte.** *Genetics* 1991, **129**:285–295.
4. Doebley J, Stec A, Hubbard L: **The evolution of apical dominance in maize.** *Nature* 1997:485–488.
5. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS: **The effects of artificial selection on the maize genome.** *Science* 2005, **308**:1310–4.
6. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright R a, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner R a, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J: **Comparative population genomics of maize domestication and improvement.** *Nature Genetics* 2012, **44**:808–811.
7. Hajjar R, Hodgkin T: **The use of wild relatives in crop improvement: a survey of developments over the last 20 years.** *Euphytica* 2007, **156**:1–13.
8. Lai Z, Kane NC, Kozik A, Hodgins KA, Dlugosch KM, Barker MS, Matvienko M, Yu Q, Turner KG, Pearl SA, Bell GDM, Zou Y, Grassa C, Guggisberg A, Adams KL, Anderson J V, Horvath DP, Kesseli R V, Burke JM, Michelmore RW, Rieseberg LH: **Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression.** *American Journal of Botany* 2012, **99**:209–218.

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER II

Text A1. Library preparation and normalization for 454 sequencing of the *C. palaestinus* transcriptome followed previously described protocols (see text). Briefly, because the 454 technology is unable to accurately sequence strings of homopolymers, a modified “broken chain” oligo-dT primer was used during cDNA synthesis to reduce the length of the mononucleotide runs associated with the 3’ poly-A tail in mRNA. The resultant cDNA was amplified and normalized using the TRIMMER DIRECT cDNA normalization kit (Evrogen, Moscow, Russia). After normalization, the cDNA was sheared into 500-800 base pair fragments and any remaining small fragments were removed using AMPure SPRI beads. The fragmented ends of the remaining cDNA were then polished and ligated to adapters and the desired ligation products were selectively amplified and size selected.

The *C. palaestinus* 454 EST library was sequenced on the GS-FLX at the David H. Murdock Research Institute (<http://www.dhmri.org/about.html>) using standard Titanium chemistry (<http://www.454.com/>). The 454 reads were cleaned using ESTclean (<http://sourceforge.net/projects/estclean/>) and the cleaned sequences were assembled in MIRA v. 3.0.3 (http://www.chevreux.org/projects_mira.html) using the assembly mode “denovo, est, accurate, 454” and the settings “-AS:mrl=50 -OUT:sssip=yes -CL:bsqc=yes.”

Table A1. Pairwise Spearman correlation coefficients of the 24 traits phenotyped in this study.

	Root	Leaf sz.	Leaf shp	Spines	Flower	Disc dia	Cap hght	No. heads	Color	Stem hght	No. ints	Int len	Low br.	No. seed	Seed wght	Seed len	Seed wid.	% via.	Dorm.	% oil	% palmitic	% stearic	% oleic
Root																							
Leaf sz.	-0.07																						
Leaf shp	0.05	-0.23																					
Spines	-0.03	0.34	-0.26																				
Flower	0.04	-0.05	-0.22	-0.05																			
Disc dia	-0.09	0.42	0.01	0.31	-0.06																		
Cap hght	-0.09	0.29	-0.22	0.16	0.22	0.20																	
No. heads	0.12	0.26	-0.01	-0.08	-0.12	-0.02	-0.09																
Color	0.13	0.01	0.21	0.10	0.03	0.08	-0.09	0.01															
Stem hght	-0.12	0.58	-0.11	0.16	-0.02	0.31	0.24	0.02	-0.08														
No. ints	-0.02	0.07	-0.04	-0.28	0.21	-0.07	0.13	0.15	-0.06	0.37													
Int len	-0.09	0.49	-0.07	0.36	-0.21	0.34	0.14	-0.05	-0.01	0.65	-0.41												
Low br.	-0.07	-0.42	-0.08	-0.08	0.31	-0.18	0.07	-0.41	-0.17	-0.14	0.05	-0.18											
No. seed	-0.06	0.39	-0.04	0.20	-0.28	0.17	0.03	0.08	-0.04	0.42	0.09	0.33	-0.19										
Seed wght	-0.07	0.12	-0.13	0.06	0.20	0.12	0.28	-0.02	-0.03	0.06	0.10	-0.04	-0.02	-0.56									
Seed len	-0.08	0.16	-0.11	0.03	0.17	0.12	0.28	0.02	-0.11	0.11	0.12	-0.01	-0.03	-0.56	0.82								
Seed wid.	-0.04	-0.08	-0.04	-0.08	0.29	0.02	0.23	-0.07	-0.11	-0.09	0.09	-0.19	0.09	0.730	0.80	0.83							
% via.	-0.03	0.09	-0.03	0.21	-0.10	0.07	-0.15	-0.17	0.29	0.11	-0.09	0.18	-0.12	0.30	-0.19	-0.20	-0.34						
Dorm.	-0.10	-0.12	0.15	-0.36	-0.31	0.034	-0.18	0.18	-0.03	-0.10	0.03	-0.10	-0.10	-0.12	0.01	-0.05	0.01	-0.18					
% oil	0.08	0.11	0.01	0.20	-0.05	0.11	-0.14	-0.07	0.10	0.18	-0.16	0.28	-0.08	0.57	-0.35	-0.49	-0.58	0.22	-0.26				
% palmitic	-0.07	0.09	-0.16	0.18	0.12	0.09	0.11	-0.13	0.11	0.15	0.13	0.07	0.08	0.01	-0.02	0.07	0.02	0.15	-0.15	-0.15			
% stearic	-0.01	-0.05	0.01	0.10	-0.04	-0.01	0.07	-0.24	-0.01	0.07	-0.07	0.12	0.05	0.05	0.03	-0.12	-0.08	0.03	-0.14	0.22	0.16		
% oleic	-0.10	-0.32	0.02	-0.19	0.11	-0.21	0.04	-0.08	-0.03	-0.20	0.14	-0.30	0.10	-0.50	0.37	0.29	0.42	-0.19	0.20	-0.47	0.24	0.27	
% linoleic	0.10	0.28	0.00	0.15	-0.11	0.17	-0.06	0.12	0.01	0.15	-0.14	0.26	-0.10	0.45	-0.35	-0.26	-0.38	0.15	-0.13	0.43	-0.36	-0.39	-0.98

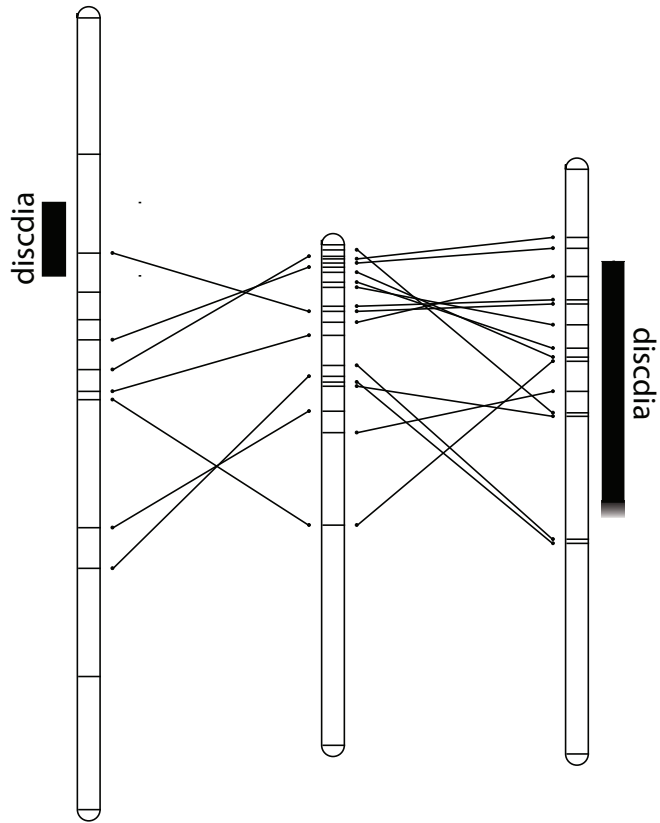
Root= rooting rate; Leaf sz = average leaf size; Leaf shp = leaf shape; Spines = leaf spininess; Flower = days to flowering; Disc dia = primary disc diameter; Cap hght = primary capitulum height; No. heads = number of heads; Color = flower color; Stem hght = stem height; No. ints = number of internodes; Low br. = lowest branch height; No. seed = number of selfed seed; Seed wght = seed weight; Seed len = seed length; Seed wid. = seed width; % via. = seed viability; Dorm = seed dormancy; % oil = % seed oil content; % palmitic = % palmitic acid; % stearic = % stearic acid; % oleic = % oleic acid; % linoleic = % linoleic acid

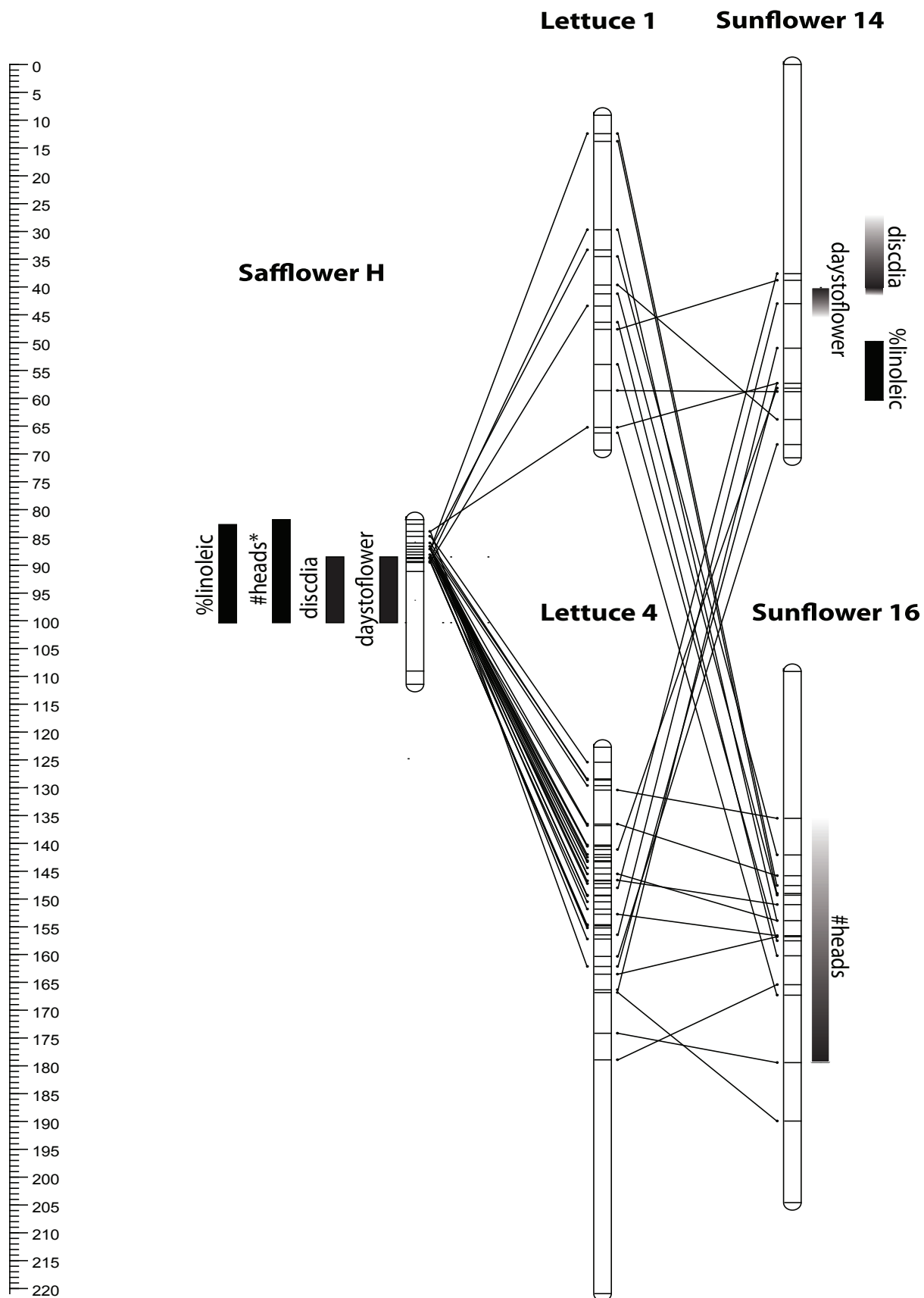
Table A2. Summary of significant interactions detected among all mapped markers.

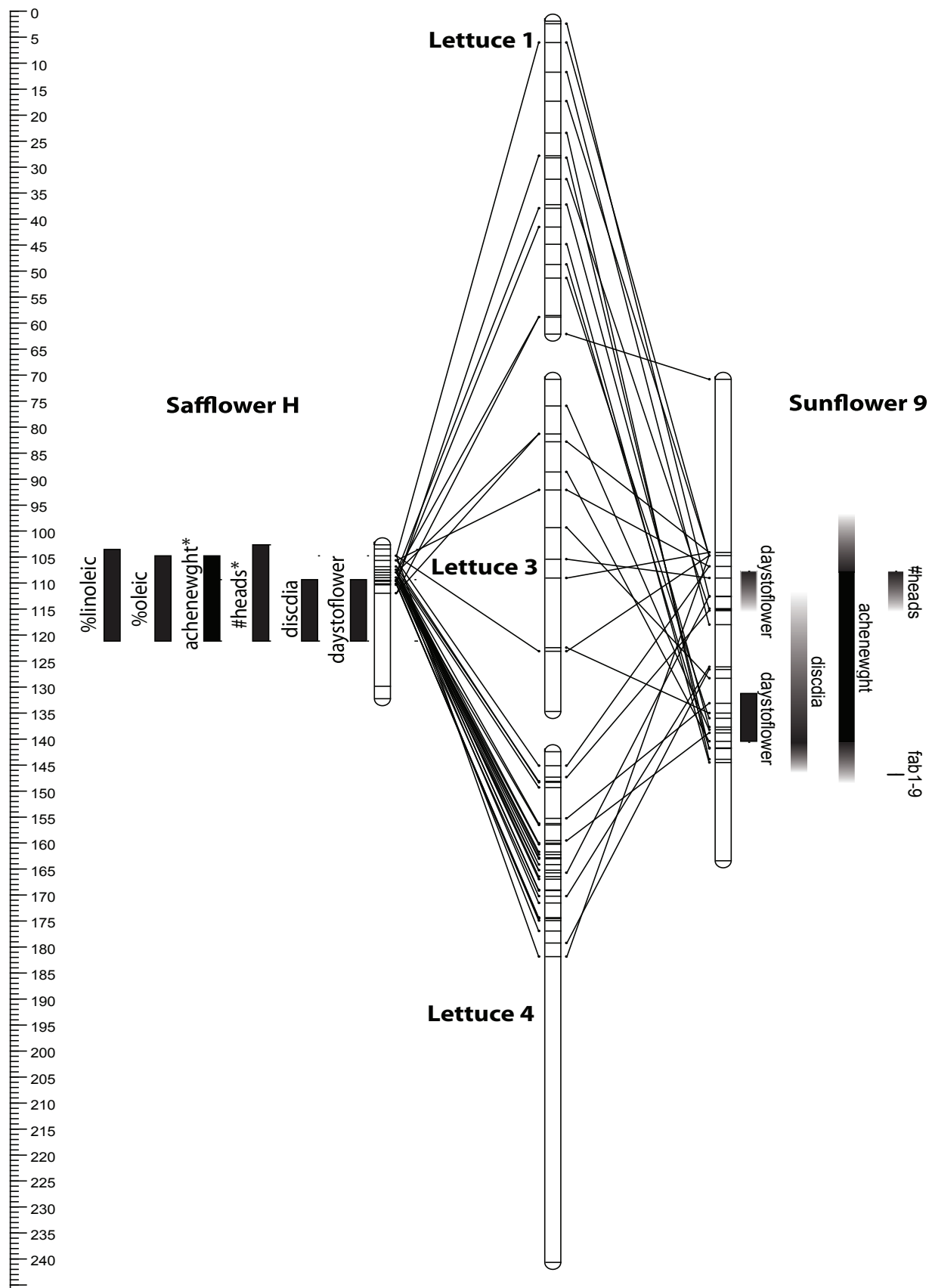
Trait	Type of interaction ^{ab}		
	A × A	A × D	D × D
Rooting rate		I × A	D × I
		J × C	
Average leaf roundness	H × K	A × I	
	H × L	L × L	
Spininess	H × L		
	I × J		
Days to flower	A × E		A × B
Primary capitulum height	E × J	A × K	G × G
	G × G	F × H	J × L
	K × L	G × G	L × K
		L × K	
Primary disc diameter	D × J	K × A	B × D
			D × J
Number of heads	A × E	E × G	C × I
Flower color	A × D	B × H	B × H
	D × E	I × H	E × I
	D × K		
	E × I		
	L × L		
Stem height			E × H
Number of internodes		E × E	G × L
Internode length	E × K	E × E	B × D
Number of selfed seed	B × E	D × D	B × L
	E × L	J × D	C × L
	H × L		
Achene weight	A × C	E × K	
Achene length	A × C	A × L	C × K
		D × E	
		F × E	
		K × C	
Achene width	A × C	A × I	
		A × L	
		D × J	
		E × K	
Seed viability	A × L	A × L	A × G

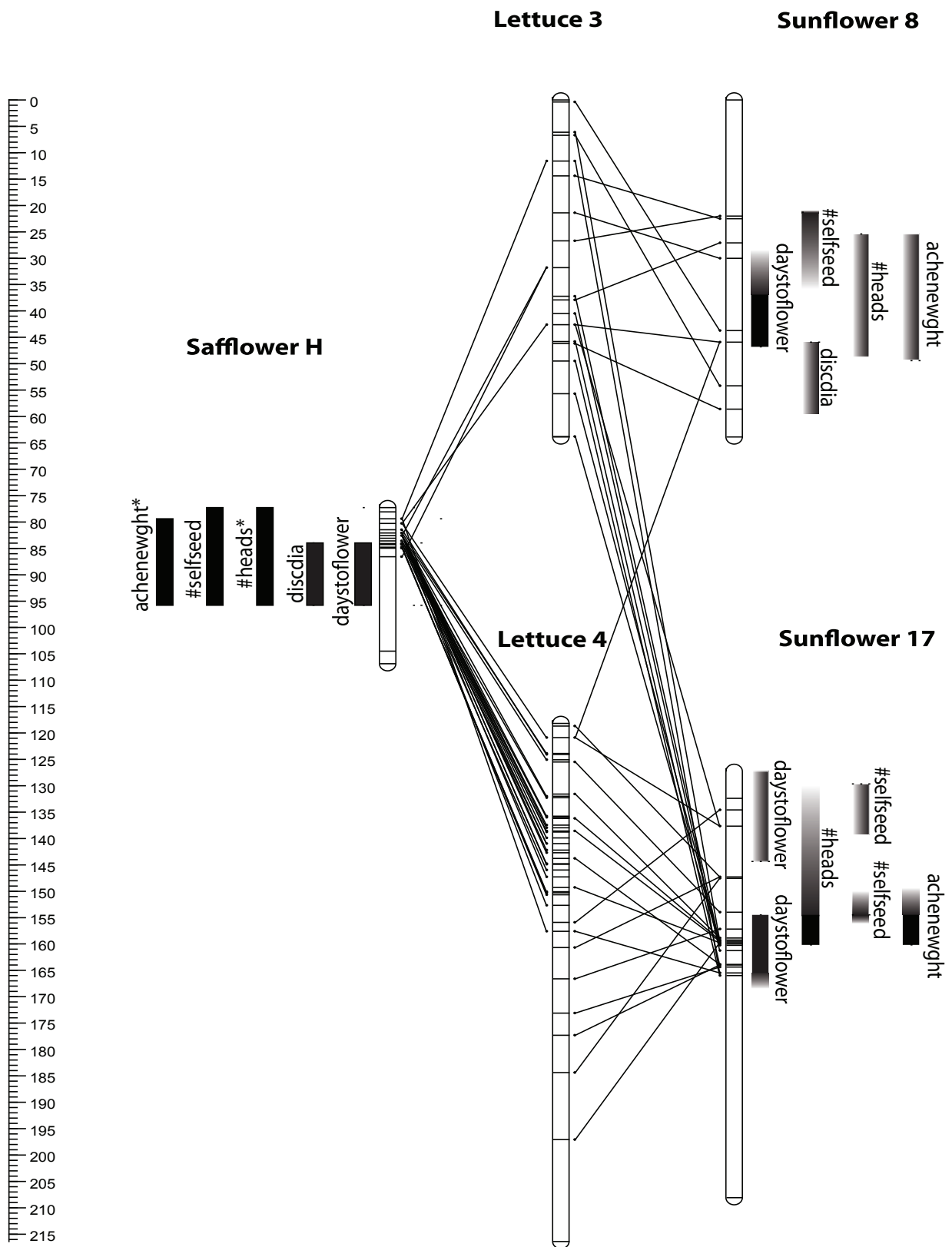
Figure A1. Additional colocalizing quantitative trait loci (QTL), following the format of Figure 2.3.

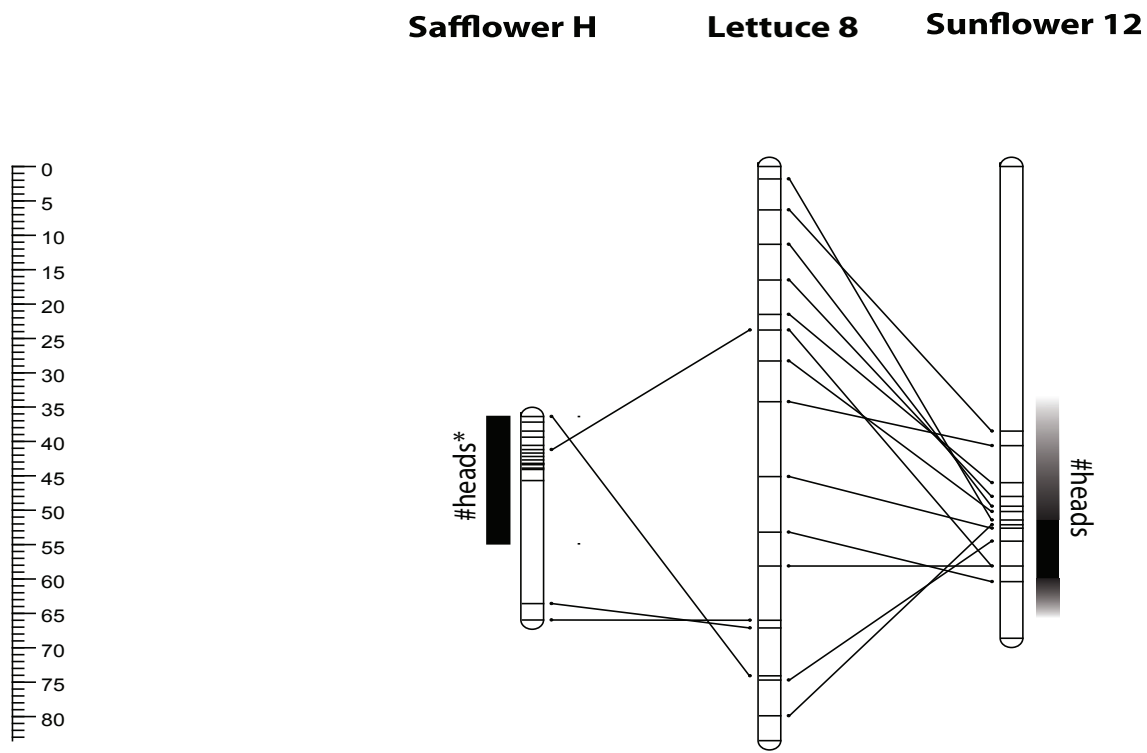
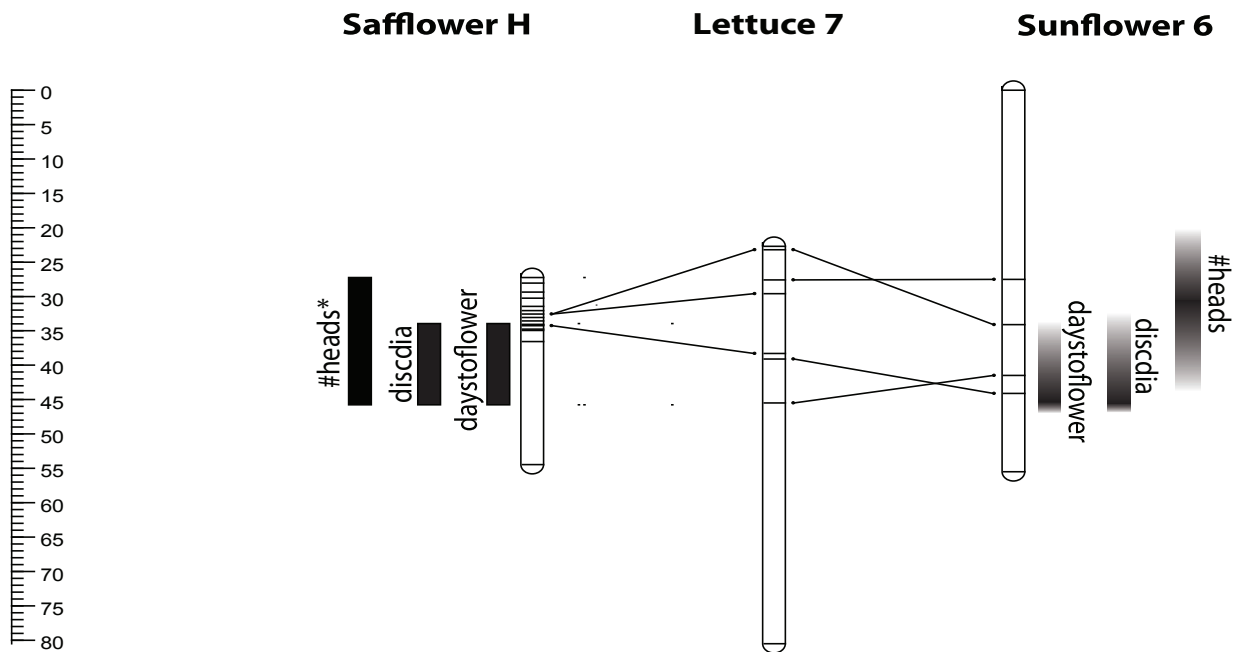
Safflower A Lettuce 9 Sunflower 1







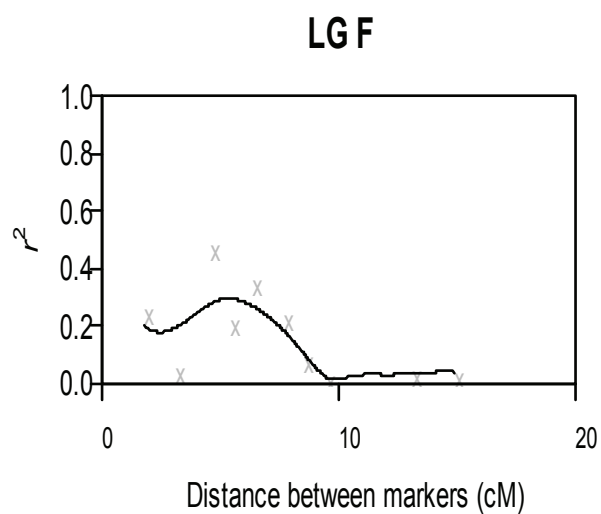
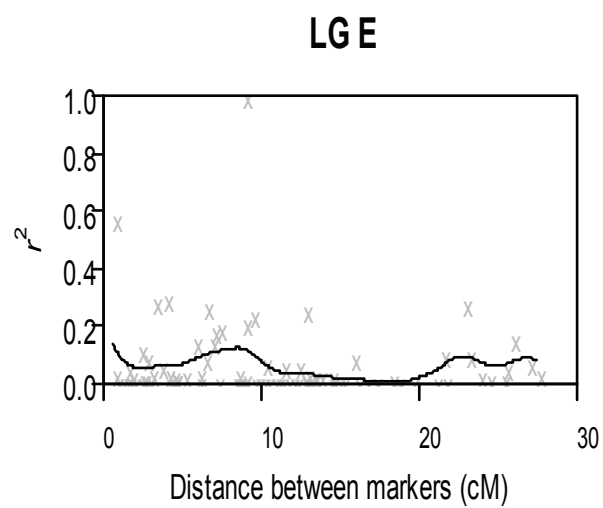
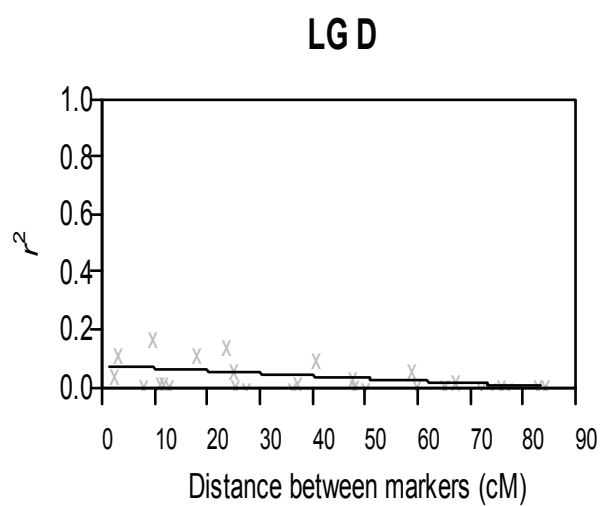
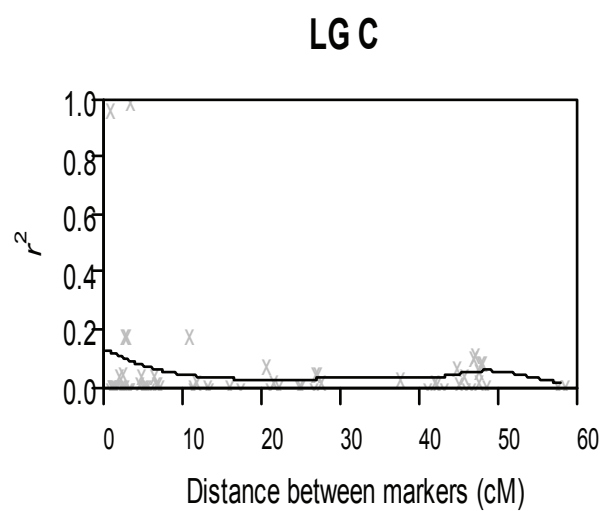
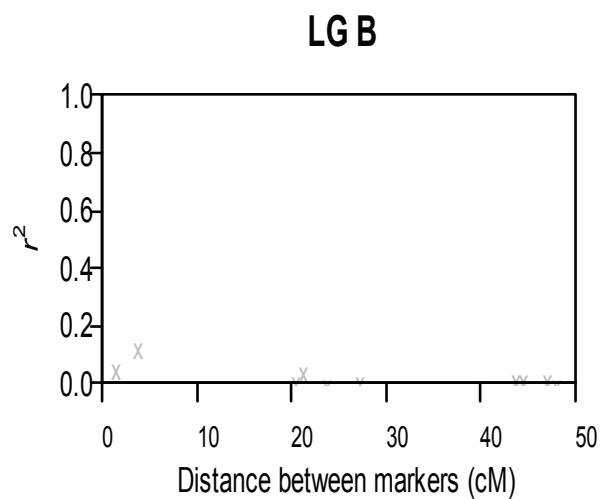
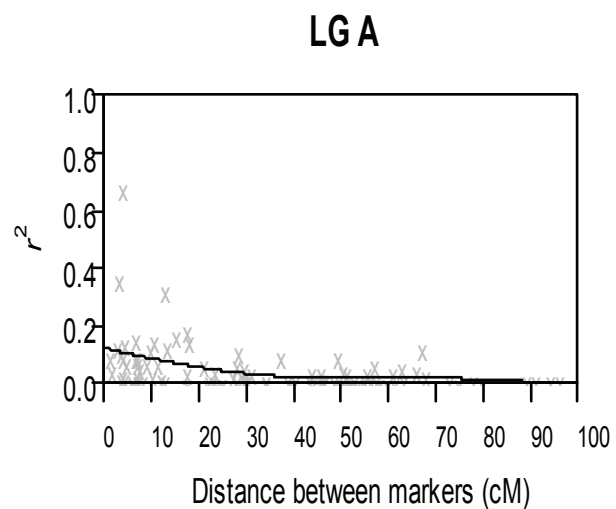


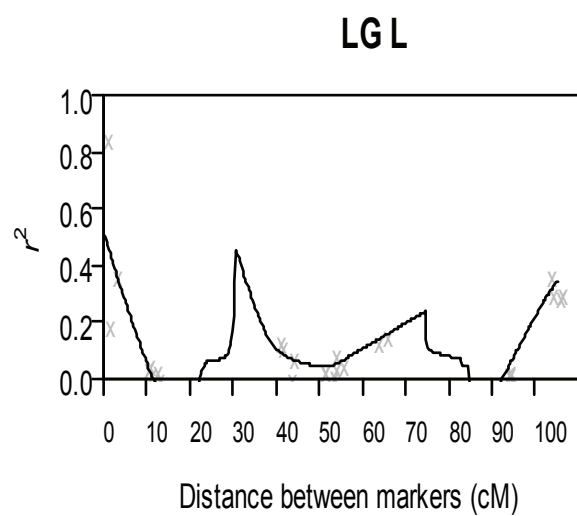
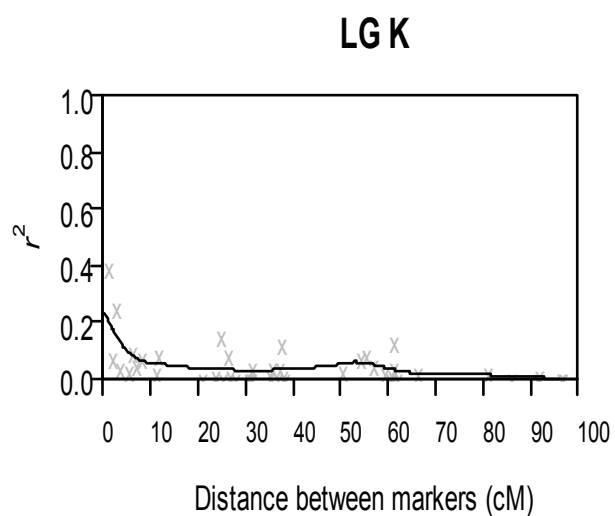
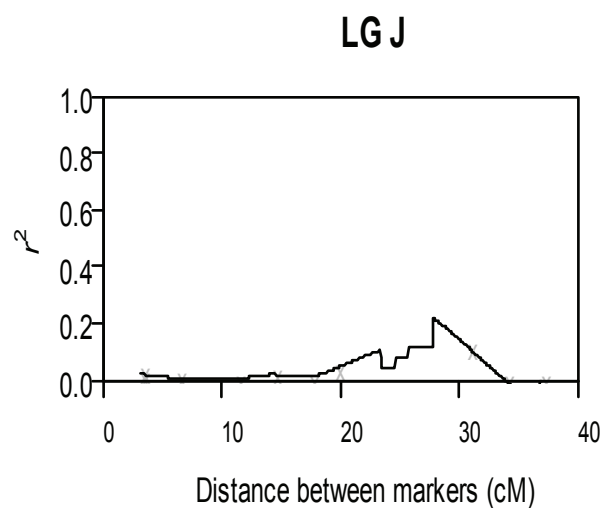
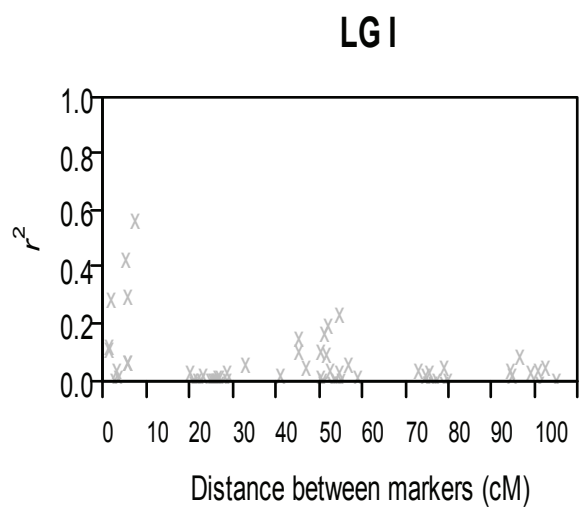
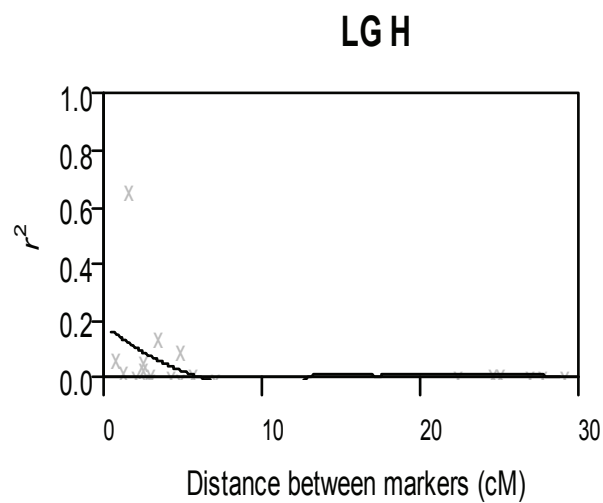
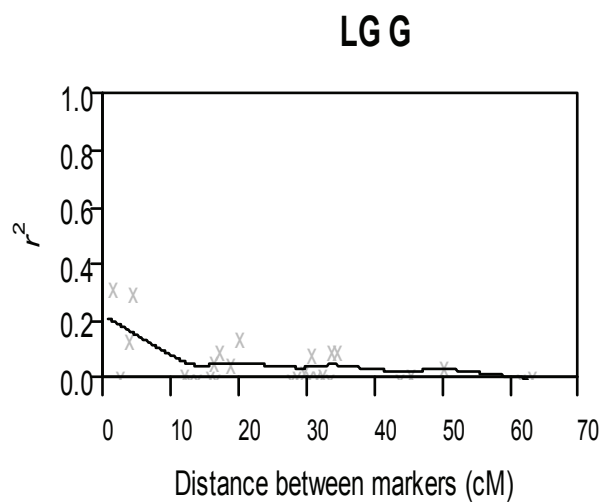


APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER III

Figure B1. Plots of the squared allele frequency correlations (r^2) as a function of genetic map distance (cM) calculated for all SNPs across all samples for each linkage group. The line summarizes the r^2 values as a function of map distance using the KernSmooth package within the R programming language (see text for details).





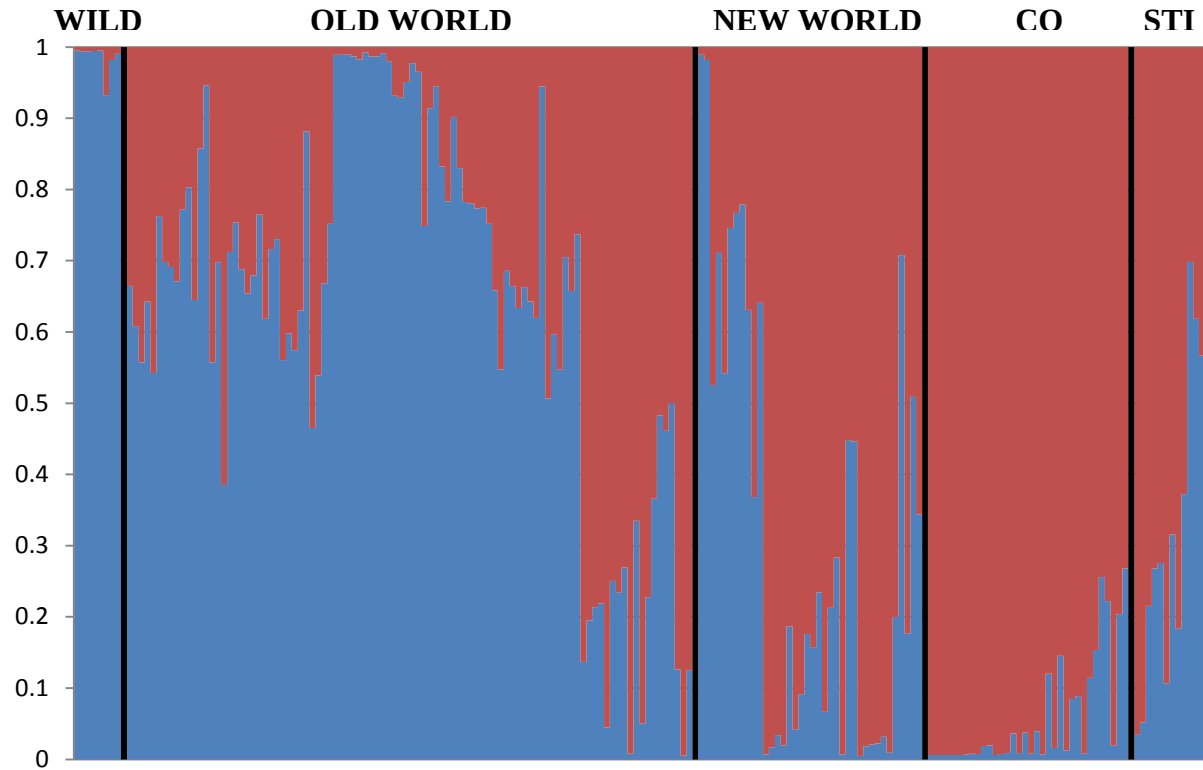
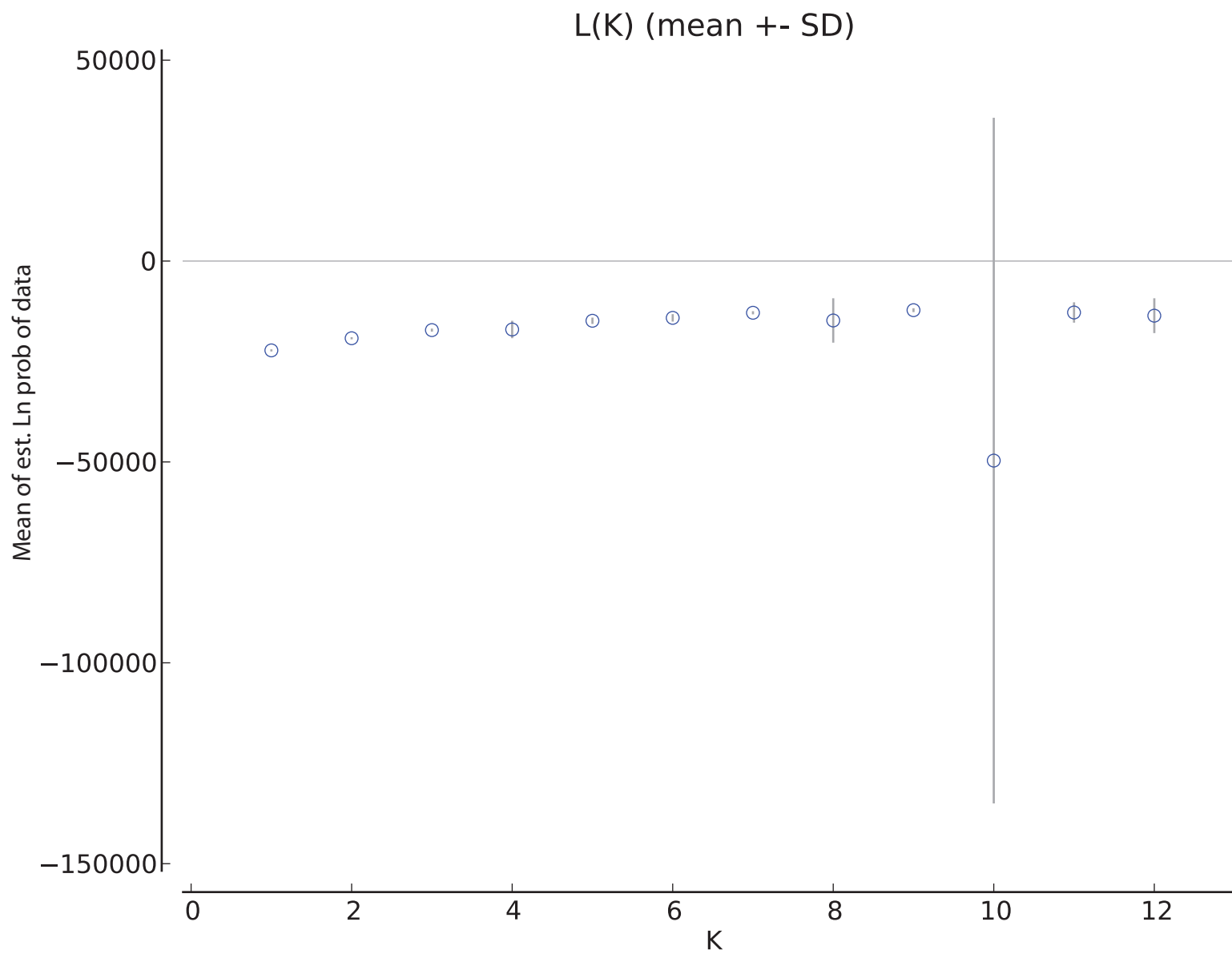
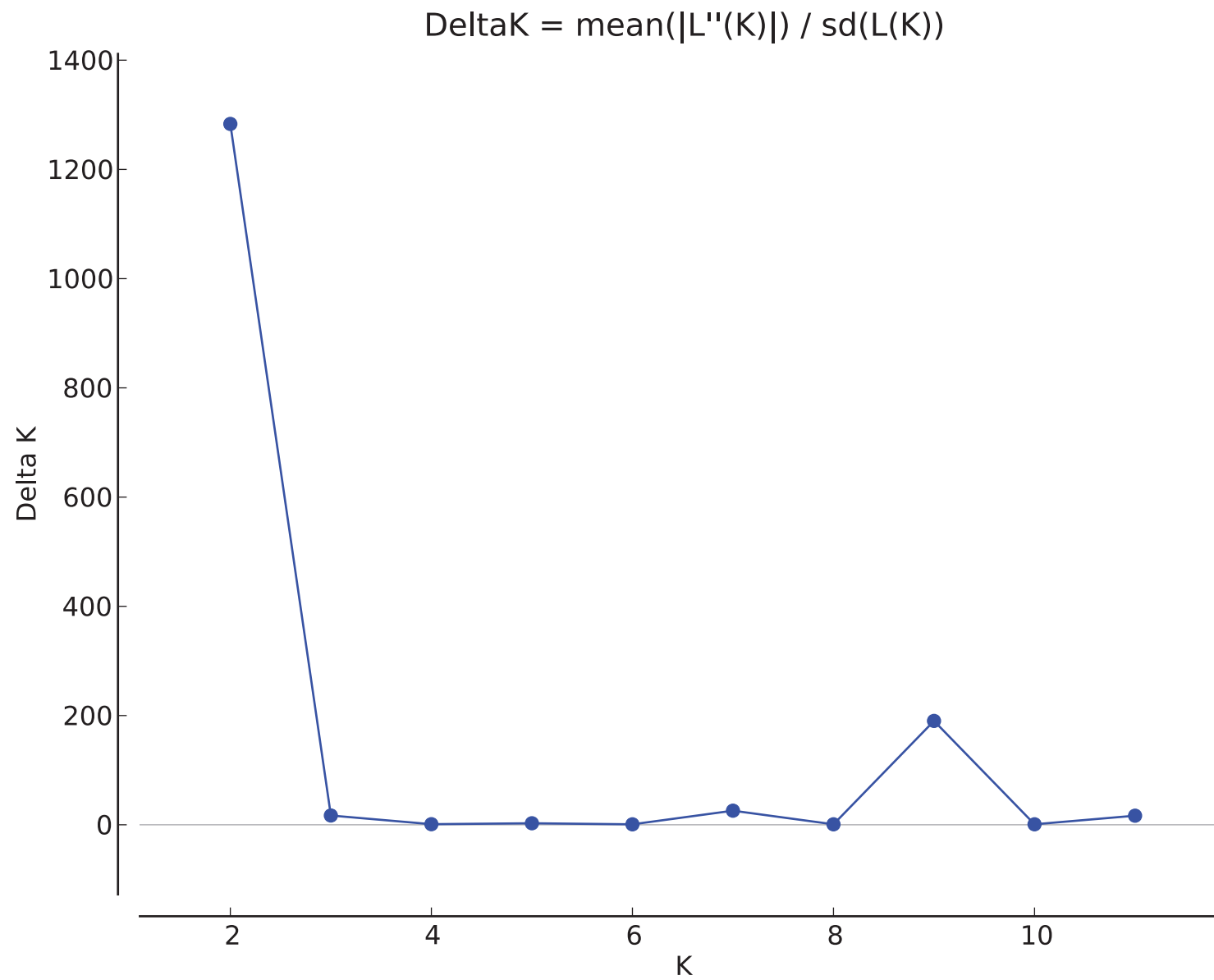
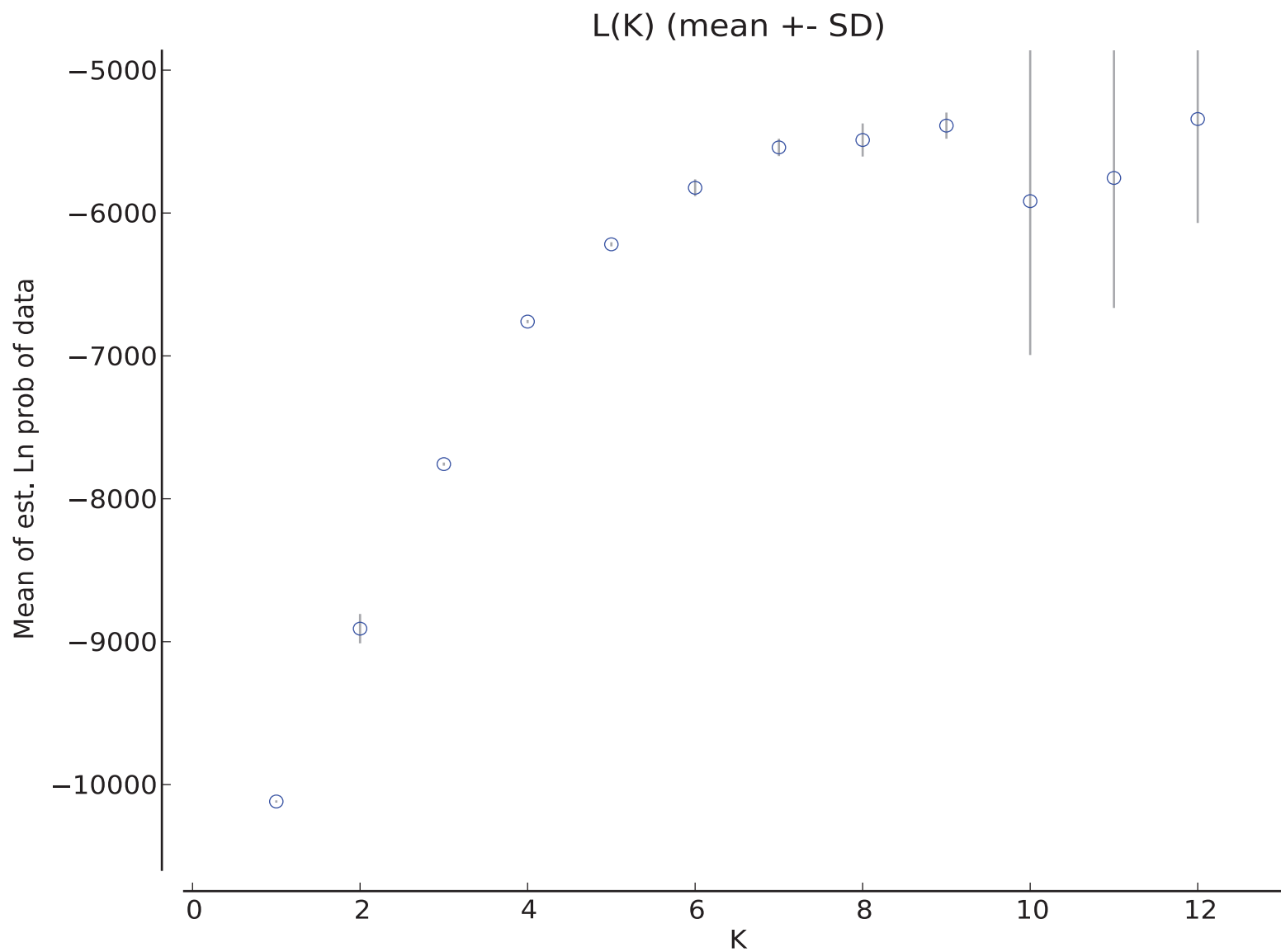


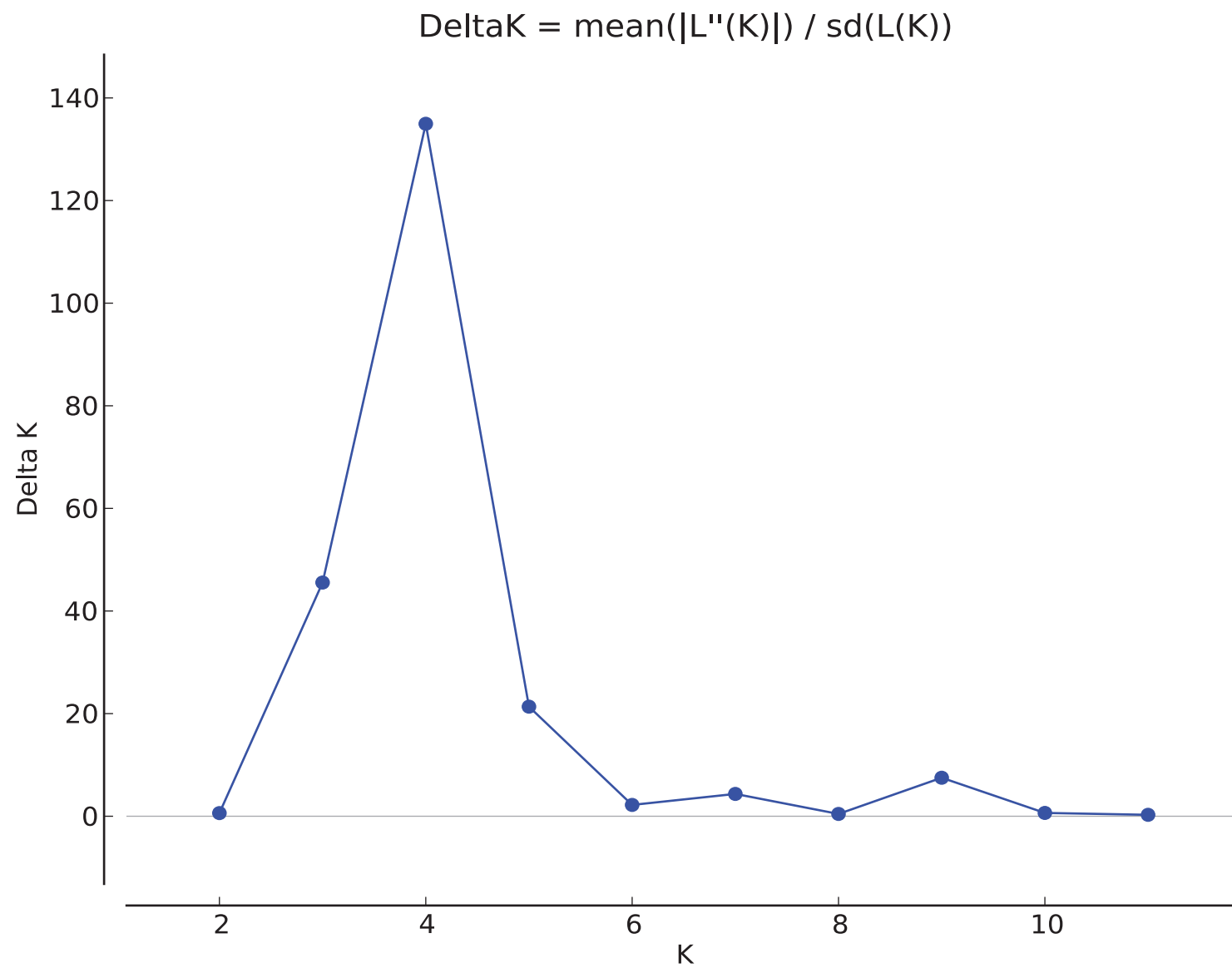
Figure B2. STRUCTURE plots of 190 *Carthamus* individuals. Black bars are used to separate pre-defined groupings of *Carthamus* (labels found at top of graphs). Each vertical bar represents a single individual, and the proportion of membership to each population is indicated on the y-axis. Here, $K = 2$ clusters; blue = cluster 1; red = cluster 2.

Figure B3. Log likelihood and *Delta K* values corresponding to STRUCTURE run of all 190 *Carthamus* individuals (a,b) and a separate analysis of 96 Old World individuals (c,d). The average of five runs is plotted in each figure.









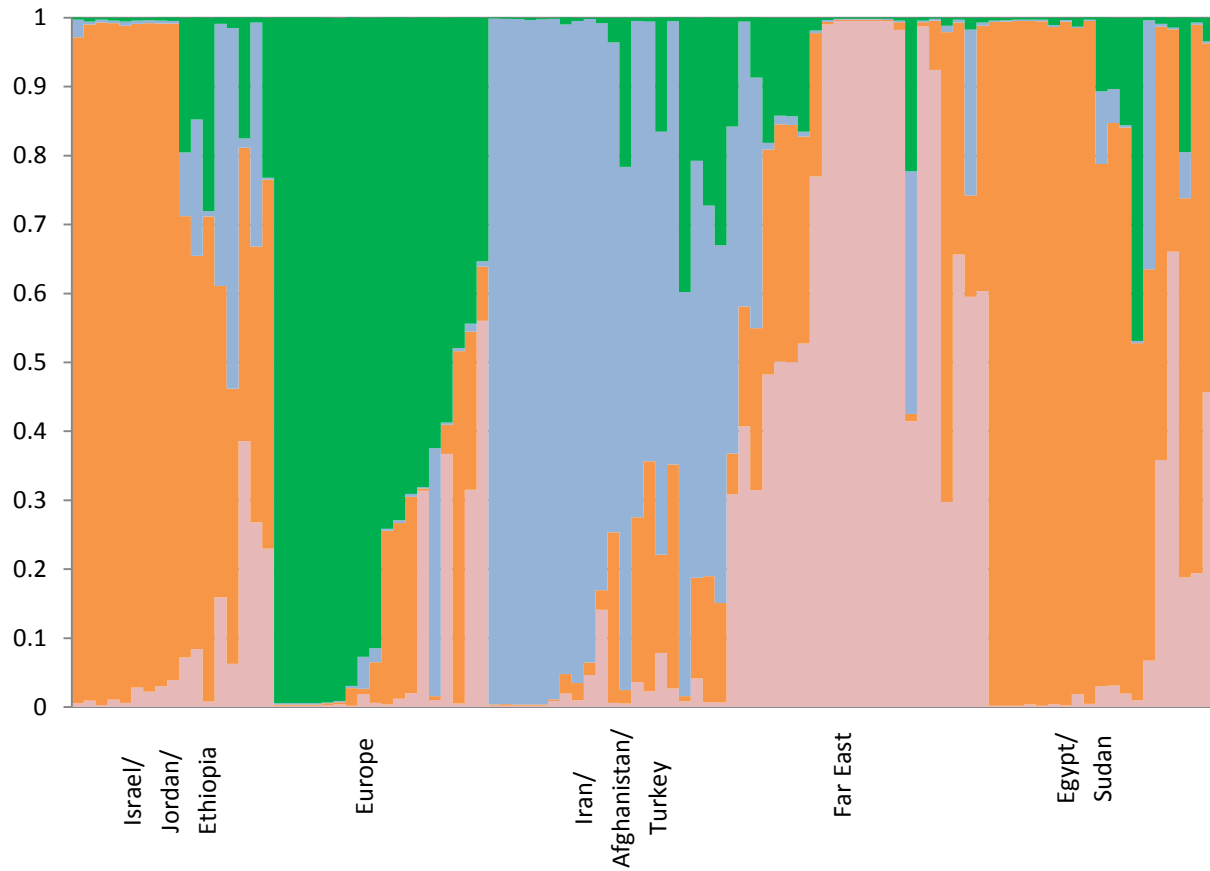


Figure B4. STRUCTURE plot of 96 Old World *C. tinctorius* individuals. Each vertical bar represents a single individual, and the proportion of membership to each population is indicated on the y-axis. Colors correspond to Figure 3.1, in which orange = cluster 2, green = cluster 3, light blue = cluster 4, and pink = cluster 5.

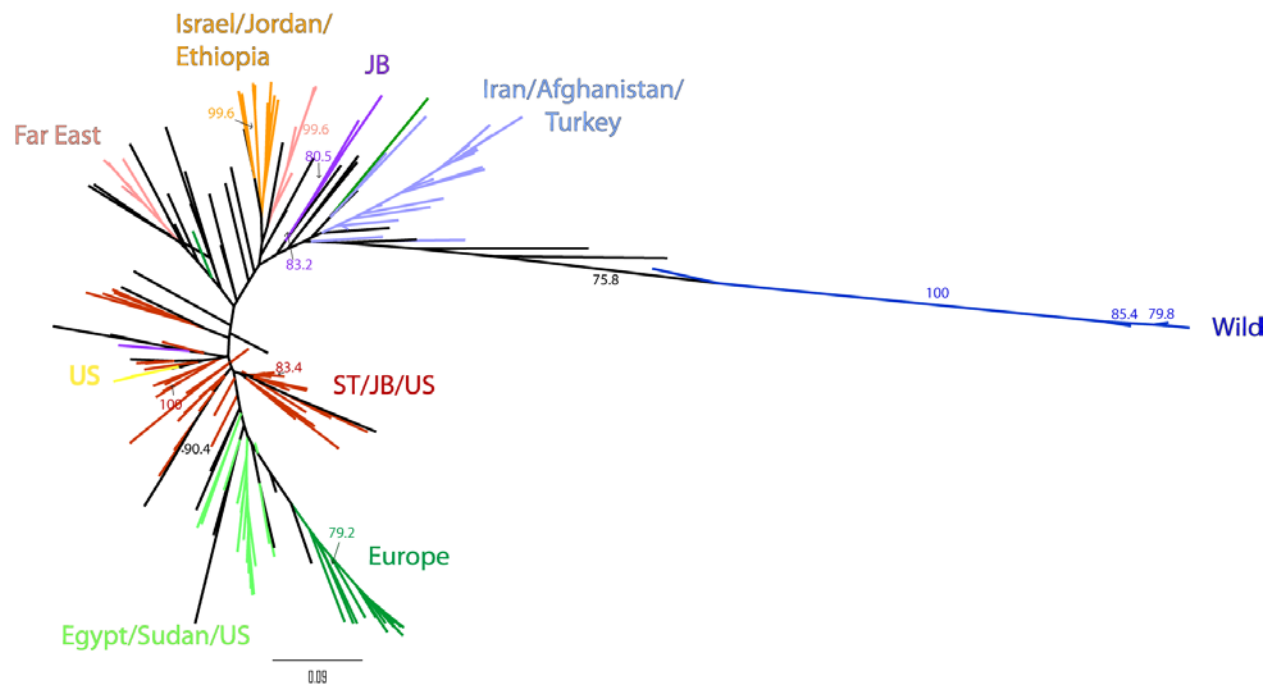


Figure B5. Neighbor-joining trees of 190 *Carthamus* individuals, color-coded to correspond to the 9 clusters appearing in Figure 1. Branches with bootstrap support values greater than 75% are labeled. Individuals with less than 50% membership to any given cluster were not assigned to any cluster and are represented by a black line.

Table B1. List of Cultivated safflower accessions obtained from the USDA for use in this study.

Ac. No.	Country of origin	Notes ^a	Cluster ^b
PI 181866	Syria	Cultivated, Core	2
PI 193473	Ethiopia	Cultivated, Core	2
PI 198844	France	Cultivated	3
PI 198990	Israel		2
PI 199889	India	Cultivated, Core	5
PI 208677	Algeria		3
PI 209287	Romania	Cultivated, Core	3
PI 209297	Kenya	Cultivated, Core	5/3
PI 209300	Kenya	Cultivated, Core	5
PI 220647	Afghanistan	Cultivated, Core	4
PI 226993	Israel	Core	2
PI 235658	Australia	Cultivated	2
PI 237547	Sudan	Cultivated, Core	6
PI 237548	Sudan	Cultivated, Core	4/5
PI 239042	Morocco		3
PI 239226	Spain	Cultivated, Core	3
PI 242419	Australia		6
PI 243070	Jordan	Cultivated	2
PI 248625	Pakistan	Cultivated, Core	5
PI 250081	Egypt	Cultivated	6
PI 250202	Pakistan	Cultivated, Core	5
PI 250533	Egypt	Cultivated	6
PI 250537	Egypt	Cultivated, Core	6
PI 250611	Egypt	Cultivated	6
PI 250833	Iran	Cultivated, Core	4
PI 251262	Jordan	Wild (Knowles)	2
PI 251290	Israel	Cultivated	2
PI 251291	Jordan	Cultivated	4
PI 251398	Iran	Wild (Knowles), Core	4
PI 251984	Turkey	Cultivated, Core	4
PI 253386	Israel	Wild (Knowles)	4
PI 253523	Italy	Cultivated, Core	4
PI 253531	Bulgaria	Cultivated, Core	3/6
PI 253538	Armenia	Cultivated	3
PI 253540	Hungary	Cultivated, Core	4
PI 253541	Hungary	Cultivated, Core	3
PI 253543	Poland	Cultivated, Core	3
PI 253548	Denmark	Cultivated	3
PI 253559	Portugal	Cultivated	3
PI 253560	Morocco	Cultivated	3
PI 253759	Iraq	Cultivated, Core	4

PI 253908	Afghanistan	Wild (Knowles), Core	4
PI 254976	Greece	Cultivated, Core	2
PI 257582	Ethiopia	Core	2
PI 258420	Portugal		3
PI 259992	Pakistan	Cultivated, Core	3
PI 260637	India	Cultivated, Core	5
PI 262420	Australia	Cultivated	6
PI 262423	Australia	Cultivated	4
PI 262430	Syria	Core	2
PI 262435	Uzbekistan		2
PI 262444	Kazakstan	Core	4
PI 268374	Afghanistan	Wild (Harlan), Core	4
PI 271070	Sudan	Cultivar, Core	5
PI 273876	Eritrea	Cultivated, Core	2
PI 279051	India	Cultivated, Core	5
PI 279342	Japan	Cultivated	5
PI 283764	India	Core	5
PI 286199	Kuwait	Core	4
PI 291600	Argentina	Cultivated, Core	6
PI 292003	Israel		6
PI 301048	Turkey	Cultivated, Core	4/6
PI 304408	Pakistan	Wild (Knowles), Core	8
PI 304503	Turkey	Wild (Knowles), Core	2
PI 304595	Afghanistan	Cultivated, Core	4
PI 305529	Sudan	Core	6
PI 305531	Sudan	Core	5
PI 305534	Sudan	Core	6
PI 305540	Kazakhstan	Core	5
PI 306599	Egypt	Wild (Knowles),	6
PI 306686	Israel		6
PI 306974	India	Wild (Knowles), Core	5
PI 307055	India	Wild (Knowles), Core	5
PI 312275	Hungary	Cultivated, Core	3
PI 314650	Kazakhstan	Wild (Jones), Core	6
PI 343930	Ethiopia	Cultivated	2
PI 348915	Canada	Cultivated, Core	4
PI 369843	Uzbekistan		2
PI 369847	Tajikistan	Core	4
PI 369853	Uzbekistan	Cultivar	4
PI 380800	Iran	Cultivated, Core	4
PI 386174	Syria	Cultivar, Core	2/4
PI 393499	Libya		4
PI 401479	Bangladesh	Wild (Hobbs), Core	5

PI 405984	Iran	Cultivated, Core	4
PI 406015	Iran	Cultivated, Core	4
PI 407624	Turkey	Cultivated, Core	4
PI 426523	Pakistan	Cultivated, Core	3
PI 451956	India	Cultivated, Core	6/2/5
PI 506427	China	Cultivar, Core	7
PI 514630	China	Cultivar, Core	5
PI 525457	US	"Girard," Cultivar, Core, Historic	7
PI 532619	Cyprus	Cultivated	2
PI 537608	US	Breeding, Core	6
PI 537626	US	Breeding, Core	4/3/7/8
PI 537636	US	Breeding, Core	6
PI 537652	Mexico	Breeding	8/5/6
PI 537659	US	Breeding, Core	5
PI 537682	US	Breeding, Core	8/5
PI 537692	US	"Gila," Cultivar, Core, Historic	8
PI 537695	US	"Ole," Cultivar, Historic	7
PI 538779	US	"Centennial," Cultivar, Pureline	7
PI 543995	China	Cultivated, Core	5
PI 544006	China	Cultivated, Core	5
PI 544033	China	Cultivated, Core	2/4
PI 544041	China	Cultivated, Core	8
PI 544052	China	Cultivated, Core	6
PI 560172	US	Breeding, Core	7/8
PI 560175	US	Breeding, Core	7
PI 560177	US	"Oleic Leed," Breeding, historic	8
PI 560192	US	Breeding, Core	9/2
PI 560200	US	Breeding, Core	6
PI 560205	US	"Mexico Dwarf," Breeding	5
PI 561703	US	"San Jose," Cultivar, Historic	7/6
PI 562638	India	Core	5
PI 568864	China	Cultivated	5
PI 572415	US	Cultivar	5/7
PI 572420	US	Cultivar	8
PI 572428	US	"Nebraska 10," Cultivar, Historic	6
PI 572435	US	"Dart," Cultivar, Historic	7
PI 572436	US	"Leed," Cultivar, Historic	8
PI 576991	Germany		3
PI 576992	North Korea		5
PI 613394	US	Cultivated, Core	4
PI 592391	Canada	"AC Sunset," Cultivar, Pureline, Historic	3
PI 601166	US	"Oker," Cultivar, Historic	7
PI 603208	US	"Lesaf," Breeding, Historic	8

PI 613465	Spain	Cultivated, Core	3
PI 613498	US	Cultivated	7
PI 613514	Australia	Cultivated	8/7
W6 39446	US	Cultivated, Winter hardy	4
PI 651878	US	Breeding pureline, Winter hardy	5
PI 651879	US	Breeding pureline, Winter hardy	5
PI 651880	US	Breeding pureline, Winter hardy	5

^aNotes regarding each accession are provided by the USDA and include accession improvement status, as follows:

Wild: not collected in a field or cultivated area

Cultivar: named cultivar developed by scientific means

Cultivated: collected from a field planting

Breeding: lines developed by scientific means and used in breeding programs

Landrace: locally adapted variety

^bNumbers correspond to the primary cluster (>50% membership) assigned to each sample by STRUCTURE. Individuals with multiple clusters listed had no predominant cluster of ancestry, and the multiple dominating clusters are listed.

Table B2. Genetic diversity statistics among subsets of CO breeding lines.

Population	<i>N</i>	%P	<i>A_g</i> (± S.E.)	<i>PAL</i>	<i>H_o</i>	<i>H_e</i>
Varieties	11	36.84	1.34 ^a (0.0015)	0.06 ^a	0.003 ^a	0.121 ^a
Elite breeding lines	17	42.11	1.35 ^a (0.0017)	0.07 ^a	0.011 ^a	0.138 ^a
Germplasm conversion	6	35.34	1.35 ^a (0.0014)	0.07 ^a	0.036 ^b	0.131 ^a

N number of plants sampled

%P percent polymorphic loci

A_g allelic richness (based on the rarefaction method) averaged across all loci

PAL private allelic richness (based on the rarefaction method) averaged across all loci

H_o observed heterozygosity averaged across all loci

H_e expected heterozygosity averaged across all loci

Superscript letters indicate differences in significance levels ($P < 0.001$)

Table B3. Pairwise F_{ST} among wild, cultivated (Old World and New World), and commercial (CO and STI) safflowers.

	Wild	Old World	New World	CO
Wild				
Old World	0.512			
New World	0.551	0.070		
CO	0.712	0.257	0.133	
STI	0.556	0.170	0.094	0.191

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER IV

A)

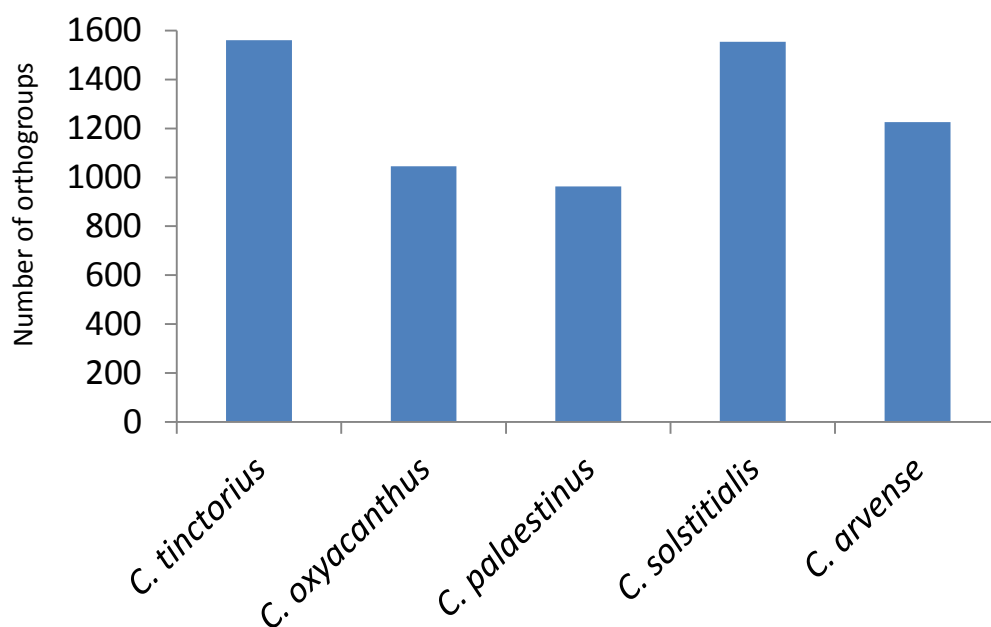
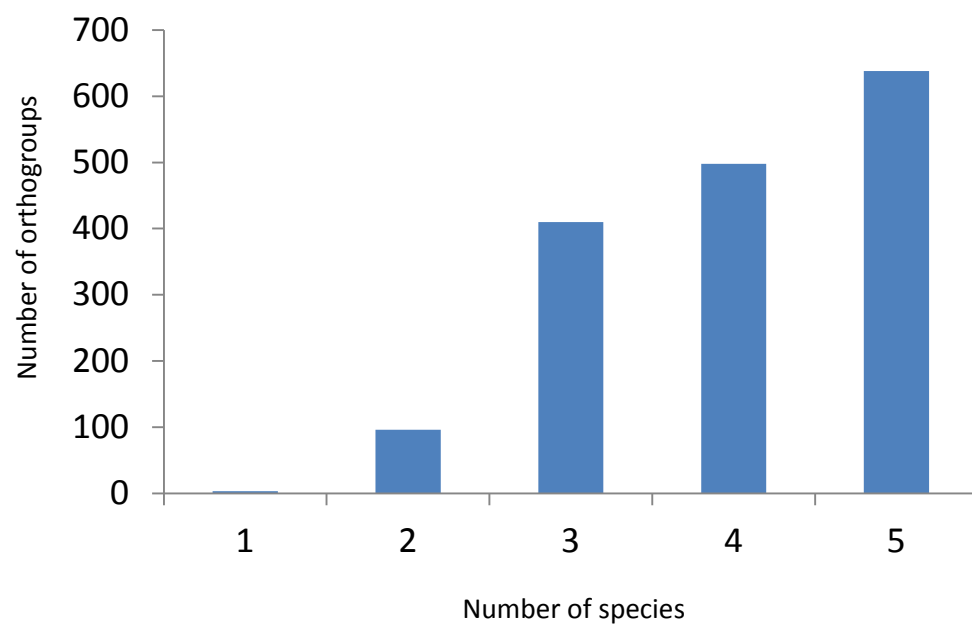


Figure C1. Number of orthogroups containing one through five species (A) and number of orthogroups of which each species is a member (B).

Table C1. Species included within the 22 genome gene family circumscription.

Family	Species
Funariaceae	<i>Physcomitrella patens</i>
Selaginellaceae	<i>Selaginella moellendorffii</i>
Amborellaceae	<i>Amborella trichopoda</i>
Muscaceae	<i>Musa acuminata</i>
Arecaceae	<i>Phoenix dactylifera</i>
Poaceae	<i>Sorghum bicolor</i>
Poaceae	<i>Brachypodium distachyon</i>
Poaceae	<i>Oryza sativa</i>
Ranunculaceae	<i>Aquilegia coerulea</i>
Nelumbonaceae	<i>Nelumbo nucifera</i>
Phrymaceae	<i>Mimulus guttatus</i>
Solanaceae	<i>Solanum tuberosum</i>
Solanaceae	<i>Solanum lycopersicum</i>
Vitaceae	<i>Vitis vinifera</i>
Malvaceae	<i>Theobroma cacao</i>
Salicaceae	<i>Populus trichocarpa</i>
Fabaceae	<i>Medicago truncatula</i>
Fabaceae	<i>Glycine max</i>
Rosaceae	<i>Fragraria vesca</i>
Caricaceae	<i>Carica papaya</i>
Brassicaceae	<i>Arabidopsis thaliana</i>
Brassicaceae	<i>Thellungiella parvula</i>

Table C2. Tree size and species included for each of the positively selected orthogroups.

Set ID	No. branches	Tree length	Average dS	Ct	Species included*			
					Cp	Co	Ce	Ci
2648	8	0.0973	0.06	x	x	x	x	x
2635	5	0.722	0.09	x	x		x	x
7215	5	0.589	0.073	x	x	x	x	x
8816	5	0.589	0.074	x	x		x	x
3402	6	0.486	0.049	x	x	x	x	x
1008	5	0.388	0.049	x		x	x	x
5547	6	0.573	0.057	x	x	x	x	x
5925	5	0.357	0.045	x	x	x	x	x
6970	5	0.359	0.045	x	x	x	x	x
1198	5	0.377	0.047	x	x	x	x	x
2039	5	0.973	0.122	x			x	x
7143	5	0.378	0.041	x	x	x	x	x
3407	5	0.402	0.05	x		x	x	x
2076	5	0.244	0.031	x	x	x	x	x
8235	5	0.238	0.03	x	x	x	x	x
12423	5	0.562	0.07	x	x	x	x	x

7594	6	0.62	0.062	x	x	x	x	x
3859	5	0.483	0.06	x	x	x	x	x
3943	6	0.441	0.044	x	x	x	x	x
8059	5	0.297	0.037	x	x	x	x	x
1780	8	0.438	0.055	x		x	x	x
11119	5	0.587	0.073	x	x		x	x
5754	6	0.282	0.028	x	x	x	x	x
1325	5	0.794	0.099	x		x	x	x
967	13	0.281	0.035	x	x	x	x	x
4768	5	0.385	0.048	x	x	x	x	x
591	5	0.997	0.125	x			x	x
948	5	0.565	0.071	x	x	x	x	
6276	7	0.515	0.043	x	x	x	x	x
3843	5	0.323	0.04	x	x	x	x	x
7305	6	0.487	0.049	x	x	x	x	x
4930	5	0.51	0.064	x		x	x	x
12570	5	0.172	0.021	x	x	x	x	
4311	5	0.606	0.076		x	x		x
8948	5	0.565	0.071	x	x	x	x	x
2238	7	0.868	0.072	x	x	x	x	x
8058	5	0.362	0.045	x	x	x	x	x
5390	5	0.41	0.051	x	x	x	x	x
2611	5	0.593	0.074	x	x	x	x	x
8990	5	0.412	0.051	x	x	x	x	x
2262	5	0.634	0.079	x			x	x

*Ct = *C. tinctorius*; Cp = *C. palaestinus*; Co = *C. oxyacanthus*; Ce = *C. solstitialis*; Ci = *C. arvensis*

Table C3. LRTs for variation in ω .

Set ID	Model	ln(Likelihood)	ω_1	ω_2	ω_3	Test	LRT statistic	P-value
2648	M0	-1574.75	0.295			M3:M0	40.29	3.76e-08
	M3	-1554.60	0.017	0.017	0.937			
	M1a	-1554.62	0.024	1.000		M2a:M1a	0.00	1.000
	M2a	-1554.62	0.024	1.000	1.000			
	MX+S1 (constraint)	-1544.32	0.001	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	80.28	3.25e-19
	Mx+S1 (selection)	-1504.18	0.001	0.001	19.999			
2635	M0	-1726.88	0.275			M3:M0	86.82	6.25e-18
	M3	-1683.47	0.000	0.836	6.731			
	M1a	-1692.31	0.000	1.000		M2a:M1a	17.64	1.48e-4
	M2a	-1683.49	0.012	1.000	7.150			
	MX+S1 (constraint)	-1691.63	0.001	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	71.43	2.87e-17
	Mx+S1 (selection)	-1655.91	0.020	0.020	19.999			
7215	M0	-1931.71	0.336			M3:M0	73.01	5.25e-15
	M3	-1895.20	0.000	0.000	2.187			
	M1a	-1901.01	0.000	1.000		M2a:M1a	11.62	0.003
	M2a	-1895.20	0.000	1.000	2.187			
	MX+S1 (constraint)	-1900.58	0.001	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	62.12	3.23e-15
	Mx+S1 (selection)	-1869.52	0.001	0.001	19.999			
8816	M0	-1875.40	0.333			M3:M0	57.72	8.74e-12
	M3	-1846.54	0.184	3.298	40.103			
	M1a	-1857.98	0.036	1.000		M2a:M1a	21.70	1.94e-5
	M2a	-1847.13	0.082	1.000	13.753			
	MX+S1	-1857.66	0.001	0.999	1.000	MX+S1 (selection):MX+S1	47.38	5.86e-12

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-1833.97	0.169	0.169	19.999			
3402	M0	-1162.90	0.130			M3:M0	48.97	5.92e-10
	M3	-1138.41	0.000	0.003	1.618			
	M1a	-1139.48	0.000	1.000		M2a:M1a	2.13	0.345
	M2a	-1138.41	0.001	1.000	1.618			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1136.79	0.001	0.999	1.000	(constraint)	46.12	1.11e-11
	Mx+S1							
	(selection)	-1113.73	0.001	0.001	19.999			
1008	M0	-2307.08	0.154			M3:M0	41.73	1.89e-8
	M3	-2286.21	0.029	0.029	1.756			
	M1a	-2286.86	0.000	1.000		M2a:M1a	1.30	0.522
	M2a	-2286.21	0.029	1.000	1.756			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-2284.65	0.001	0.999	1.000	(constraint)	36.05	1.92e-9
	Mx+S1							
	(selection)	-2266.62	0.029	0.032	19.999			
5547	M0	-851.71	0.477			M3:M0	26.67	2.32e-5
	M3	-838.38	0.000	1.036	7.403			
	M1a	-841.76	0.000	1.000		M2a:M1a	6.75	0.034
	M2a	-838.38	0.000	1.000	7.213			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-841.54	0.001	0.999	1.000	(constraint)	28.19	1.10e-7
	Mx+S1							
	(selection)	-827.45	0.001	0.150	19.999			
5925	M0	-1953.86	0.206			M3:M0	34.46	5.99e-7
	M3	-1936.63	0.000	1.058	6.233			
	M1a	-1938.42	0.000	1.000		M2a:M1a	3.57	0.167
	M2a	-1936.63	0.000	1.000	5.853			
	MX+S1	-1938.06	0.001	0.999	1.000	MX+S1 (selection):MX+S1	27.43	1.63e-7

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-1924.35	0.001	0.022	19.999			
6970	M0	-1703.26	0.132			M3:M0	31.32	2.64e-6
	M3	-1687.60	0.000	0.035	2.432			
	M1a	-1688.68	0.000	1.000		M2a:M1a	7.01	0.030
	M2a	-1685.18	0.000	1.000	31.899			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1687.32	0.001	0.999	1.000	(constraint)	27.03	2.01e-7
	Mx+S1							
	(selection)	-1673.80	0.001	0.001	19.999			
1198	M0	-1392.51	0.157			M3:M0	21.07	3.07e-4
	M3	-1381.97	0.000	0.000	1.377			
	M1a	-1382.31	0.000	1.000		M2a:M1a	0.66	0.717
	M2a	-1381.97	0.000	1.000	1.377			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1380.37	0.001	0.999	1.000	(constraint)	23.29	1.39e-6
	Mx+S1							
	(selection)	-1368.72	0.001	0.001	19.999			
2039	M0	-2627.69	0.096			M3:M0	60.58	2.19e-12
	M3	-2597.40	0.000	0.327	3.267			
	M1a	-2600.10	0.034	1.000		M2a:M1a	0.00	N/A
	M2a	-2600.10	0.034	1.000	57.578			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-2596.66	0.001	0.001	1.000	(constraint)	22.66	1.93e-6
	Mx+S1							
	(selection)	-2585.32	0.001	0.349	19.999			
7143	M0	-1625.60	0.277			M3:M0	14.74	0.005
	M3	-1618.23	0.175	0.175	4.838			
	M1a	-1619.68	0.000	1.000		M2a:M1a	2.89	0.236
	M2a	-1618.23	0.175	1.000	4.838			
	MX+S1	-1619.16	0.001	0.999	1.000	MX+S1 (selection):MX+S1	21.82	2.99e-6

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-1608.25	0.001	0.001	19.999			
3407	M0	-1753.74	0.267			M3:M0	16.45	0.002
	M3	-1745.51	0.033	0.033	1.300			
	M1a	-1745.60	0.000	1.000		M2a:M1a	0.17	0.917
	M2a	-1745.51	0.033	1.000	1.300			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1744.14	0.001	0.999	1.000	(constraint)	20.81	5.07e-6
	Mx+S1							
	(selection)	-1733.74	0.012	0.054	19.999			
2076	M0	-582.26	1.180			M3:M0	26.29	2.77e-5
	M3	-569.12	0.019	0.019	18.991			
	M1a	-578.92	0.000	1.000		M2a:M1a	19.60	5.54e-5
	M2a	-569.12	0.019	1.000	18.991			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-578.88	0.001	0.999	1.000	(constraint)	20.76	5.20e-6
	Mx+S1							
	(selection)	-568.50	0.001	0.001	19.999			
8235	M0	-1849.91	0.053			M3:M0	27.95	1.28e-5
	M3	-1835.93	0.000	0.018	6.110			
	M1a	-1840.04	0.000	1.000		M2a:M1a	8.21	0.017
	M2a	-1835.93	0.018	1.000	6.110			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1838.12	0.001	0.999	1.000	(constraint)	20.33	6.52e-6
	Mx+S1							
	(selection)	-1827.95	0.001	0.001	19.999			
12423	M0	-1047.84	0.122			M3:M0	26.54	2.47e-5
	M3	-1034.57	0.000	0.000	1.076			
	M1a	-1034.59	0.000	1.000		M2a:M1a	-6.45	N/A
	M2a	-1037.81	0.000	1.000	1.712			
	MX+S1	-1033.63	0.001	0.999	1.000	MX+S1 (selection):MX+S1	20.31	6.58e-6

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-1023.47	0.001	0.520	19.999			
7594	M0	-1476.36	0.247			M3:M0	23.09	1.21e-4
	M3	-1464.81	0.165	0.165	2.999			
	M1a	-1467.02	0.098	1.000		M2a:M1a	4.41	0.110
	M2a	-1464.81	0.165	1.000	2.999			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1466.89	0.079	0.999	1.000	(constraint)	19.74	8.86e-6
	Mx+S1							
	(selection)	-1457.02	0.139	0.139	19.999			
3859	M0	-1412.00	0.296			M3:M0	37.65	1.32e-7
	M3	-1393.18	0.073	0.074	2.809			
	M1a	-1395.86	0.000	1.000		M2a:M1a	5.97	0.051
	M2a	-1392.88	0.000	1.000	5.726			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1395.62	0.001	0.999	1.000	(constraint)	18.82	1.43e-5
	Mx+S1							
	(selection)	-1386.21	0.001	0.001	19.999			
3943	M0	-1453.77	0.148			M3:M0	22.99	1.27e-4
	M3	-1442.28	0.058	0.058	2.455			
	M1a	-1443.26	0.000	1.000		M2a:M1a	1.96	0.375
	M2a	-1442.28	0.058	1.000	2.455			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1442.90	0.001	0.999	1.000	(constraint)	18.59	1.62e-5
	Mx+S1							
	(selection)	-1433.61	0.001	0.001	19.999			
8059	M0	-1201.29	0.409			M3:M0	32.13	1.80e-6
	M3	-1185.23	0.000	0.314	37.797			
	M1a	-1193.82	0.000	1.000		M2a:M1a	20.24	4.0e-5
	M2a	-1183.71	0.000	1.000	47.687			
	MX+S1	-1193.53	0.001	0.999	1.000	MX+S1 (selection):MX+S1	17.70	2.58e-5

	(constraint) Mx+S1 (selection)					(constraint)		
		-1184.68	0.001	0.871	19.999			
1780	M0	-2847.18	0.033			M3:M0	33.59	9.06e-7
	M3	-2830.39	0.002	0.002	0.409			
	M1a	-2832.38	0.012	1.000		M2a:M1a	0.00	1.000
	M2a	-2832.38	0.012	1.000	74.188			
	MX+S1 (constraint)	-2827.27	0.001	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	15.81	7.00e-5
	Mx+S1 (selection)	-2819.37	0.001	0.048	19.999			
11119	M0	-1322.98	0.321			M3:M0	25.89	3.33e-5
	M3	-1310.04	0.000	0.216	2.796			
	M1a	-1312.09	0.048	1.000		M2a:M1a	4.10	0.129
	M2a	-1310.04	0.149	1.000	2.725			
	MX+S1 (constraint)	-1311.64	0.001	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	15.54	8.09e-5
	Mx+S1 (selection)	-1303.87	0.001	0.674	19.999			
5754	M0	-1018.93	0.096			M3:M0	35.63	3.46e-7
	M3	-1001.12	0.000	0.042	10.906			
	M1a	-1006.20	0.004	1.000		M2a:M1a	11.61	0.003
	M2a	-1000.40	0.020	1.000	29.686			
	MX+S1 (constraint)	-1006.07	0.004	0.999	1.000	MX+S1 (selection):MX+S1 (constraint)	12.49	4.10e-4
	Mx+S1 (selection)	-999.82	0.021	0.937	19.999			
1325	M0	-1700.52	0.207			M3:M0	26.41	2.61e-5
	M3	-1687.32	0.164	0.171	6.672			
	M1a	-1692.09	0.123	1.000		M2a:M1a	9.84	0.007
	M2a	-1687.17	0.159	1.000	8.357			
	MX+S1	-1692.01	0.124	0.999	1.000	MX+S1 (selection):MX+S1	12.04	0.001

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-1685.99	0.152	0.154	11.732			
967	M0	-1068.18	0.036			M3:M0	17.82	0.001
	M3	-1059.27	0.007	0.007	0.372			
	M1a	-1061.27	0.015	1.000		M2a:M1a	0.00	N/A
	M2a	-1061.27	0.015	1.000	46.515			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1049.15	0.001	0.999	1.000	(constraint)	11.68	0.001
	Mx+S1							
	(selection)	-1043.32	0.001	0.001	9.825			
4768	M0	-1420.24	0.124			M3:M0	28.52	9.79e-6
	M3	-1405.98	0.000	0.000	1.360			
	M1a	-1406.35	0.000	1.000		M2a:M1a	0.74	0.689
	M2a	-1405.98	0.000	1.000	1.360			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1406.12	0.001	0.999	1.000	(constraint)	11.65	0.001
	Mx+S1							
	(selection)	-1400.29	0.001	0.001	19.999			
591	M0	-2952.19	0.089			M3:M0	37.55	1.38e-7
	M3	-2933.41	0.038	0.329	1.079			
	M1a	-2933.46	0.044	1.000		M2a:M1a	0.00	1.000
	M2a	-2933.46	0.044	1.000	40.399			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-2929.54	0.001	0.001	1.000	(constraint)	11.42	0.001
	Mx+S1							
	(selection)	-2923.83	0.002	0.668	19.999			
948	M0	-1006.56	0.121			M3:M0	36.24	2.58e-7
	M3	-988.44	0.000	0.023	5.368			
	M1a	-993.03	0.000	1.000		M2a:M1a	9.19	0.010
	M2a	-988.44	0.023	1.000	5.368			
	MX+S1	-992.48	0.001	0.999	1.000	MX+S1 (selection):MX+S1	11.20	0.001

	(constraint)					(constraint)		
	Mx+S1							
	(selection)	-986.88	0.003	0.008	10.886			
6276	M0	-886.24	0.151			M3:M0	22.60	1.52e-4
	M3	-874.94	0.000	0.355	3.612			
	M1a	-876.23	0.029	1.000		M2a:M1a	2.28	0.320
	M2a	-875.09	0.053	1.000	3.622			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-875.49	0.001	0.999	1.000	(constraint)	11.03	0.001
	Mx+S1							
	(selection)	-869.97	0.001	0.316	19.999			
3843	M0	-2134.14	0.074			M3:M0	26.47	2.54e-5
	M3	-2120.90	0.033	2.996	3.224			
	M1a	-2121.96	0.000	1.000		M2a:M1a	2.18	0.336
	M2a	-2120.87	0.024	1.000	4.810			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-2121.54	0.001	0.999	1.000	(constraint)	10.80	0.001
	Mx+S1							
	(selection)	-2116.13	0.025	0.037	19.999			
7305	M0	-741.79	0.200			M3:M0	26.77	2.21e-5
	M3	-728.41	0.000	0.072	3.553			
	M1a	-730.23	0.000	1.000		M2a:M1a	4.75	0.093
	M2a	-727.85	0.000	1.000	6.800			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-730.21	0.001	0.999	1.000	(constraint)	10.60	0.001
	Mx+S1							
	(selection)	-724.91	0.001	0.001	19.999			
4930	M0	-1445.87	0.412			M3:M0	20.89	3.33e-4
	M3	-1435.43	0.000	0.000	1.357			
	M1a	-1435.99	0.000	1.000		M2a:M1a	1.13	0.569
	M2a	-1435.43	0.000	1.000	1.357			
	MX+S1	-1435.58	0.001	0.999	1.000	MX+S1 (selection):MX+S1	9.64	0.002

	(constraint)					(constraint)		
	MX+S1							
	(selection)	-1430.76	0.001	0.001	7.062			
12570	M0	-800.26	0.485			M3:M0	19.39	0.001
	M3	-790.57	0.000	4.422	46.671			
	M1a	-795.34	0.000	1.000		M2a:M1a	8.72	0.013
	M2a	-790.98	0.000	1.000	7.512			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-795.32	0.001	0.999	1.000	(constraint)	9.41	0.002
	MX+S1							
	(selection)	-790.62	0.001	0.999	18.372			
4311	M0	-1172.96	0.077			M3:M0	19.96	0.001
	M3	-1162.98	0.026	0.026	1.731			
	M1a	-1163.23	0.015	1.000		M2a:M1a	0.51	0.776
	M2a	-1162.98	0.026	1.000	1.731			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1161.47	0.001	0.999	1.000	(constraint)	8.56	0.003
	MX+S1							
	(selection)	-1157.19	0.024	0.028	16.361			
8948	M0	-1208.50	0.377			M3:M0	29.44	6.38e-6
	M3	-1193.78	0.000	1.132	12.106			
	M1a	-1197.72	0.000	1.000		M2a:M1a	7.80	0.020
	M2a	-1193.82	0.000	1.000	11.371			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1197.69	0.001	0.999	1.000	(constraint)	7.90	0.005
	MX+S1							
	(selection)	-1193.73	0.001	0.999	11.879			
2238	M0	-1730.52	0.082			M3:M0	31.67	2.24e-6
	M3	-1714.69	0.042	0.998	11.146			
	M1a	-1716.68	0.038	1.000		M2a:M1a	3.97	0.138
	M2a	-1714.69	0.042	1.000	11.148			
	MX+S1	-1715.03	0.010	0.999	1.000	MX+S1 (selection):MX+S1	7.59	0.006

	(constraint)					(constraint)		
	MX+S1							
	(selection)	-1711.23	0.038	0.999	6.233			
8058	M0	-1359.61	0.224			M3:M0	26.97	2.02e-5
	M3	-1346.12	0.000	0.000	2.222			
	M1a	-1348.13	0.000	1.000		M2a:M1a	4.01	0.135
	M2a	-1346.12	0.000	1.000	2.222			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1347.91	0.001	0.999	1.000	(constraint)	5.50	0.019
	MX+S1							
	(selection)	-1345.16	0.001	0.001	16.517			
5390	M0	-1844.81	0.121			M3:M0	5.38	0.250
	M3	-1842.12	0.090	0.090	1.902			
	M1a	-1842.29	0.072	1.000		M2a:M1a	0.34	0.842
	M2a	-1842.12	0.090	1.000	1.902			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1841.34	0.001	0.001	1.000	(constraint)	5.35	0.021
	MX+S1							
	(selection)	-1838.66	0.082	0.082	19.999			
2611	M0	-1402.83	0.124			M3:M0	49.68	4.21e-10
	M3	-1377.99	0.026	0.026	2.410			
	M1a	-1380.00	0.000	1.000		M2a:M1a	5.59	0.061
	M2a	-1377.21	0.000	1.000	74.510			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1379.27	0.001	0.999	1.000	(constraint)	5.32	0.021
	MX+S1							
	(selection)	-1376.61	0.017	0.290	3.773			
8990	M0	-1534.96	0.109			M3:M0	32.46	1.54e-6
	M3	-1518.73	0.000	1.106	22.798			
	M1a	-1521.35	0.000	1.000		M2a:M1a	5.19	0.075
	M2a	-1518.75	0.000	1.000	22.273			
	MX+S1	-1521.15	0.001	0.999	1.000	MX+S1 (selection):MX+S1	4.96	0.026

	(constraint)					(constraint)		
	MX+S1							
	(selection)	-1518.67	0.001	0.999	19.999			
2262	M0	-1415.32	0.029			M3:M0	42.32	1.43e-8
	M3	-1394.16	0.000	0.001	1.666			
	M1a	-1394.75	0.000	1.000		M2a:M1a	1.19	0.553
	M2a	-1394.16	0.000	1.000	1.880			
	MX+S1					MX+S1 (selection):MX+S1		
	(constraint)	-1394.81	0.001	0.999	1.000	(constraint)	4.92	0.026
	MX+S1							
	(selection)	-1392.34	0.001	0.999	16.789			