

2020

Speciation Dynamics Of Diverging Allopolyploid Monkeyflower (Mimulus)

Caroline Victoria Schlutius

William & Mary - Arts & Sciences, cvschlutius@gmail.com

Follow this and additional works at: <https://scholarworks.wm.edu/etd>



Part of the [Developmental Biology Commons](#), and the [Evolution Commons](#)

Recommended Citation

Schlutius, Caroline Victoria, "Speciation Dynamics Of Diverging Allopolyploid Monkeyflower (Mimulus)" (2020). *Dissertations, Theses, and Masters Projects*. Paper 1616444419.
<http://dx.doi.org/10.21220/s2-px7c-bx06>

This Thesis is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Speciation dynamics of diverging allopolyploid monkeyflower (*Mimulus*)

Caroline V. Schlutius

Sebastopol, California

Bachelor of Science, Yale University, 2018

A Thesis presented to the Graduate Faculty of
The College of William & Mary in Candidacy for the Degree of
Master of Science

Department of Biology

College of William & Mary
August 2020

APPROVAL PAGE

This Thesis is submitted in partial fulfillment of
the requirements for the degree of

Master of Science



Caroline V. Schlutius

Approved by the Committee July, 2020

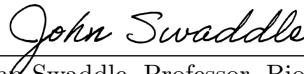


Committee Chair

Joshua Puzey, Assistant Professor, Biology
College of William & Mary



Helen Murphy, Assistant Professor, Biology
College of William & Mary



John Swaddle, Professor, Biology
College of William & Mary

ABSTRACT

Our understanding of speciation has been greatly improved with advances in genomic technology, but most of our knowledge of speciation is still built on research of diploid systems. Polyploids, however, are found in many lineages across the tree of life and exhibit considerably different evolutionary dynamics than diploids. Here, we investigate patterns of population structure and divergence in a system of two allopolyploid species of *Mimulus* (monkeyflower) that occur sympatrically in Chile: *M. luteus* and *M. cupreus*. We find that while the two species have consistent phenotypic differences across the range, they are genetically clustered into a northern and southern population (rather than by species), based on a STRUCTURE analysis of 48 whole-genome paired-end sequences of the two species across six populations in Chile. Using LUMPY and DELLY2 to locate chromosomal structural variants (SVs), we identify hundreds of SVs unique to one species or the other both across the entire range and just within the north or south. We also calculated metrics of divergence (F_{ST} and D_{XY}) in 10 kbp regions across the genome and find that these metrics were not greater within SV regions than across the whole genome. However, we did find that inversions occurred at 100–150X greater frequency within the regions of top 1% of F_{ST} and D_{XY} values compared to the across the entire genome, indicating that inversions may promote divergence. Overall, we find evidence to suggest that *M. luteus* and *M. cupreus* are currently undergoing sympatric speciation, and that inversions may help promote divergence in this system while deletions and duplications likely do not. Additionally, SV diversity is much higher than generally assumed, perhaps due to increased genomic instability in these allopolyploids, warranting future studies looking into the effects of SVs on species divergence.

TABLE OF CONTENTS

Acknowledgments	ii
Dedication	iii
List of Tables	iv
List of Figures	v
Chapter	
1 Polyploid speciation	2
1.1 Introduction	2
1.2 Methods	10
1.3 Results	17
1.4 Discussion	29
Bibliography	37

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Josh Puzey for all of his guidance during the course of my degree, both as a student and as a person. He has been incredibly supportive of my aspirations since the first day I arrived in Williamsburg, and his ability to stay grounded through all the ups and downs of this project has been invaluable. I would also like to thank Dr. Helen Murphy and Dr. John Swaddle for all of their advice and guidance through this project. Dr. Matthias Leu has been a fantastic mentor to have in the department, and I would like to thank him for the time and effort he has put into helping me grow. Dr. Arielle Cooley has also been a wonderful mentor with this project, and without her help it could not have happened. I would also like to thank Cici Zheng and Lizzie Davies for all their help in plant care.

Additionally, I would like to thank my parents and my sisters for all of their love and support that's gotten me to where I am today.

I would especially like to thank my wonderful boyfriend, Bob Galvin, for supporting me and keeping me going throughout this process. His unflagging support has meant the world to me.

Finally, I give thanks to my dog, Tots, for keeping me sane while I finished this thesis in COVID19 lockdown.

NSF Award Number 1754080.

I dedicate this thesis to the women who have come before me and paved the way for future generations of women scientists.

LIST OF TABLES

1.1	Origin and species identity of phenotyped samples.	11
1.2	Metrics of divergence between <i>M. luteus</i> and <i>M. cupreus</i> and diversity within <i>M. luteus</i> and <i>M. cupreus</i> by region.	25
1.3	Structural variant diversity by species and region.	26
1.4	Structural variant frequency across the entire genome and in genomic regions of highest divergence.	27

LIST OF FIGURES

1.1	A hypothetical demonstration of how SV accumulation may prevent recombination when different haplotypes come back into contact.	6
1.2	Hypothetical process of allopolyploid formation.	7
1.3	Study species.	9
1.4	Map of sampled populations.	12
1.5	Phenotypic differences between <i>M. luteus</i> and <i>M. cupreus</i> .	18
1.6	Dimensionality reduction of phenotype traits using PCA.	19
1.7	Genetic map of newly constructed linkage groups.	20
1.8	Synteny of the constructed <i>M. luteus</i> genome to the <i>M. guttatus</i> genome.	21
1.9	Genome-wide population structure by species and region.	23
1.10	Distributions of Metrics of Divergence (F_{ST} and D_{XY}) and Diversity(π).	24
1.11	Population structure by SV type and comparison.	28

SPECIATION DYNAMICS OF DIVERGING ALLOPOLYPLOID MONKEYFLOWER
(*MIMULUS*)

Chapter 1

Polyploid speciation

1.1 Introduction

1.1.1 Speciation

The diversity of life has long been of fascination to biologists. Many famous early biologists were concerned with just this, both documenting it – as in the case of Carl Linnaeus and Alexander von Humboldt – and investigating the process of speciation itself – as in the case of Charles Darwin and Alfred Russel Wallace. Countless studies have since been conducted to better understand how species come to be, and with the advent of large-scale, affordable sequencing technology, studies have begun to address this process from a genomic perspective as well (Marques et al., 2019; Mallet, 2007; Feder et al., 2012). One key area of research opened up by these advances is understanding the dynamics of sympatric speciation (Foote, 2018).

Sympatric speciation is a type of speciation where the diverging species remain in contact with each other throughout the speciation process, and was dismissed as impossible by several preeminent biologists such as Ernst Mayr (Mayr, 1963) and Theodosius

Dobzhansky (Kastritsis & Dobzhansky, 1967). The prevailing biological species concept (BSC; Mayr (1942)) stresses reproductive isolation as the indicator of species; allopatric speciation, in which a geographic barrier prevents gene flow, was treated as the default process. Indeed, as Mayr (1963) put it,

The mechanisms that isolate one species reproductively from others are perhaps the most important set of attributes a species has, because they are, by definition, the species criteria.

However, advances in our understanding of and ability to observe genetic processes has led some scientists to suggest that sympatric speciation may be more common than previously thought (Dieckmann & Doebeli, 1999; Foote, 2018). Moreover, researchers are also questioning the basic tenet of BSC: that speciation happens at the level of the genome. Wu (2001) argues that under the BSC, reproductive isolation must be complete across the entire genome, since introgression in parts of the genome means, by definition, that reproductive isolation is not complete. However, given that adaptation happens at the level of genes, divergence may proceed through sustained selection on the genes underlying the adaptive traits (Wu, 2001). Under this view, speciation may be considered complete when populations will not lose their divergence, and will in fact continue to diverge, when they come into contact (Wu, 2001). Wu summarizes this idea by describing two hypothetical populations of a species inhabiting different slopes of a hill, with three loci across the genome better suited to one or the other slope. The rest of the genome, however, is equally fit across the populations, and very low levels of migration are sufficient to prevent population differentiation across the genome. Consequently, the genome would vary in the extent of differentiation, with regions near the three adaptive loci being highly differentiated while the rest of the genome is not.

Studies have since explored this genic view of speciation more thoroughly, both theo-

retically and empirically (Xu et al., 2012; Doellman et al., 2018). One key area of interest in this research is the role of genetic interactions during speciation. Many traits are known to have complex genetic underpinnings [e.g. flower size (Galliot et al., 2006) and self-pollination (Sicard et al., 2011)], and it is important to understand how these genetic interactions influence the process of divergence (Feder et al., 2012). For example, if a trait under divergent selection is controlled by multiple interacting genes, how are those genes structured in the genome? Are they physically linked on a chromosome? Or do they reside further apart in the genome? What are the implications of that for divergence at those loci? While our understanding has grown substantially in animal systems such as flies (Doellman et al., 2018), mosquitoes (Turner et al., 2005), and fish (Hohenlohe et al., 2010), much less is known about speciation genomics in plants (Lexer & Widmer, 2008). Furthermore, even within plant systems, these questions have largely been restricted to diploid systems (Lexer & Widmer, 2008), with little known about how genic processes affect polyploid speciation.

1.1.2 Implications of Polyploidy

Polyploidy, the condition of having more than two sets of chromosomes, is widespread throughout the tree of life. It is found in animal lineages, such as fishes, insects, crabs and amphibians (Kenny et al., 2016; Mable et al., 2011; Leggatt & Iwama, 2003; Otto & Whitton, 2000), and is especially common in plant lineages. All seed and flowering plants share an ancient whole genome duplication (WGD) event (Jiao et al., 2011), with many lineage-specific events since (Alix et al., 2017). Polyploidy has been shown to have benefits over diploids, such as increased resistance to pathogens (Burdon & Marshall, 1981; Hannweg et al., 2016) and increased emergence of novel phenotypes (Lynch & Conery, 2000). One potential contribution to speciation dynamics in polyploids is the accumulation

of large structural variants (SVs), such as inversions, deletions, and duplications. SVs can accumulate quickly in polyploids. In *Brassica napus*, large DNA fragment losses are present after only a few generations and are widespread in induced polyploids (Gaeta et al., 2007). A study of yeast found the same phenomenon occurred after several hundred generations. Whereas diploid and haploid clones accumulated no copy number variants in their chromosomes by after 250 generations, all clones of tetraploid yeast accumulated multiple copy number variants, with chromosomes having anywhere between two and six copies, and none of the clones had a consistent number of copies between all their chromosomes (Selmecki et al., 2015). Polyploids are able to withstand such drastic changes where diploids are not because their genetic material is duplicated. A large scale deletion (or, at an extreme, the loss of an entire chromosome copy) would be fatal to a diploid, but polyploids retain another copy of those lost genes and can thus maintain their function. Given that these large scale SVs can accumulate so quickly, should populations from newly forming polyploids be separated for even several dozen generations, when they come back into contact homologous chromosomes may be so different that they can no longer recombine (Fig. 1.1).

One can think of polyploidy itself as the biggest structural change a genome can undergo. Not only is the entire genome duplicated, but there is instability in the genome as a consequence, which can lead to a greater accumulation of smaller structural variants as well. Despite these known effects of polyploidization, we know very little about the dynamics of speciation within polyploid systems. Most of what we know is in relation to the immediate reproductive isolation of allopolyploids (recent hybrids which have undergone a WGD) from their parents or the comparison of speciation rates between polyploid and diploid lineages (Van de Peer et al., 2017). For example, species radiations in polyploid lineages are associated with evolutionary times of extreme stress, such as the Cretaceous-Paleogene mass extinction (Van de Peer et al., 2017; Vanneste et al., 2014). However,

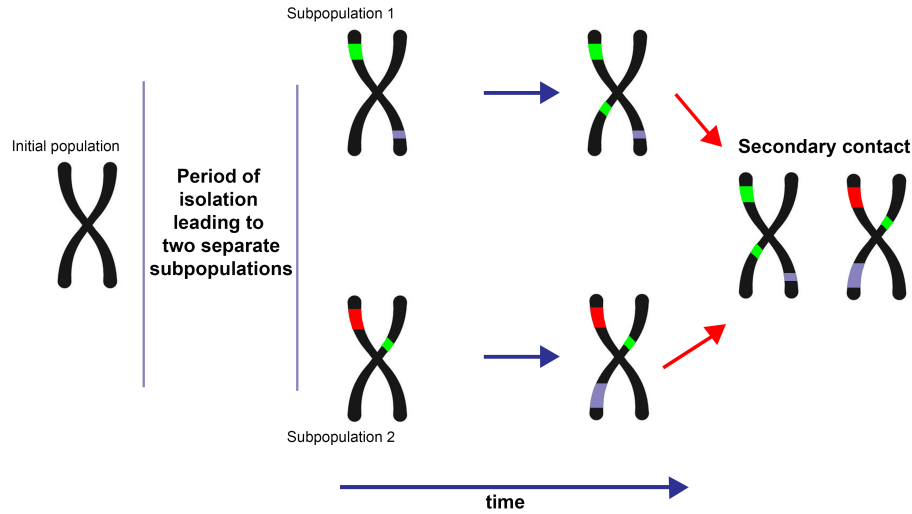


FIG. 1.1: A hypothetical demonstration of how SV accumulation may prevent recombination when different haplotypes come back into contact. Focusing on one chromosome (shown in black) which is separated in two isolated subpopulations (top and bottom), different SVs begin to accumulate in the two subpopulations (colors correspond to different SV types). This continues over time, with additional SVs accumulating on the chromosome. When the two subpopulations come back into contact, the chromosome copies may look so different that recombination between the SV regions is suppressed, thus preventing gene flow and promoting divergence in those regions.

polyploidy can also lead to different short-term speciation dynamics than in the diploid paradigm. Genome duplication has been shown to slow (Levin, 1983) or increase (Stanley et al., 1984) growth rates, change gene expression (Auger et al., 2005; Wang et al., 2006; Gaeta et al., 2007), and change gene dosage (Shaked et al., 2001; Hegarty et al., 2006). In allopolyploids, this may be particularly true.

Hybridization, the merger of two genomes within one nucleus, is an important mechanism for creating genetic diversity. Genome merger allows for novel genetic interactions and allele combinations, which in turn can lead to new phenotypes for selection to act upon. This potential for phenotypic novelty is enhanced in allopolyploids. Allopolyploids are formed from a WGD in a recent hybrid, such that there are now two copies of each parental contribution to the genome (Fig. 1.2). Since all genes exist in duplicate, there is less purifying selection acting on each copy to select against changes (Lynch & Conery,

2000). Consequently, mutations can accumulate more rapidly on one of the copies, since the other copy can perform the original function (Lynch & Conery, 2000). In this way, the functionality of gene copies may be lost, changed to a new function, or partially lost in each copy such that both copies must now be inherited to perform the original function.

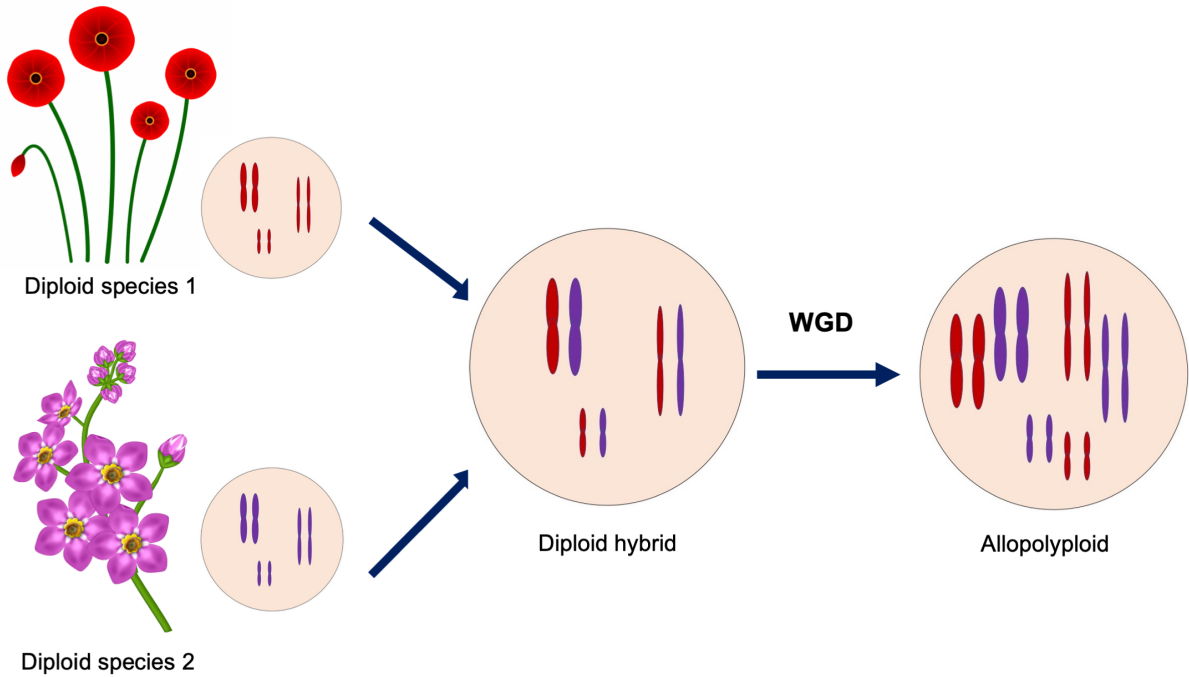


FIG. 1.2: Hypothetical process of allopolyploid formation. Here, two diploid ($2n$) species hybridize with each other, creating a diploid hybrid which has inherited one copy of its chromosomes from each parent. This recent hybrid then undergoes a WGD, whereby its entire set of chromosomes are duplicated, resulting in an allopolyploid ($4n$) with two copies of its chromosomes inherited from parent 1 (in red), and the other two inherited from parent 2 (in purple). In an allopolyploid, recombination between these two parental halves of the genome is rare, resulting in two distinct subgenomes.

In addition, the two parental contributions to the hybrid genome become two distinct, non-recombining subgenomes within the allopolyploid genome. Just like a diploid hybrid, the allopolyploid inherits half of its genome from each parent. However, after the WGD, each parental half (now called a subgenome) is doubled and then no longer recombines with the other parental half. The inability of these two subgenomes to recombine effectively

with one another means the genetic diversity between them is protected from homogenization. In other words, the each subgenome evolves somewhat independently from one another. Moreover, changes in gene expression and fractionation (gene and regulatory element loss) frequently affect the two subgenomes differently, such that one subgenome tends to become dominant (Thomas et al., 2006; Schnable et al., 2011; Pophaly & Tellier, 2015). The dominant subgenome suffers less gene loss and for duplicate copies of genes found on both subgenomes, the copy found on the dominant subgenome tends to be preferentially expressed. The sensitive (non-dominant) subgenome is subject to weaker purifying selection, and mutations accumulate more rapidly. Subgenome dominance can affect evolution in many ways (Bird et al., 2018), such as biasing gene dosage (Wright et al., 1998), reciprocal gene silencing (Werth & Windham, 1991), sequence elimination and methylation changes (Shaked et al., 2001), and novel gene expression (Osborn et al., 2003). All of these effects are known to be consequences of genome duplication, yet the effects they may have on species divergence has not been investigated.

1.1.3 Study System

In this study, we look at the speciation dynamics of two currently diverging monkeyflower species in Chile. These two putative species, *Mimulus luteus* and *Mimulus cupreus* (Fig. 1.3), are sister taxa formed from a relatively recent shared allopolyploidization event in which the widely studied *M. guttatus* was one of the parents. The ranges of the two putative species overlap, and both species prefer premontane stream habitats (Grant, 1924; von Bohlen V., 1995). The two taxa have been described as separate species for over a century (Grant, 1924) and are morphologically distinct from each other (von Bohlen V., 1995; Cooley et al., 2008). Both are pollinated by bumblebees, but *M. cupreus* receives far fewer visits and appears to be evolving a selfing life history strategy (Cooley

et al., 2008). *M. cupreus* also has lower nectar sugar content and nectar volume, as well as a different flower shape from *M. luteus*. However, the two species exhibit considerable environmental plasticity and have low genetic divergence (Beardsley et al., 2004). The taxa also readily hybridize in the greenhouse and in the field; hybrid swarms have been found in their native populations (Cooley et al., 2008). Given this, these two taxa provide a suitable model in which to investigate the effect of polyploidy on ongoing speciation. Moreover, *Mimulus* has been utilized as a model system for many different research areas (Wu et al., 2008; Yuan, 2019), and a large bank of genomic resources now exist for the genus.

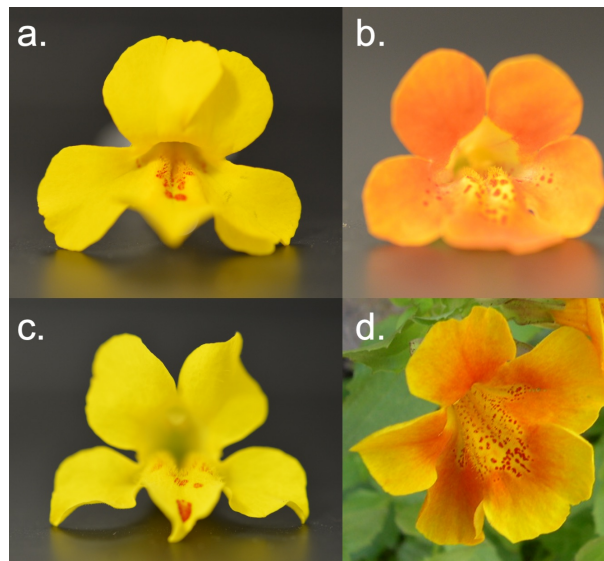


FIG. 1.3: Study species. The focal species, *M. cupreus* and *M. luteus*, are shown here along with their hybrid. *M. cupreus* has two morphs: (a) a yellow morph and (b) an orange morph. (c) *M. luteus* is always yellow. (d) The hybrid between the orange morph of *M. cupreus* and *M. luteus* has a gradient of pigment in its petals.

Here, we aim to elucidate the mechanism for morphological differentiation in the face of gene flow in a system of sympatric polyploids. Specifically, we are interested in understanding: (1) what are patterns of population structure in these neo-speciating *Mimulus*? (2) what is the diversity of SVs in these populations of diverging polyploids?

(3) what are the patterns of SV sharing in these populations?,(4) what genes underly outlier regions of divergence metrics in the genome? and (5) are those same outlier regions localized to one of the two subgenomes shared between the taxa?

1.2 Methods

1.2.1 Phenotyping

Growth Conditions

All seeds were originally collected from wild plants in 6 populations in Chile (Fig. 1.4), and the seeds used for phenotyping were inbred one to three generations prior to this experiment. Seeds were planted for 15 *M. cupreus* individuals and 17 *M. luteus* individuals. Ten pots were planted per individual, and all individuals with at least four germinated plants were included in this study (Table 1.1).

All seeds were planted in Jolly Gardener Pro-line C/B Growing Mix soil in 3-in pots. Plants were kept in trays filled with approximately 2 cm of water which was regularly refilled. All plants were housed in a growth chamber held at 21 °C with 16-hour days. Approximately three seeds were sown in each pot, and after germination plants were thinned such that no more than one plant grew in each pot.

Measurements and Analysis

Measurements were taken on the first two flowers for each plant and included: flower width, flower length, peduncle length, length of the upper calyx, length of the lower calyx, pistil length, length of the upper two stamen, and length of the two lower stamen. All measurements were taken with Neiko 0-150 mm digital calipers.

The mean was taken for the first two flowers of each plant, and then all of the replicate

TABLE 1.1: Origin and species identity of phenotyped samples.

Species	Population	Individual	Number replicates	Generations inbred
<i>M. cupreus</i>	Las Cayenas	LC11	5	1
		LC22	9	1
		LC24	8	1
		LC39	4	3
	Laguna de Laja	LL360	10	1
		LL366	6	3
	Laguna de Maule	LM418	8	2
		LM444	9	2
		LM473	10	1
	Termas de Chillan	TC287	4	3
		TC309	6	3
		TC324	9	1
	Termas del Flaco	TF385	8	1
<i>M. luteus</i>	Laguna de Laja	LL340	6	1
		LL349	10	1
	Laguna de Maule	LM409	5	2
		LM446	9	2
		LM447	5	2
	Los Queñes	LQ97	6	1
	Termas de Chillan	TC294	4	2
		TC311	9	2
		TC314	10	1
		TC318	10	2
		TC322	7	3
	Termas del Flaco	TF260	10	1
		TF262	10	1
		TF267	7	2
		TF269	10	1
		TF270	4	1

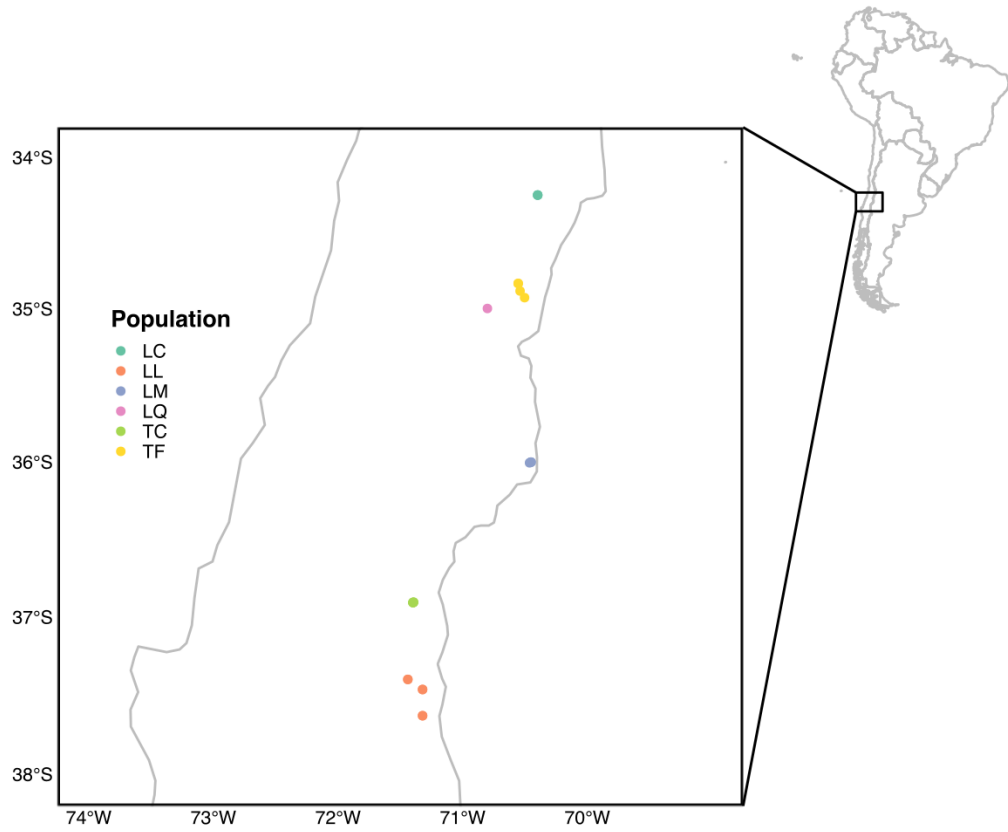


FIG. 1.4: Map of sampled populations. Seeds were collected from *M. luteus* and *M. cupreus* individuals in six populations across a latitudinal gradient in Chile.

plants for each individual were averaged to give a single final measurement for each trait for each individual. All values were then scaled and centered, and a principle component analysis was then implemented using scikit-learn version 0.23.1 (Pedregosa et al., 2011) in Python. Differences in PC1 values and all individual traits between *M. luteus* and *M. cupreus* were evaluated using a t-test.

1.2.2 Constructing Linkage Map

To improve the quality of our reference genome, we anchored the existing scaffolds of the *M. luteus* genome onto a new linkage map. To generate SNP markers, we sequenced the whole genomes of 373 F2 individuals produced from a cross between highly inbred lines of *M. l. variegatus* (a subspecies of *M. luteus*) and *M. cupreus*. DNA was extracted using the Qiagen DNeasy Plant Mini Kit (Germantown, MD, USA). For each plant, 0.09 - 0.10 g of fresh leaf tissue were collected and snap frozen in liquid nitrogen. Once extracted, the DNA was double-eluted in 30-35 uL of warm dH2O, and checked for purity and concentration using both a Nanodrop Lite (Thermo Fisher Scientific, Waltham, MA, U.S.A.) and a Qubit 4 Fluorometer (Invitrogen, Carlsbad, CA, U.S.A.). Illumina genome sequencing was performed by the Duke University Center for Genomic and Computational Biology. Sequences were demultiplexed using Stacks version 2.1 (Catchen et al., 2011) and aligned to the *M. luteus* genome (Edger et al., 2017) using bowtie2 version 2.3.4.2 (Langmead & Salzberg, 2012). Sequences were sorted and read groups were added with Picard tools version 2.18.11. GATK Unified Genotyper (DePristo et al., 2011) was used to call SNPs with a minimum base quality score of 25.

We then used these SNPs to construct a linkage map for *M. luteus* using the R/qtln package implemented in R (R Core Team, 2013; Broman et al., 2003). SNPs were filtered to exclude any sites with less than 25% genotyped individuals, and individuals were filtered to exclude individuals with more than 25% missing data. Duplicate individuals were then dropped, as were markers with distorted segregation patterns using an alpha value of 1×10^{-10} . This left 1226 markers to use in the linkage map. Markers were ordered into linkage groups using a maximum recombination fraction of 0.35 and a minimum LOD score of 6, resulting in 24 linkage groups. For each linkage group produced, the ripple function was then applied in windows of 6 markers to refine the marker order, and the

genetic distance was calculated using the Kosambi function. New orders with a positive delta LOD score compared to the original LOD score with an error probability of 0.01 were retained for the final linkage map. All markers on the 24 linkage groups were matched to locations on the *M. luteus* genome assembly to create pseudochromosomes using AllMaps (Tang et al., 2015), and annotations were lifted over using the LiftOver tool from the UCSC Genome Browser.

1.2.3 Broad Scale Genetic Structure Analysis

Extraction and Sequencing

DNA was extracted from 24 *M. luteus* and 24 *M. cupreus* plants using the Qiagen DNeasy Plant Mini Kit (Germantown, MD, USA) for whole genome sequencing. DNA was extracted using the same protocol as in 1.2.2. The samples were made into Illumina libraries using Nextera DNA Flex (Illumina, San Diego, CA, USA) and sequenced using an Illumina Novaseq 6000 with paired-end 150 basepair reads. The paired-end reads were then aligned to the constructed *M. luteus* reference genome using bowtie2 version 2.3.4.2 (Langmead & Salzberg, 2012). Mate-pairs were validated and fixed, duplicate reads were removed, and read groups were added using Picard tools version 2.18.11. GATK unified genotyper (DePristo et al., 2011) was used to call SNPs with a minimum base quality score of 25.

Genetic Clustering

We examined genetic clustering of our samples using a maximum-likelihood approach implemented in STRUCTURE version 2.3.2 (Pritchard et al., 2000). STRUCTURE runs were informed by a priori assumptions about origin from LOCPRIOR and using correlated allele frequency. Each run had a burn-in period of 100,000 generations and 1,000,000

generations of data collection. We ran STRUCTURE for K 1 to 8 with 10 iterations per K . CLUMPP version 1.1.2 (Jakobsson & Rosenberg, 2007) was used to align clusters across runs and Structure Harvester (Earl et al., 2012) was used to determine the optimal number of clusters using the second order rate of change of the log probabilities of the data.

Divergence

Using the clustering results from the STRUCTURE analysis, *M. luteus* and *M. cupreus* samples were grouped into a north and south population. For northern and southern *M. luteus* and *M. cupreus*, genetic diversity (π) was measured within each of the four populations (northern *M. luteus*, southern *M. luteus*, northern *M. cupreus*, and southern *M. cupreus*), and genetic divergence (D_{XY}) was measured between each north-south comparison for each species and between each species within the northern or southern grouping. These analyses were performed using custom scripts implemented in Python and were calculated at each position along the genome. Weir and Cockerham's F_{ST} (Weir & Cockerham, 1984) was also measured for each D_{XY} comparison using VCFtools version 0.1.16 (Danecek et al., 2011). In addition to a by-site measure, these statistics were also calculated in 10 kbp windows. All of these statistics were calculated for only variant sites across the genome.

To identify genes within the most divergent regions, the 10 kbp windows of the top 1% of F_{ST} and D_{XY} values between *M. luteus* and *M. cupreus* in the north and south were expanded by 10 kbp on each side, and these expanded regions were then annotated with genes which overlapped at least 50% with those windows. Gene annotations were taken from the *M. luteus* genome (Edger et al., 2017). This was implemented using custom scripts written in Python. All genes were compared to *Arabidopsis thaliana* genes using the NCBI BLAST tool to find the best match, and the functions of the *M. luteus* genes

were then inferred from the functions of the best match *A. thaliana* genes.

1.2.4 Structural Variant Analysis

We detected structural variants (SVs) using two softwares, LUMPY (Layer et al., 2014) and DELLY2 (Rausch et al., 2012). To prepare samples for LUMPY analysis, the paired-end reads were aligned to the *M. luteus* reference genome using the BWA-MEM algorithm (BWA 0.7.17-r1188: Li & Durbin (2009)). Discordant and split reads were identified and written into separate files using samtools version 1.9 (Li, 2011) and samblaster version 0.1.24 (Faust & Hall, 2014). Size statistics were calculated for each sample using a LUMPY script. The full reads were sorted using samtools. We then ran LUMPY according to Layer et al. (2014). To prepare samples for DELLY2 analysis, samples were sorted using samtools and duplicates were marked using Picard tools. DELLY2 was then run using the default parameters.

LUMPY and DELLY2 outputs were then grouped into the same four groups as above (see 1.2.3). The LUMPY analysis produced a separate output file for each individual, which were then grouped using svtools v0.5.1. The DELLY2 analysis was run for each group separately, producing an output for each group. The LUMPY outputs from svtools were filtered such that all SVs are supported by at least half of the samples in that group and breakend (BND) SVs were excluded. The DELLY2 outputs were filtered to exclude any SVs that did not meet the overall pass (PASS) filter, and were subsequently filtered such that at least half of the samples in the group met the pass filter for samples and BND SVs were excluded. BND SVs were excluded since it could not be confidently determined whether these were accurate or due to linkage map construction.

SVs were then further filtered and retained according to Lucek et al. (2019). Specifically, only SVs between 10kbp and 10Mbp in length that were supported by both programs

with at least 80% reciprocal coordinate overlap were retained. SVs from *M. luteus* and *M. cupreus* were then merged into a northern dataset and a southern dataset, and from these combined outputs, only SVs unique to either *M. luteus* or *M. cupreus* were kept in each of the northern and southern datasets. These datasets were compared to find unique SVs shared by northern and southern *M. luteus* and *M. cupreus*, creating datasets for SVs unique to *M. luteus* or *M. cupreus* across the entire range.

To test for genetic divergence within SV regions, a STRUCTURE analysis using $K = 2$ populations was performed for deletions, duplications, and inversions separately in the northern, southern, *M. luteus*, and *M. cupreus* datasets. SVs of each type were combined between the northern and southern datasets and between the *M. luteus* and *M. cupreus* datasets. Only bi-allelic sites were included and sites were thinned to 10 kbp except in the case of inversions found in the combined *M. luteus* and *M. cupreus* datasets, where this left very few sites and sites with more than 25% missing data were filtered out instead. These runs were performed with a burn-in period of 100,000 iterations and 1,000,000 iterations of data collection using an admixture model with no prior population assumptions, and 8 runs were performed per analysis.

Finally, to test for subgenome differences in SV location on the genome, gene annotations were mapped to the unique SVs identified in the northern and southern datasets where they existed, and genes were identified as belonging to either the A (*M. guttatus*-like) or the B (other) subgenome where possible.

1.3 Results

1.3.1 Phenotyping

A total of 17 *M. luteus* and 13 *M. cupreus* individuals were included in the phenotype analysis. Each individual had between four and ten replicates measured. *M. luteus* is larger than *M. cupreus* in all measured phenotypes (t-test: $p < 0.01$ for all, Fig. 1.5). This reinforces the findings of Cooley et al. (2008), who also found *M. luteus* to have longer, wider, and taller corollas than *M. cupreus*. Here, we've also found other floral traits, like stamen and pistil length, to be larger in *M. luteus* than *M. cupreus*. Cooley et al. (2008) also found *M. cupreus* to have developed a self-pollinating life history, which is typically accompanied by smaller flower size (Ornduff, 1969), as we see here.

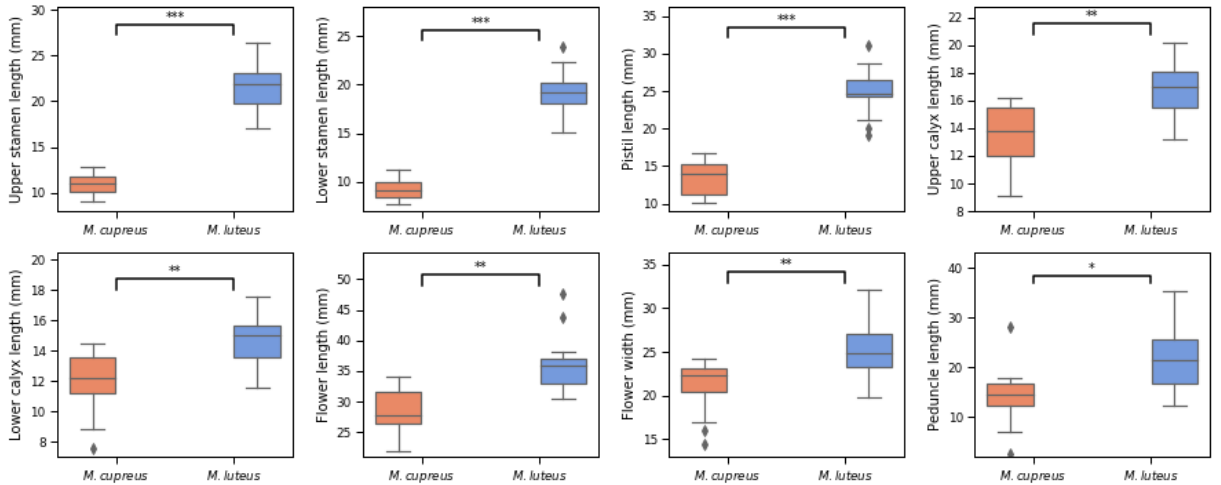


FIG. 1.5: Phenotypic differences between *M. luteus* and *M. cupreus*. Boxplots show the distribution of phenotype values for the two putative species, with the median denoted by the middle black line, the whiskers showing the interquartile range, and diamonds depicting outliers. Asterisks denote significance based on a t-test between the values of *M. luteus* and *M. cupreus* ($p < 0.01^*$, $p < 0.001^{**}$, $p < 1 \times 10^{-10}^{***}$).

M. luteus and *M. cupreus* differed significantly in their first principal component (PC) values (t-test: $p < 0.001$, Fig. 1.6a). Almost 80% of the variance in phenotypes was

explained by the first PC, indicating a high amount of collinearity among the variables. This first PC explained the variance for five of the eight measured phenotypes, with greater PC values indicating larger measurements. The second PC explained 9.3% of the variance in phenotypes, and those values were significantly different between individuals from the northern and southern populations defined by the results of the STRUCTURE analysis (see 1.3.3; t-test: $p < 0.01$, Fig. 1.6b). The three traits whose variance was most greatly explained by the second PC were peduncle length and the lengths of the upper and lower calyx, with northern populations tending to have longer calyces and shorter peduncles.

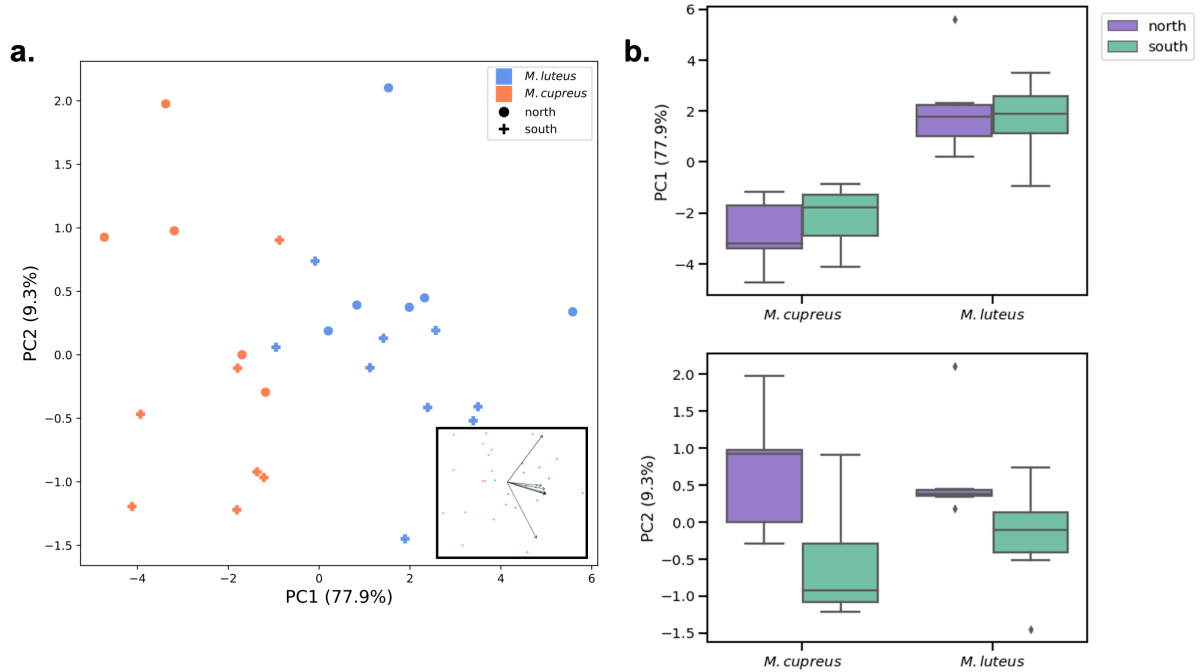


FIG. 1.6: Dimensionality reduction of phenotype traits using PCA. (a) The first two principle components (PC1 and PC2) are plotted against one another, with each point corresponding to the averaged values of an individual and the color of each point corresponding to species. The shapes show differences between northern and southern individuals, and the inset at the bottom right shows the loadings. (b) For PC1 and PC2, boxplots split by species and colored by location (north or south) show the distribution of values along the principle component. The central black bar shows the median, and the whiskers show the interquartile range. Diamonds depict outliers.

1.3.2 Linkage Map

The final linkage map consists of 24 linkage groups with a total length of 13,000 cM (Fig. 1.7). While this genome length is quite large, it was the best possible arrangement of markers to minimize overall LOD scores of marker order. However, it does suggest that homoeologous scaffolds from the two subgenomes were collapsed into a single linkage group. Of the markers used to create the linkage groups, the average inter-marker spacing is 11 cM. The largest linkage group contained 214 markers and spanned 2000 cM, while the three smallest contain just one marker. After assembling the linkage groups into pseudochromosomes using AllMaps, one linkage group was collapsed into another, leaving a total of 23 pseudochromosomes spanning 185 Mbp. This is relatively close to, but less than, the estimated number of chromosomes in *M. luteus* ($2n = 4x = 60-62$, (Vallejo-Marin, 2012)). The largest pseudochromosome is 27 Mbp long, while the smallest is 163 kbp long.

Of the 6439 scaffolds in the existing reference genome, 621 scaffolds are included in the 23 linkage groups. This accounts for 185 Mbp in the new genome compared to the 410 Mbp in the previous assembly. This new assembly also captures about half (23,318 of 46,855) of the genes in the previous assembly, which is believed to have captured nearly the entire gene space of the *M. luteus* genome (Edger et al., 2017). The median size of the excluded scaffolds is 2500 kbp. The new assembly is also largely syntenic with the *M. guttatus* genome (one of the parents of the initial allopolyploidization event; Fig. 1.8).

1.3.3 Genetic Clustering

The optimal number of clusters in the STRUCTURE analysis is $K = 2$. These two clusters separate out a northern and southern group (Fig. 1.9) and do not appear to be related to species identity (Fig. 1.9); *M. luteus* and *M. cupreus* individuals belong both

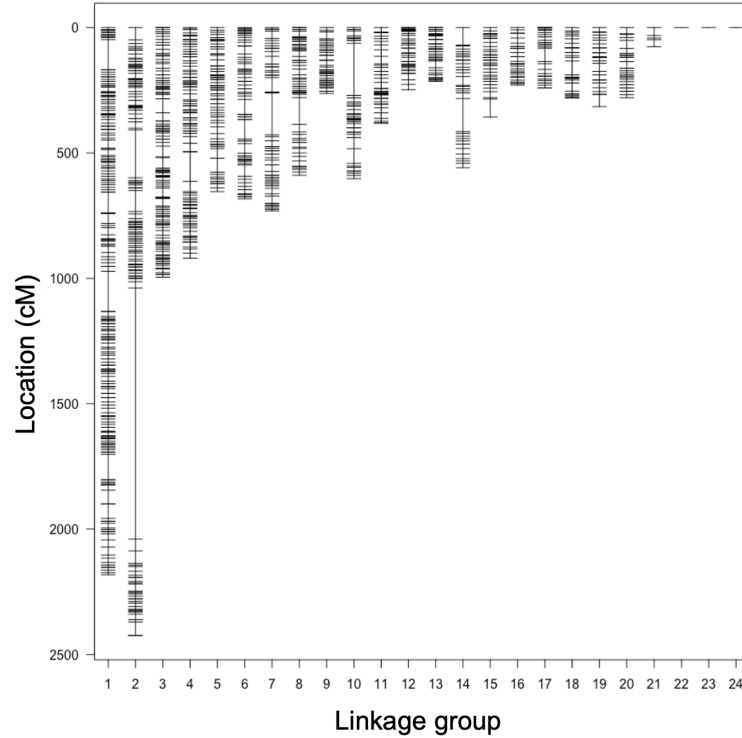


FIG. 1.7: Genetic map of newly constructed linkage groups. The markers used to create the linkage groups are shown here as horizontal bars. Linkage groups are on the x-axis, and genetic distance in cM is shown on the y-axis.

primarily to cluster one and cluster two. Most individuals belonged primarily to one cluster or the other, with relatively few individuals split roughly evenly between the two clusters. From these results, we defined a northern and a southern group, determined by the majority identity of each population with cluster one (southern) or cluster two (northern). Las Cayenas (LC) and Termas del Flaco (TF) make up the northern populations; Laguna de Maule (LM), Termas de Chillan (TC), and Laguna de Laja (LL) make up the southern populations; and Los Queñes (LQ), which is located between these two groups geographically, is split with one individual in each group. Interestingly, the only other populations which contained individuals belonging multiple groups were Las Cayenas, the northernmost population, which had one individual primarily belonging to cluster two, and Laguna de Laja, the southernmost population, which had one individual belonging

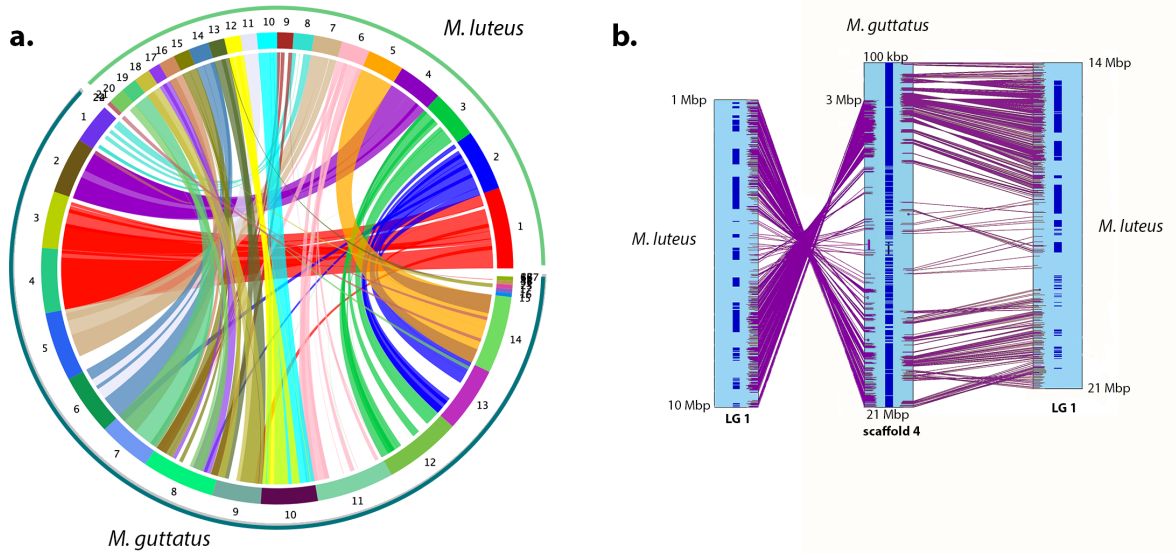


FIG. 1.8: Synteny of the constructed *M. luteus* genome to the *M. guttatus* genome. (a) The synteny between the 23 pseudochromosomes of the reconstructed *M. luteus* genome assembly and the scaffolds of the *M. guttatus* genome assembly is shown, with each color corresponding to a *M. luteus* pseudochromosome. Darker regions indicate duplications or overlapping matches. (b) The synteny between *M. luteus* linkage group 1 and *M. guttatus* scaffold 4 is highlighted, emphasizing the synteny between the same region on *M. guttatus* with multiple regions on *M. luteus*. The dark blue bars indicate the presence of genes in that region.

primarily to cluster one. There are six sequenced *M. luteus* individuals and seven sequenced *M. cupreus* individuals in the northern group, and sixteen sequenced *M. luteus* individuals and thirteen sequenced *M. cupreus* individuals in the southern group.

1.3.4 Patterns of Divergence

Standing nucleotide diversity (π) is greater in the northern populations than in the south (Fig. 1.10b, Table 1.2). In the north, the average value of π is 0.40 for both *M. luteus* and *M. cupreus*, while in the south the average value of π is 0.31 and 0.25 for *M. luteus* and *M. cupreus*, respectively. The northern and southern populations of each species are also more diverged from each other than the two species are from each other within either the north or south by both metrics of divergence (Fig. 1.10a). F_{ST} and

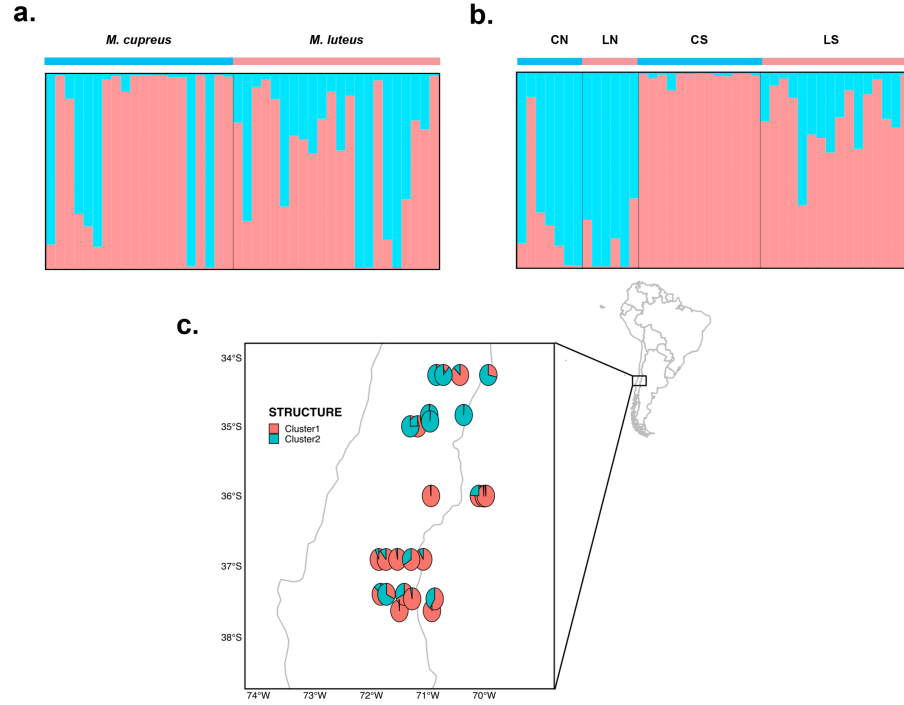


FIG. 1.9: Genome-wide population structure by species and region. (a–b) STRUCTURE plots are shown for the optimal number of clusters ($K = 2$), with individuals depicted as vertical bars. The colors correspond to the two clusters, with the relative height of each color depicting proportion cluster identity within an individual. The plots contrast cluster identity for (a) all *M. cupreus* and all *M. luteus* individuals and (b) cluster identity within the northern and southern groupings of the two putative species. (c) Map of the study region in Chile showing the locations of the sampled individuals across the sampled range. Percent cluster identity is shown for each individual in a pieplot. The longitudes of the sampled individuals have been jittered for easier comparison.

D_{XY} both measure the divergence of the species, but F_{ST} reflects differences in allele frequencies between populations and thus can be thought of as a relative measure of divergence, whereas D_{XY} shows absolute divergence between populations. The average F_{ST} value is 0.23 between the north and south populations of *M. luteus* and 0.36 between the two *M. cupreus* populations, while the average F_{ST} value is 0.05 and 0.09 between *M. luteus* and *M. cupreus* in the north and south, respectively. The average D_{XY} value show similar trends, with an average value of 0.40 and 0.44 between the northern and southern populations of *M. luteus* and *M. cupreus*, respectively. The average D_{XY} values between

M. luteus and *M. cupreus* in the north and south are 0.27 and 0.24, respectively.

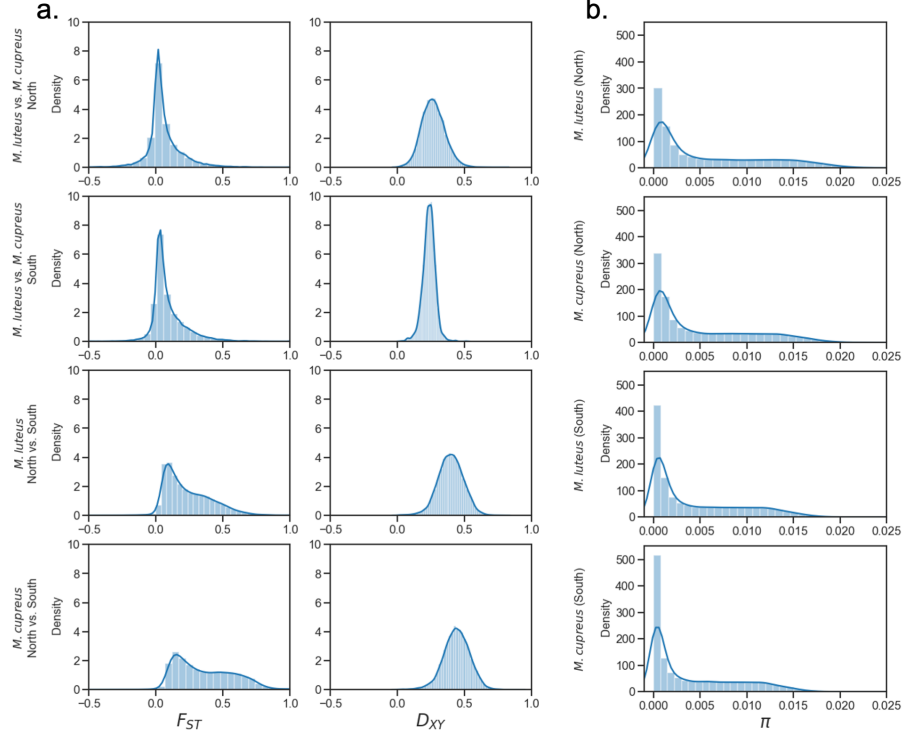


FIG. 1.10: Distributions of Metrics of Divergence (F_{ST} and D_{XY}) and Diversity (π). The distributions of F_{ST} and D_{XY} (a) are shown for *M. luteus* and *M. cupreus* in the north and south, and within each putative species between the north and south. In general, divergence is higher within each species between the north and south than it is between the two species within the north or south. The distributions of π (b) are shown for *M. luteus* and *M. cupreus* in the north and south. In general, the northern populations tend to have higher levels of genetic diversity than the southern populations for both putative species.

There were many genes found in the most divergent regions of the genome. In the expanded windows containing the top 1% of D_{XY} values between *M. luteus* and *M. cupreus* in the north and south, there were 372 unique genes, and in the windows containing the top 1% of F_{ST} values there were 1022 unique genes. Of these genes, several are likely to regulate floral organ size based on the closest *A. thaliana* gene match. For example, the genes Mlu_05771 and Mlu_17264 are among the top 1% F_{ST} regions in the north and match to AT1G59640.1 (acts to control petal size) and AT4G04885.1 (acts to regulate flower development), respectively. Many other identified genes act to control flowering

TABLE 1.2: Metrics of divergence between *M. luteus* and *M. cupreus* and diversity within *M. luteus* and *M. cupreus* by region.

Comparison	F_{ST}	D_{XY}	Group	π
<i>M. cupreus</i> (north) vs. <i>M. cupreus</i> (south)	0.36	0.45	<i>M. cupreus</i> (north)	0.40
<i>M. cupreus</i> (north) vs. <i>M. luteus</i> (north)	0.05	0.27	<i>M. luteus</i> (north)	0.40
<i>M. cupreus</i> (south) vs. <i>M. luteus</i> (south)	0.09	0.24	<i>M. cupreus</i> (south)	0.25
<i>M. luteus</i> (north) vs. <i>M. luteus</i> (south)	0.23	0.40	<i>M. luteus</i> (south)	0.31

time. Mlu_16286, Mlu_41093, and Mlu_39112 best match AT1G77300.1, AT2G06210.1, and AT3G33520.1, respectively – all of which act to regulate flowering time.

1.3.5 Structural Variants

In general, there were more SVs found in *M. cupreus* than *M. luteus*, and more SVs found in the north than the south. The difference in SVs found between *M. luteus* and *M. cupreus* may be due to *M. luteus* being used as the reference genome. Within the northern group, there are 141 SVs unique to *M. luteus* and 143 SVs unique to *M. cupreus* (Table 1.3. Within the southern group, there are 83 SVs unique to *M. luteus* and 185 SVs unique to *M. cupreus*. For each species in each group, the most common type of SV is deletions, followed by duplications, with least common being inversions. Northern *M. luteus* has the most inversions, numbering nine across the genome. There are 87 total SVs unique to *M. luteus* across the entire sampled range, and 116 total SVs unique to *M. cupreus* across the entire sampled range.

Divergence

We were interested in looking at the relationship between divergence and SVs unique to each species in two ways: (1) are SVs sufficient for divergence (i.e. is divergence greater

TABLE 1.3: Structural variant diversity by species and region.

Region	Species	Deletions	Duplications	Inversions	Total
North	<i>M. luteus</i>	82	53	6	141
	<i>M. cupreus</i>	92	45	6	143
South	<i>M. luteus</i>	48	31	4	83
	<i>M. cupreus</i>	98	77	10	185
Entire Range	<i>M. luteus</i>	47	39	1	87
	<i>M. cupreus</i>	73	41	2	116

in regions with SVs than without), and (2) do SVs promote divergence (i.e. do regions of divergence disproportionately occur in SV regions). In looking at the first relationship, we do not find evidence divergence is greater in regions of the genome containing SVs unique to one of the two putative species. For both species in the north and south, D_{XY} or F_{ST} values are not greater in regions of the genome which contained SVs compared to those that did not contain any SVs (t-test, $p > 0.05$). This is true overall and for deletions, duplications, and inversions separately.

For the second relationship, we do not find that unique deletions or duplications promote divergence, but we do see that unique inversions may promote divergence. Of the regions of the genome which contain the greatest values of D_{XY} or F_{ST} (top 1% and 5% of values), the frequency at which deletions and duplications occur is generally the same as their frequency across the whole genome. The frequency of deletions and duplications is slightly higher in regions with the highest D_{XY} or F_{ST} values, but the difference is not great (Table 1.4). In the south, though, the frequency of deletions in regions of high D_{XY} does appear to be higher than across the entire genome. The difference in frequencies for inversions in regions of divergence versus overall, however, is much more clear, especially

when looking at SVs unique to either species across the entire range. For this comparison, inversions occur at 100–150 x greater frequency in regions of high D_{XY} and F_{ST} than across the entire genome. Despite this, a chi-square test shows no significant differences by group in any of these comparisons, including inversions ($p > 0.05$ for all).

TABLE 1.4: Structural variant frequency across the entire genome and in genomic regions of highest divergence.

Region	SV Type	Overall	F_{ST} (1%)	F_{ST} (5%)	D_{XY} (1%)	D_{XY} (5%)
North	Deletions	56%	66%	60%	62%	61%
	Duplications	53%	66%	60%	57%	58%
	Inversions	8%	16%	16%	14%	17%
South	Deletions	49%	61%	62%	76%	73%
	Duplications	42%	61%	61%	52%	63%
	Inversions	14%	18%	17%	18%	22%
Entire Range	Deletions	31%	36%	36%	36%	36%
	Duplications	27%	36%	35%	31%	33%
	Inversions	0.009%	1.2%	1.5%	0.9%	1.3%

We also do not see evidence for increased population structure by species in SV regions in the genome. The STRUCTURE analysis of SV regions showed no distinction in cluster identity between species for all three types of SVs. This was true for SVs unique to *M. luteus* or *M. cupreus* in the north and south and in SVs unique to *M. luteus* or *M. cupreus* across the range sampled (Fig. 1.11). For each analysis, this was confirmed using a Mann-Whitney U Test ($p > 0.05$).

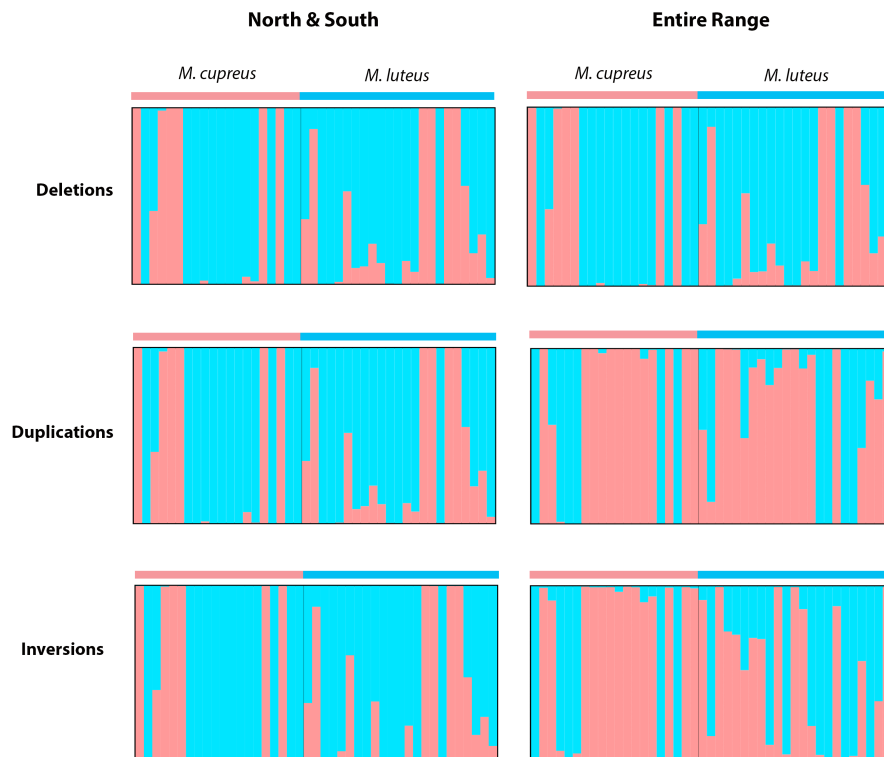


FIG. 1.11: Population structure by SV type and comparison. Population structure results from a STRUCTURE analysis with $K = 2$ are shown for deletions, duplications, and inversions unique to one of the two putative species in either the north or south (left) or the entire range sampled (right). The input loci for each analysis were taken from the regions of the genome in which an SV type of interest resided. Each vertical bar represents one individual, with the two colors showing percent identity of cluster one (pink) and cluster two (blue). These two clusters are unique to these analyses and do not correspond to either species or region. *M. cupreus* and *M. luteus* individuals are separated by a black line. For all SV types across the two comparison types, population structure does not differentiate *M. luteus* from *M. cupreus* (Mann-Whitney U Test: $p > 0.05$).

Subgenome Analysis

Structural variants did not reside preferentially on one subgenome or the other. Of the genes assigned to one subgenome or the other, 670 mapped to subgenome A and 596 to subgenome B in the SVs unique to *M. cupreus* across the sampled range, and 497 mapped to subgenome A and 435 to subgenome B in the SVs unique to *M. luteus* across the sampled range. While this indicates a slightly higher number of genes assigned to subgenome A across the SVs, most SVs contained both subgenome A and B genes, often in roughly equal

proportions. This may indicate that our linkage map collapsed homoeologous chromosomes together. Moreover, the vast majority of genes mapped to the SVs could not be assigned to one subgenome or the other. More than 99% of genes located on the SVs in both *M. cupreus* and *M. luteus* could not be assigned to a subgenome. Consequently, broader trends about subgenome dominance cannot be assessed.

1.4 Discussion

1.4.1 Population Structure

Here, we found no clear population structure distinguishing *M. luteus* from *M. cupreus*, and instead found genetic structuring differentiating a northern and southern population. However, there are clear phenotypic differences between the two putative species, with very little phenotypic variance explained by the northern and southern divide, indicating consistent genetic differentiation between *M. luteus* and *M. cupreus* at some scale. These phenotypic differences confirm and expand on previously observed differences between the two taxa (Cooley et al., 2008), and support the assertion that *M. cupreus* is moving toward a selfing habit. Consequently, this system is also a promising candidate for a case of sympatric speciation – an exciting finding in itself, as there have been few documented cases of sympatric speciation (Bolnick & Fitzpatrick, 2007). The four criteria for sympatric speciation (modified in Bolnick & Fitzpatrick (2007) from Coyne et al. (2004)) are: (1) overlapping ranges at the scale of dispersal distance for the two species, (2) speciation is complete, (3) the two species must be sister species, and (4) allopatric speciation is very unlikely. *M. luteus* and *M. cupreus* certainly have overlapping ranges in the areas sampled for this research, and their species descriptions also describe overlapping ranges (von Bohlen V., 1995). However, their ranges are not identical; *M. luteus* can

occupy a greater range of altitudes as *M. cupreus*, and its range extends slightly further to the north and south (von Bohlen V., 1995). Despite this, the large majority of their ranges overlap, suggesting the first criterion is met. As for the third condition, the two species are at least in the same species complex as each other within the Simiolus group of *Mimulus* (Vickery, 1966). Only one other species (*M. tigrinus*) is also a part of this complex, indicating they share at least a recent speciation event, if not the most recent. The lack of genome-wide divergence and the genetic structuring at the population-level found here, however, suggests that it is unlikely an additional speciation event separates the two species and thus this system likely meets the third criterion. This would need to be more rigorously tested before it can be confirmed, however. It also appears unlikely that these two species have experienced allopatric speciation and are now in secondary contact, indicating the fourth condition is met. Were this a case of secondary contact, we would expect that genetic clustering would be evident at the species level, since the genomes of the two species would have had a chance to evolve in isolation from one another. Instead, we see no evidence of genome-wide structuring by species. This brings us to the second criterion: is the speciation process complete?

By the standard BSC, the answer is no. The two species still hybridize, demonstrating a lack of reproductive isolation between them, and at a genome level they are not genetically distinct. However, the answer is less certain when considering the more recently proposed genic view of speciation. This view states that complete reproductive isolation is not necessary to demarcate species, but rather the condition that two species will not lose their divergence when they come into contact and will instead continue to diverge (Wu, 2001). Given the clear and consistent phenotypic differences between the two putative species, there must be some fraction of the genome which is differentiated between *M. luteus* and *M. cupreus*. This shows that while mating may be random with respect to most loci in the genome (demonstrated by the lack of species-level structuring across the

genome), it is necessarily nonrandom at the loci responsible for the observed phenotypic differences. By the genic species definition, whether or not *M. luteus* and *M. cupreus* are species depends on whether the divergence in these differentiated regions is broken down by hybridization.

Determining this is a necessary next step to understanding whether this is a case of sympatric speciation. The consistency of the phenotypic differences across the range suggest that these differences arose before the northern and southern groups differentiated, but despite this, we see much higher levels of divergence between the north and south than between species. To understand whether this is due to current gene flow breaking down genome differentiation or whether this is a result of past gene flow (which has since decreased) requires subsequent studies looking at current levels of gene flow between the two taxa. Since we have observed that *M. cupreus* is evolving a selfing habit and is visited less by its bumblebee pollinator than *M. luteus* (Cooley et al., 2008), one such study could look at gene flow at a pollinator-level by collecting pollen samples from bumblebees. It would also be interesting to observe the fitness of the hybrids between the two taxa, since they have been observed in the wild, to see if divergence is reinforced by decreased fitness in the hybrids. Additionally, we found here that several genes in the most divergent regions of the genome may be related to flowering time. It would thus be interesting to look at phenological differences between the two taxa to see if gene flow is reduced by mismatched phenologies. These questions must be answered before it can be determined if these two taxa are species or are on their way to becoming species, since it is possible that they are not on the path to speciation and are instead in equilibrium as different races of the same species (Matessi et al., 2002).

1.4.2 Structural Variant Diversity

We found a very high number of SVs present in this system; there were roughly 200 SVs unique to either *M. luteus* or *M. cupreus* across the sampled range and even more within just the northern or southern populations. Some of the deletions or duplications in particular may be an artifact of the linkage map assembly; the high cM length of the map (reflecting high recombination rates), in conjunction with the subgenome analysis showing genes assigned to both subgenomes residing on the same SV, indicate a high likelihood of homoeologous chromosomes (the comparable chromosome copies from the two subgenomes) getting collapsed within the same pseudochromosome in our assembly. This collapse would likely result in false deletions or duplications, though false inversions are less likely to be due to homoeologous chromosome collapse. Despite this, our SV filtering was performed such that all SVs must be shared by a majority of individuals in each species and that the SVs must be found in only one of the two species. This means that while all SVs may not reflect biologically accurate chromosomal rearrangements, they are all found in the majority of the individuals of one species and not the other. Moreover, our linkage map only accounted for roughly half of the estimated genome size of *M. luteus* and half of its genes. Consequently, there may be many more SVs across the genome that we were unable to detect.

Additionally, these findings give merit to the idea that chromosomal structural rearrangements should be explicitly taken into account when studying divergence. While numerous studies have reflected on the transition from small regions of divergence in the genome to larger genomic islands of divergence (Turner et al., 2005; Riesch et al., 2017), some even explicitly acknowledging the roles that chromosomal rearrangements may play (Ortíz-Barrientos et al., 2002), few speciation studies have directly address the role that SVs play in this process (Wellenreuther et al., 2019). Indeed, one review of plant specia-

tion makes the direct assumption that chromosomal rearrangements play a negligible role in the process (Lexer & Widmer, 2008). However, even in diploid systems, such as walking sticks (Lucek et al., 2019) and sunflowers (Rieseberg et al., 1995), structural variants have been shown to be quite common and likely involved in divergence. As genomic technologies continue to advance, scientists will be able to more directly consider the diversity of SVs and their role in speciation (Wellenreuther et al., 2019).

1.4.3 Structural Variant Divergence

Here, we found some evidence of increased divergence in inversion loci, but not in deletion or duplication loci. F_{ST} and D_{XY} values were not greater in regions containing unique deletions or duplications, nor were the genomic regions with the greatest F_{ST} and D_{XY} values enriched for deletions or duplications. For unique inversions, F_{ST} and D_{XY} values were also not greater, but we did observe an enrichment of inversions in the genomic regions with the greatest F_{ST} and D_{XY} values. This largely corroborates what Lucek et al. (2019) found in SV differentiation in two ecotypes of walking sticks (*Timema*), where there was no observed difference in F_{ST} values was found between SV regions and the null distribution across the genome for deletions, duplications, and inversions.

The lack of differentiation in SV regions could be due to the fact that there are so many of them. Given that the SVs number in the hundreds, it is not surprising that divergence is not necessarily increased in SVs. It is also perhaps unsurprising that regions of the genome with the top 1% of F_{ST} and D_{XY} values are enriched for inversions, but not deletions or duplications. Inversions are effective mechanisms for reducing recombination in inverted loci, and have been shown to capture genes related to adaptation in several cases of the closely related *Mimulus guttatus* (Twyford & Friedman, 2015; Lee et al., 2016), as well as in other systems (e.g. Ullastres et al., 2014; Kapun & Flatt, 2019). With

reduced recombination, and thus less gene flow, these regions then have a greater chance of differentiation. Ultimately, though, a better-resolved genome assembly for *M. luteus* will allow us to better understand its genome structure and assess with greater accuracy the distribution and diversity of SVs. With this information, we will also be better able to identify regions of divergence and their associations with SVs.

1.4.4 Genes of Divergence

In the regions of divergence identified here, we found several genes related to the phenotypic differences we measured in this study. The closest *A. thaliana* matches for two *M. luteus* genes found in these regions, Mlu_05771 and Mlu_17264, are AT1G59640.1 and AT4G04885.1, regulate flower size, which we found to be significantly different between *M. luteus* and *M. cupreus*. Interestingly, many of the genes in these regions of divergence also regulate flowering time. For example, AT1G77300.1, AT2G06210.1, and AT3G33520.1 are the closest *A. thaliana* matches for Mlu_16286, Mlu_41093, and Mlu_39112, respectively, and have been shown to regulate flowering time in different ways. While flowering time was not measured here and would need to be studied further to confirm that it varies between *M. luteus* and *M. cupreus*, flowering time divergence has been shown to rapidly develop in *Brassica rapa* allopolyploids (Pires et al., 2004) and *M. guttatus* populations (Hall & Willis, 2006). The genes found here which relate to flowering time suggest that this may be a mechanism by which gene flow is reduced between the two species, in addition to the shift towards selfing in *M. cupreus* shown previously (Cooley et al., 2008). However, all of these inferences are based on the closest *A. thaliana* genes, which is not a close relative of *Mimulus*. Consequently, the function of genes within regions of divergence should be tested either within *Mimulus* or a more closely related relative before making conclusions about their true functions. Regardless, the functions of the *A. thaliana* matches found

here give us a good starting point for further exploration of the differences of *M. luteus* and *M. cupreus*.

1.4.5 Patterns of Subgenome Divergence

Finally, we did not find evidence of subgenome dominance in this study, but it is likely that this is grounded in the incomplete linkage map than in biological reality. Subgenome dominance has been shown to rapidly develop in resynthesized allopolyploids (Edger et al., 2017) and is documented in many older allopolyploid genomes (Flagel et al., 2008; Cheng et al., 2012; Doyle et al., 2008). Consequently, it is likely that there is some level of subgenome dominance in this system, but we are unable to observe it because of an incorrectly assembled genome. As mentioned previously, it is likely that homoeologous chromosomes were collapsed in this constructed genome assembly, meaning that the related regions on the chromosome copy belonging to subgenome A and B were likely condensed into a single region or consecutive regions on the created linkage map and resulting genome assembly. Because of this, the two subgenomes cannot be evaluated separately, as we observed with genes belonging to both subgenomes, often in roughly equal numbers, located on a single SV. Even beyond this genome assembly, however, most genes annotated for *M. luteus* have yet to be assigned to a subgenome. Of the genes included here, less than 1% were mapped to a subgenome. Therefore, to properly evaluate subgenome dominance, a better genome assembly must be created, and more genes must be mapped to their respective subgenomes.

1.4.6 Conclusions

Overall, we found substantial evidence that *M. luteus* and *M. cupreus* are phenotypically differentiated but not genomically differentiated. Further studies looking at current

levels of gene flow and mechanisms of divergence, such as pollinator or phenology differences, must be done before we can conclude whether or not these species are in the process of sympatric speciation. Our results show that structural variants are numerous in this system and suggest that it may be important to explicitly study their role in divergence in other systems. In this system, our results indicate that inversions may contribute to divergence, but this should be more thoroughly tested with a more complete and better annotated reference genome for *M. luteus*. We identified several genes in regions of elevated F_{ST} which controlled flower size, which we measured to differ between the species, as well as many which controlled flowering time, which could contribute to limiting gene flow between the species. Future work to improve the reference genome will allow for a more thorough search for the genes which are being divergently selected upon in the two species, as well as allow for an analysis of subgenome dominance.

Bibliography

- Alix, K., Gérard, P. R., Schwarzacher, T., & Heslop-Harrison, J. (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany*, 120(2), 183–194.
- Auger, D. L., Gray, A. D., Ream, T. S., Kato, A., Coe, E. H., & Birchler, J. A. (2005). Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, 169(1), 389–397.
- Beardsley, P. M., Schoenig, S. E., Whittall, J. B., & Olmstead, R. G. (2004). Patterns of evolution in western north american mimulus (phrymaceae). *American Journal of Botany*, 91(3), 474–489.
- Bird, K. A., VanBuren, R., Puzey, J. R., & Edger, P. P. (2018). The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist*, 220(1), 87–93.
- Bolnick, D. I., & Fitzpatrick, B. M. (2007). Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Evol. Syst.*, 38, 459–487.
- Broman, K. W., Wu, H., Sen, S., & Churchill, G. A. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7), 889–890.
- Burdon, J., & Marshall, D. (1981). Inter-and intra-specific diversity in the disease-response of glycine species to the leaf-rust fungus phakopsora pachyrhizi. *The Journal of Ecology*, (pp. 381–390).

- Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, genomes, genetics*, 1(3), 171–182.
- Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., & Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PloS one*, 7(5), e36442.
- Cooley, A., Carvallo, G., & Willis, J. (2008). Is floral diversification associated with pollinator divergence? flower shape, flower colour and pollinator preference in Chilean *Mimulus*. *Annals of Botany*, 101(5), 641–650.
- Coyne, J. A., Orr, H. A., et al. (2004). *Speciation*, vol. 37. Sinauer Associates Sunderland, MA.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15), 2156–2158.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), 491.
- Dieckmann, U., & Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, 400(6742), 354–357.
- Doellman, M. M., Ragland, G. J., Hood, G. R., Meyers, P. J., Egan, S. P., Powell, T. H., Lazorachak, P., Glover, M. M., Tait, C., Schuler, H., et al. (2018). Genomic differentiation

- during speciation-with-gene-flow: Comparing geographic and host-related variation in divergent life history adaptation in *rhagoletis pomonella*. *Genes*, 9(5), 262.
- Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., & Wendel, J. F. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annual review of genetics*, 42, 443–461.
- Earl, D. A., et al. (2012). Structure harvester: a website and program for visualizing structure output and implementing the evanno method. *Conservation genetics resources*, 4(2), 359–361.
- Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., Bowman, M. J., et al. (2017). Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *The Plant cell*, 29(9), 2150–2167.
- Faust, G. G., & Hall, I. M. (2014). Samblaster: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17), 2503–2505.
- Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in genetics*, 28(7), 342–350.
- Flagel, L., Udall, J., Nettleton, D., & Wendel, J. (2008). Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC biology*, 6(1), 16.
- Foote, A. D. (2018). Sympatric speciation in the genomic era. *Trends in ecology & evolution*, 33(2), 85–95.
- Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E., & Osborn, T. C. (2007). Genomic

- changes in resynthesized brassica napus and their effect on gene expression and phenotype. *The Plant Cell*, 19(11), 3403–3417.
- Galliot, C., Hoballah, M. E., Kuhlemeier, C., & Stuurman, J. (2006). Genetics of flower size and nectar volume in petunia pollination syndromes. *Planta*, 225(1), 203–212.
- Grant, A. L. (1924). A monograph of the genus mimulus. *Annals of the Missouri Botanical Garden*, 11(2/3), 99–388.
- Hall, M. C., & Willis, J. H. (2006). Divergent selection on flowering time contributes to local adaptation in mimulus guttatus populations. *Evolution*, 60(12), 2466–2477.
- Hannweg, K., Steyn, W., & Bertling, I. (2016). In vitro-induced tetraploids of plectranthus esculentus are nematode-tolerant and have enhanced nutritional value. *Euphytica*, 207(2), 343–351.
- Hegarty, M. J., Barker, G. L., Wilson, I. D., Abbott, R. J., Edwards, K. J., & Hiscock, S. J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. *Current Biology*, 16(16), 1652–1659.
- Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS genetics*, 6(2).
- Jakobsson, M., & Rosenberg, N. A. (2007). Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), 1801–1806.
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345), 97–100.

- Kapun, M., & Flatt, T. (2019). The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Molecular ecology*, *28*(6), 1263–1282.
- Kastritsis, C. D., & Dobzhansky, T. (1967). *Drosophila pavlovskiana*, a race or a species? *American Midland Naturalist*, (pp. 244–248).
- Kenny, N. J., Chan, K. W., Nong, W., Qu, Z., Maeso, I., Yip, H. Y., Chan, T. F., Kwan, H. S., Holland, P. W., Chu, K. H., et al. (2016). Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*, *116*(2), 190–199.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). Lumpy: a probabilistic framework for structural variant discovery. *Genome biology*, *15*(6), R84.
- Lee, Y. W., Fishman, L., Kelly, J. K., & Willis, J. H. (2016). A segregating inversion generates fitness variation in yellow monkeyflower (*Mimulus guttatus*). *Genetics*, *202*(4), 1473–1484.
- Leggatt, R. A., & Iwama, G. K. (2003). Occurrence of polyploidy in the fishes. *Reviews in Fish Biology and Fisheries*, *13*(3), 237–246.
- Levin, D. A. (1983). Polyploidy and novelty in flowering plants. *The American Naturalist*, *122*(1), 1–25.
- Lexer, C., & Widmer, A. (2008). The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 3023–3036.

- Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25, 1754–1760.
- Lucek, K., Gompert, Z., & Nosil, P. (2019). The role of structural genomic variants in population differentiation and ecotype formation in *timema cristinae* walking sticks. *Molecular ecology*, 28(6), 1224–1237.
- Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *science*, 290(5494), 1151–1155.
- Mable, B., Alexandrou, M., & Taylor, M. (2011). Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology*, 284(3), 151–182.
- Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133), 279.
- Marques, D. A., Meier, J. I., & Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends in ecology & evolution*, 34(6), 531–544.
- Matessi, C., Gimelfarb, A., & Gavrillets, S. (2002). Long-term buildup of reproductive isolation promoted by disruptive selection: How far does it go? *Selection*, 2(1-2), 41–64.
- Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press.
- Mayr, E. (1963). *Animal Species and Evolution*, chap. 5. Harvard University Press.
- Ornduff, R. (1969). Reproductive biology in relation to systematics. *Taxon*, 18(2), 121–133.

- Ortíz-Barrientos, D., Reiland, J., Hey, J., & Noor, M. A. (2002). Recombination and the divergence of hybridizing species. In *Genetics of Mate Choice: From Sexual Selection to Sexual Isolation*, (pp. 167–178). Springer.
- Osborn, T. C., Pires, J. C., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H.-S., Comai, L., Madlung, A., Doerge, R., Colot, V., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in genetics*, 19(3), 141–147.
- Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics*, 34(1), 401–437.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pires, J. C., Zhao, J., Schranz, M. E., Leon, E. J., Quijada, P. A., Lukens, L. N., & Osborn, T. C. (2004). Flowering time divergence and genomic rearrangements in resynthesized brassica polyploids (brassicaceae). *Biological Journal of the Linnean Society*, 82(4), 675–688.
- Pophaly, S. D., & Tellier, A. (2015). Population level purifying selection and gene expression shape subgenome evolution in maize. *Molecular biology and evolution*, 32(12), 3226–3235.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R

Foundation for Statistical Computing, Vienna, Austria.

URL <http://www.R-project.org/>

- Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339.
- Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A. A., Farkas, T. E., Lucek, K., Hellen, E., Soria-Carrasco, V., Dennis, S. R., et al. (2017). Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology & Evolution*, 1(4), 0082.
- Rieseberg, L. H., Van Fossen, C., & Desrochers, A. M. (1995). Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature*, 375(6529), 313–316.
- Schnable, J. C., Springer, N. M., & Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, 108(10), 4069–4074.
- Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shores, N., Sorenson, A. L., De, S., Kishony, R., Michor, F., Dowell, R., et al. (2015). Polyploidy can drive rapid adaptation in yeast. *Nature*, 519(7543), 349.
- Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., & Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *The Plant Cell*, 13(8), 1749–1759.
- Sicard, A., Stacey, N., Hermann, K., Dessoly, J., Neuffer, B., Bäurle, I., & Lenhard, M. (2011). Genetics, evolution, and adaptive significance of the selfing syndrome in the genus *capsella*. *The Plant Cell*, 23(9), 3156–3171.

- Stanley, J. G., Hidu, H., & Allen Jr, S. K. (1984). Growth of american oysters increased by polyploidy induced by blocking meiosis i but not meiosis ii. *Aquaculture*, *37*(2), 147–155.
- Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015). Allmaps: robust scaffold ordering based on multiple maps. *Genome biology*, *16*(1), 3.
- Thomas, B. C., Pedersen, B., & Freeling, M. (2006). Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research*, *16*(7), 934–946.
- Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in anopheles gambiae. *PLoS biology*, *3*(9), e285.
- Twyford, A. D., & Friedman, J. (2015). Adaptive divergence in the monkey flower *mimulus guttatus* is maintained by a chromosomal inversion. *Evolution*, *69*(6), 1476–1486.
- Ullastres, A., Farré, M., Capilla, L., & Ruiz-Herrera, A. (2014). Unraveling the effect of genomic structural changes in the rhesus macaque-implications for the adaptive role of inversions. *BMC genomics*, *15*(1), 530.
- Vallejo-Marin, M. (2012). *Mimulus peregrinus* (phrymaceae): A new british allopolyploid species. *PhytoKeys*, *14*, 1–14.
- Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411.
- Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the cretaceous–paleogene boundary. *Genome research*, *24*(8), 1334–1347.

- Vickery, R. K. (1966). Speciation and isolation in section *simiolus* of the genus *mimulus*. *Taxon*, (pp. 55–63).
- von Bohlen V., C. (1995). The genus *mimulus* l. (scrophulariaceae) in chile. *Gayana Bot.*, 52(1), 7–28.
- Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R., Comai, L., et al. (2006). Genomewide nonadditive gene regulation in *arabidopsis* allotetraploids. *Genetics*, 172(1), 507–517.
- Weir, B. S., & Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *evolution*, 38(6), 1358–1370.
- Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond snps: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology*, 28(6), 1203–1209.
- Werth, C. R., & Windham, M. D. (1991). A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *The American Naturalist*, 137(4), 515–526.
- Wright, R. J., Thaxton, P. M., El-Zik, K. M., & Paterson, A. H. (1998). D-subgenome bias of xcm resistance genes in tetraploid *gossypium* (cotton) suggests that polyploid formation has created novel avenues for evolution. *Genetics*, 149(4), 1987–1996.
- Wu, C., Lowry, D., Cooley, A., Wright, K., Lee, Y., & Willis, J. (2008). *Mimulus* is an emerging model system for the integration of ecological and genomic studies. *Heredity*, 100(2), 220.
- Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14, 851–865.

- Xu, S., Schlüter, P. M., & Schiestl, F. P. (2012). Pollinator-driven speciation in sexually deceptive orchids. *International Journal of Ecology*, 2012.
- Yuan, Y.-W. (2019). Monkeyflowers (mimulus): new model for plant developmental genetics and evo-devo. *New Phytologist*, 222(2), 694–700.