INVESTIGATING INTROGRESSION AT THE MAMYB2 LOCUS IN MIMULUS PUNICEUS

By

Kayden Elijah Kendrick

A THESIS

Presented to the Department of Biology
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

May 2021

An Abstract of the Thesis of

Kayden Kendrick for the degree of Bachelor of Science
in the Department of Biology to be taken June 2021

Title: Investigating Introgression at the *MaMyb2* Locus in *Mimulus Puniceus*

Approved: *Matthew Streisfeld, Associate Professor/Director of the Institute of Ecology and Evolution*
Primary Thesis Advisor

The red ecotype of *Mimulus puniceus* gained its ability to grow red flowers through the introgression of *MaMyb2* which functions in the anthocyanin production pathway required to produce red pigments. This thesis provides a brief literature review on adaptive introgression and radiations as well as background information pertaining to the monkeyflower radiation of Southern California. Utilizing genomic analysis methods, this thesis investigates the 100 kilobase region surrounding the *MaMyb2* locus. Through this analysis, it is shown that the region surrounding *MaMyb2* in the red ecotype of *M. puniceus* is genetically more similar to *Mimulus rutilus,* rather than to the accepted donor taxa of the introgressed gene, *Mimulus aridus.* This thesis also discusses the surrounding genomic landscape and its evolutionary implications.

Acknowledgements

I would first and foremost like to thank Dr. Matthew Streisfeld for serving as my primary thesis advisor. His guidance, patience and help has been paramount to completing this thesis, especially during the challenging times that the Covid-19 pandemic has presented. I would like to thank Aidan Short for his assistance in completing the Twisst analysis and serving as my Second reader. I would like to thank Dr. Corinne Bayerl for serving as my CHC representative and for being an excellent professor within the CHC and helping to hone my critical thinking and writing skills in years past.

I would also like to thank those in my personal life who helped me get through this challenging year and my academic career at University of Oregon. First, I would like to thank Carlin Otterstedt and Zach Hedeen for coaching me through this year as part of the University of Oregon rowing team. I would like to thanks my friends Catherine O'Hara, Isaac Estrada, Jose Romero, Emily Mauelshagen, Leon Burris and Hannah Sebring. And finally I would like to thank Olivia Stankey for her help in adjusting to the University and life outside of rural eastern Oregon.

Introduction

Modern genetic analysis has created greater opportunities for the research into the biological world at a scale that was previously thought impossible. Genetic sequencing technologies have become increasingly cheaper and more accessible to scientists around the world. This expansion has given scientists the ability to ask more nuanced questions about the details that drive evolution and the divergence of taxa into species. One method of studying this divergence has been the close study of radiations, groupings of closely related taxa that have not yet fully diverged and still undergo genetic exchange amongst one another. The most famous of which motivated the study of natural selection under Charles Darwin through his study of the Galapagos Finches. These radiations are extremely important to the current understanding of the mechanisms of evolution and in particular, how members of these radiations exchange beneficial genetic information with one another through a process called adaptive introgression.

Literature Review of Radiations and Adaptive Introgression

Radiations hold a place of great interest in evolutionary biology. Most radiations occur in such a way that speciation is able to occur quickly (Harrison and Larson 2014). Speciation is the process in which two closely related populations become reproductively isolated. This can occur due to geographic separation, genetic separation, or in other ways involving differing mechanics of mating. Radiations are often driven to this separation through their ability to fill ecological niches through quick adaptations. The

classic example of a radiation is Darwin's finches and their adaptations to the various food types available on the Galapagos islands that helped each species adapt into a unique niche. These ecological niches would normally take thousands of generations of adaption to fill , however, radiations often occur upon much smaller time scales (Losos 2010). These relatively short time scales are often pushed by ecological opportunity that allows for the rapid divergence of a species into new taxa. This could be caused by a new way to utilize resources or the extinction of a species that previously filled the niche. Regardless of the conditions that opens up a new niche, any circumstance will be limited by the presence of the necessary genetic traits that allows for the divergence of the taxa (Losos 2010).

These genetic traits can arise in a variety of ways, and amongst the slowest of these are random *de novo* mutations. These *de novo* mutations, however, are not a likely explanation of rapid adaptive radiations (Marques, Meier and Seehausen 2019). The time in which it takes for such mutations to arise and effect radiations is extensive. It is much more likely that the variations required for the rise of new traits already exist as standing genetic variation in a population (Marques, Meier and Seehausen 2019) (Nelson and Cresko 2018). An expedient route in which this standing variation can be made available to other taxa within a radiation is through the transfer of useful traits from one related taxa to another through adaptive introgression (Suarex-Gonzalez, Lexer and Cronk 2018). Introgression is the process in which two closely related taxa interbreed and share genetic information. It is considered adaptive introgression if the traits result in an increase of fitness for the recipient population (Burgarella, et al. 2019). This can result in somewhat immediate benefits being introduced to a population that leads to greater

levels of fitness and the ability to capitalize on ecological opportunities that arise (Losos 2010).

Understanding the effects of introgression and how it occurs is important to the overall study of radiations and evolution. Many radiations that are actively studied by evolutionary biologists have been shown to have utilized standing genetic variation, either from within their own genome or through introgression with other closely related taxa, that predates the beginning of their radiations such as Darwin's finches, threespine stickleback and heliconius butterflies (Marques, Meier and Seehausen 2019). However, introgression is not just of interest to evolutionary biologists. Introgression can also occur between cultivated crop plants and their wild relatives (Burgarella, et al. 2019). As climate change and human-population expansion continues, the conditions in which crops are grown will likely require the use of new methods to increase the hardiness of crop plants and their genetic diversity. This will require many methods of manipulation of crop plants and can include the use of introgression to introduce variants that can allow for increased fitness for the recipient crop plants (Burgarella, et al. 2019) (Wang, Chen and Ma 2019).

To find areas where introgression has occurred it is usually useful to look for regions along the genome that show signatures of natural selection taking place. Selection on a genetic scale occurs when a variant confers a benefit and is thus passed on to more offspring, increasing its frequency within a population. Signatures of selection are often looked for through genomic approaches. Regions that show signatures of selection are typically indicated by regions of lowered genetic diversity within populations whereas regions that are not under the effects of selection are often more

genetically diverse (Martin and Jiggins, Interpreting the genomic landscape of introgression 2017). In Darwin's finches it was determined that beak shape was largely influenced by the *ALX1* locus, which also shows signatures of selection through a low divergence within populations of the finches (Lamichhaney, et al. 2015). This was determined by processing the genome of the finches and looking for sites that were fixed within *ALX1*, where it was found that there was a high degree of similarity between species with blunt beaks. This pattern of lowered diversity is often spread beyond just the single locus where selection is active. This is due to an effect deemed the selective sweep, which occur when areas close to a selected locus are preserved, often showing signals of selection (Barton, et al. 2013). In another case of introgression, *Arabidopsis arenosa* populations that had adapted to live in serpentine soils, which have high concentrations of heavy metals that most plants cannot withstand, showed evidence of selective sweeps around alleles that conferred traits for survival within the inhospitable soils. Through statistical modeling, it was determined that most likely source for these genes was from the closely related *Arabidopsis lyrata* (Arnold, et al. 2016). Without the introgressed genetic material from *A. lyrata*, it likely would take a much greater amount of time for the necessary mutations to arise *de novo* within the populations of *A. arenosa* leaving those populations unable to take advantage of the serpentine soil niche. Understanding how these selective sweeps impact the evolutionary history of populations can aid in understanding the importance of introgression in both evolutionary biology and with the manipulation of crop plants.

The Monkeyflower Radiation of California

One such radiation that has been studied by modern day evolutionary biologists is the Monkeyflower radiation of Southern California. Depending on the taxonomist questioned, there are up to 12 taxa present within the radiation (Chase, Stankowski and Streisfeld 2017) (Tulig and Nesom 2012). Amongst the many taxa present is *Mimulus puniceus*, which grows with red flowers in populations around San Diego, CA and with yellow flowers to the east and the mountains (Figure 1) (Streisfeld, Young and Sobel 2013). These are called the red and yellow ecotypes as the two have distinguished areas of growth and have different pollinators (Streisfeld and Kohn 2007). Determining the cause of separation between these ecotypes has been a primary research topic for evolutionary biologists that study this radiation. Recently, genetic analysis has uncovered a section of interest within the genome of the *M. puniceus* that confers the different flower colors. As the two ecotypes of *M. puniceus* are most closely related to one another, it is expected that there will be a high degree of similarity when comparing the two genomes, which holds true throughout most of their genome. However, there is one section of the genome around the *MaMyb2* gene that shows high differentiation between the red and yellow ecotypes (Streisfeld, Young and Sobel 2013).

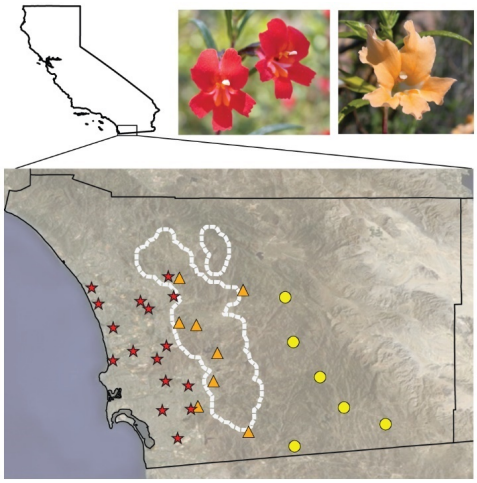


Figure 1. Map of populations from the red ecotype (red stars), the yellow ecotype (yellow circles), and their natural hybrids (orange triangles) within the hybrid zone denoted by the white circle. Images show representative pictures of the red ecotpye (left) and yellow ecotype (right). Taken as figure 1 from Streisfeld, Young and Sobel (2013). https://doi.org/10.1371/journal.pgen.1003385.g001

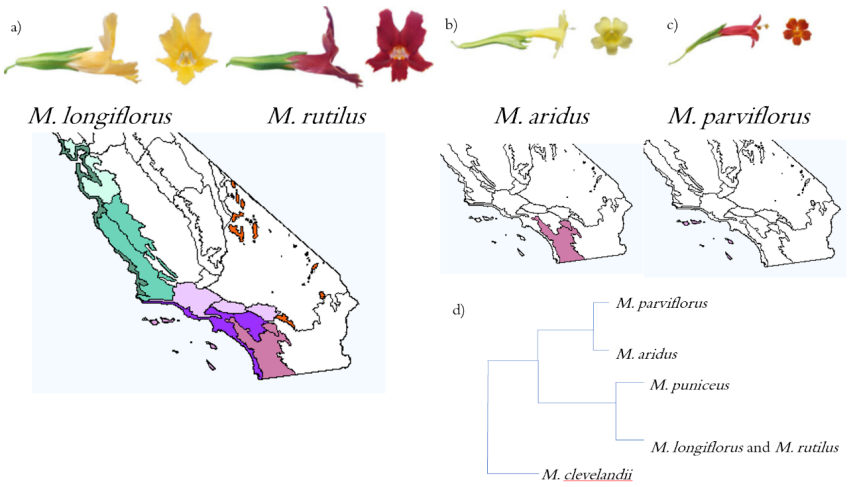


Figure 2. a) Profile and range maps for M. longiflorus and M. rutilus. b) profile and range map for M. aridus. c) profile and rang map for M. parviflorus. d) Phylogenetic relationships for the taxa discussed in this thesis, lines do not represent divergence and clades not discussed within this thesis are excluded. Flower profiles and phylogenetic relationships are from figure 1 and 3 of Chase et al. (2017) respectively and range maps are taken from the Jepson Herbarium at the University of California, Berkeley (N. S. Fraga 2018).

Along with *M. puniceus,* this thesis will examine four other taxa within the mimulus radiation. The first is *Mimulus longiflorus* and it's red-flowered counterpart *Mimulus rutilus*. These two taxa grow in populations together through much of coastal southern California (figure 2a). It is important to note that there is discussion of whether *M. rutilus* should be considered its own taxon or a variant of *M. longiflorus* as independent *M. rutilus* populations have not yet been identified. *Mimulus aridus* is a yellow-flowered taxon that occurs in the mountainous terrain east of San Diego (figure 2b). *M. parviflorus* is a red-flowered taxon that grows on the islands off the coast of California (figure 2c). Finally, *Mimulus clevelandii* is a species outside of the *Mimulus* radiation that is often used as an outgroup taxon for phylogenetic mapping (figure 2d).

The *MaMyb2* gene is part of the pathway used for anthocyanin production, a compound that is necessary to produce red pigments and thus create red flowers. This

gene affirms a genetic difference in the flower color between the red and yellow ecotype, however, *M. aridus* grows with yellow flowers which complicates the evolutionary question of how the two taxa came to be in their current state. This sharing of genetic information in the *MaMyb2* region has been shown to have occurred through introgression, and the current literature marks *M. aridus* as the most likely donor taxa. For example, it is likely that the ancestors of both *M. puniceus* and *M. aridus* had some geographical overlap that allowed for interbreeding. During this interbreeding, the *MaMyb2* allele that gives the ability to create red flowers could have been transferred from the ancestor of *M. aridus* to the ancestor of *M. puniceus,* thus providing the needed genetic material for the red ecotype of *M. puniceus* to produce red flowers. However, this hypothesis creates only more questions for the mystery of why the red and yellow ecotypes of *M.* puniceus have geographically distinct colors around San Diego.

Stankowski and Streisfeld (2015) discuss that there are several possibilities for the introgression that has occurred around *MaMyb2*. The mostly likely is as discussed above, that the ancestor of *M. aridus* introduced the *MaMyb2* region that conferred red flowers to the ancestors of *M. puniceus*. However, Stankowski and Streisfeld (2015) did not include another red-flowered taxon that has the potential to play a role in the introgression at *MaMyb2*. *Mimulus rutilus* is the red-flowered variant of the yellow-flowered *Mimulus longiflorus* and is important to the understanding of the evolutionary history of the radiation. As *M. rutilus* is also a red-flowered taxon with a yellow-flowered close relative, understanding the relationship between the red ecotype and *M. rutilus* can give insights into how the development of red flowers has effected the evolution of these two taxa. Should the red flowers in *M. rutilus* also have been caused

by the presence of the *MaMyb2* gene then it is possible that *M. rutilus* is also a recipient of introgressed genes. However, there are not populations of *M. rutilus* that occur without the yellow-flowered *M. longiflorus.* Whereas all of the western populations of *M. puniceus* are of the red ecotype. The presence of selective sweeps at the *MaMyb2* locus in the red ecotype and the omnipresence of the red ecotype in *M. puniceus'* western range indicates that the red allele rose to fixation rather quickly. This opens the door to many questions, such as: how does the relationship between *M. rutilus and M. longiflorus* relate to the relationship between the red and yellow ecotypes of *M.* puniceus? And does *M. rutilus* show signs of being genetically similar to the red ecotype in the Myb2 locus which indicate that *M. rutilus* is also under the effect of the *MaMyb2 gene?*

The idea of selective sweeps can also be utilized in determining the bounds of introgression in this region. By applying genomic analysis, the statistical differences between the genetic code of different individuals can be determined. When considering the red ecotype, it is expected that it will be most genetically similar to the yellow ecotype. As discussed above, this pattern is broken at the *MaMyb2* locus, where the red ecotype is more genetically similar to the *M. aridus* than the yellow ecotype due to the introgression at play in this locus (Stankowski and Streisfeld 2015). The bounds of this introgressed region can be inferred by looking for the points along the genome where the red ecotype becomes most genetically similar to the yellow ecotype as opposed to the other taxa of interest. These bounds are affected by many of the naturally acting drivers of divergence such as random mutation and recombination which cause the current taxa to be more divergent from one another.

In this thesis, I will test the area surrounding the *MaMyb2* locus to determine: a) what are the bounds of introgression that maintain the pattern of the red ecotype being genetically distinct from the yellow ecotype? b) do the patterns of divergence between *M. rutilus and M. longiflorus* follow closely with those of the red and yellow ecotype? and c) which taxa show the highest level of similarity to the red ecotype around the *MaMyb2* locus? I hypothesize that there will be a somewhat large bounds of introgression, as the *Mimulus* radiation is 'young', it is likely that the introgressed region will not have experienced major disruptions from random mutations and recombination events. I also hypothesize that the data will show similar patterns of divergence when comparing *M. puniceus with M. rutilus* and *M. longiflorus.* Finally, I hypothesize that the red ecotype will be more genetically similar to *M. rutilus* than it is to *M. aridus* in the *MaMyb2* region.

Methods

The V1 and V2 assembly

Genomic sequencing had already occurred for 47 samples of the monkeyflower radiation prior to the beginning of this project that were used for genomic analysis. Of those 47, 28 were used for the bulk of this project, of which 8 were of the red ecotype of *Mimulus puniceus* (5 from San Diego and 3 from Orange County), 8 were of the yellow ecotype of *M. puniceus* (5 from San Diego and 4 from Orange County), 4 were of *Mimulus aridus*, 4 were of *Mimulus rutilus*, and 4 were of *Mimulus parviflorus.* There are two versions of the genome's assembly. The V1 assembly was used for the twisst

analysis, while the V2 assembly was used for the haplotype network analysis and the Dxy analysis.

Haplotype Networks

Haplotype networks are a tool for analysis that seeks to 'map' out the differences in the genome. It works by comparing haplotypes from each individual and determining the most parsimonious connections from each haplotype. A haplotype in this respect is a unique genetic sequence that is shared by one or more individuals. The purpose of this analysis is to determine the relationships amongst the taxa of interest and to determine the locations in which these patterns shift. In areas where the genetic sequence is shared, taxa will group together and often share the same haplotype which is represented by larger circles with the sections of the circles representing different haplotypes. The closer to the *MaMyb2* gene a window is, the more tightly the red taxa are expected to group together. As the windows move away from the gene, it is expected that the red ecotype and the yellow ecotype will become more similar, while the red ecotype and other red taxa will become more dissimilar. If the stated hypotheses are supported by the data, it is predicted for *M. rutilus* to group strongly with the red ecotype throughout the different haplotype networks. It is also predicted that within the 30 kilobase region tested there will not be any areas that the red ecotype and yellow ecotype group most closely together indicating a bound of introgression.

The first step in this process is to phase the genome, creating two haplotypes that represent each of the chromosomes from a sample. After phasing, the 30 kilobase (30,000 base pairs) region surrounding *MaMyb2* was extracted and separated into 2 kilobase sections. 2 kilobase windows were selected as 2 kb gives an adequate number of base

pairs to create a digestible signal that is accurate to its locality and not obscured by including too many base pairs. 2 kb is also close to the size of the gene (~1,500 base pairs). These sections were then processed through the modelling software pegas (Paradis 2010), using Rstudio to create the haplotype networks (Figures 2 and 3).

D_{xy} Analysis in pixy

D_{xy} is a measurement of the pairwise differences between two genome sequences. A high D_{xy} indicates a high level of divergence while a low value indicates a low level of divergence. The purpose of this analysis is twofold. First, D_{xy} gives a statistic that can be used to compare the divergence between 1 main taxon (the red ecotype) and multiple others along the genome to determine which taxon is most closely related to the red ecotype. Secondly, it can be used to infer the bounds of introgression in a similar manner to the haplotype networks with a much finer scale. When looking at the D_{xy} plots there is likely to be a position in which the series with the yellow ecotype and red ecotype switches patterns with the series of *M. aridus* and the red ecotype. This position would be indicative of a bound of introgression. Additionally, if *M. rutilus* is the most genetically similar taxa to the red ecotype, it is expected that the series showing their D_{xy} will be the lowest at the gene and surrounding it.

This D_{xy} analysis was done through pixy. Pixy is a command line software designed for the calculation of population genomics statistics such as D_{xy}, π, and F_{st}. The data inputted for this analysis was filtered prior to this analysis and was left unphased. The analysis was performed in sliding windows of 4 kilobases and slid by 200 base pairs for each window such that each window overlapped with the previous one by 3,800 base pairs. This was done for comparisons of the red ecotype against the yellow ecotype, *M.*

parviflorus, M. rutilus and *M. aridus* (figure 5) and for comparisons of the red taxa, *M. rutilus* and the red ecotype populations of both San Diego and Orange county, and their yellow taxa counterparts, *M. longiflorus* and the yellow ecotype populations of both San Diego and Orange County (figure 6). In comparing the red ecotype against the other taxa, it can be seen where the similarities between the red ecotype and the red taxa begins and where that pattern is overtaken by the yellow ecotype and the red ecotype becoming more similar. It also should provide insight to which taxon is most similar to the red ecotype in the region. The prediction is that this will show *M. rutilus* as having he lowest D_{xy} values, especially within the gene. It also predicted that there will be significant portion of the 100kb region outside of the gene where the red ecotype shows a higher degree of similarity to the other red taxa than to the yellow ecotype. In comparing the red taxa with their counterpart yellow taxa, the data will give insight to the divergence of *M. rutilus* from *M. longiflorus* compared to that of the red from the yellow ecotype.

Twisst

The phylogenetic tree is a classic tool for understanding the relationships between taxa. It operates on a similar basis as the haplotype networks as it seeks to provide the most parsimonious relationship between the involved taxa (figure 7b). Twisst is a software that uses this model to estimate the relationships between the taxa across the genome by taking windows of 50 base pairs and determining the likelihood that a specific topology of a phylogenetic tree is the most parsimonious within the window assigning it a value from 0 to 1 such that the likelihood of the three possible topologies equals 1. (Martin

and Van Belleghem, Exploring Evolutionary Relationships Across the Genome Using Topology Weighting 2017). The purpose of this analysis is to find areas within the *MaMyb2* region that show a high confidence in relating the taxa of interest (*M. rutilus, M, parviflorus, and M. aridus)* to the red ecotype to determine signature of possible introgression. As with the other two analyses this analysis will also help to identify the bounds of the introgressed region by determining where the red and yellow ecotype once again become the most similar. When viewing the results of this analysis, it is expected to see high weights for topologies that group the red ecotype and the taxa of interest around the gene. Moving away from the gene, these weights will likely fall, if there is a strong drop off that can aid in identifying a bound of introgression. We can also use the number of high confidence windows (points where the weight is greater than 0.9) to determine which taxa can be considered the most genetically related. If the twisst analysis supports the given hypothesis, there will be more high confidence sites in the *M. rutilus* topologies compared to the *M. parviflorus* and *M. aridus topologies.*

For this analysis, 3 different relationships were evaluated. *M. rutilus, M. aridus* and *M. parviflorus* were all compared against both the red and yellow ecotypes from San Diego and Orange County using *M. clevelandii* as an outgroup. These relationships were evaluated as comparing the weightings between the three can show which among the taxa has the most consistent values of high confidence and help to illustrate the bounds of introgression along with the other analyses. The weights for topo2 (the phylogenetic tree which grouped the red ecotype and the taxa of interest together) were plotted along the 100kb region. Should the hypothesis of *M. rutilus* being the most similar to the red ecotype be supported by the data, it would be expected to see the most consistency in values of high confidence in the rutilus series.

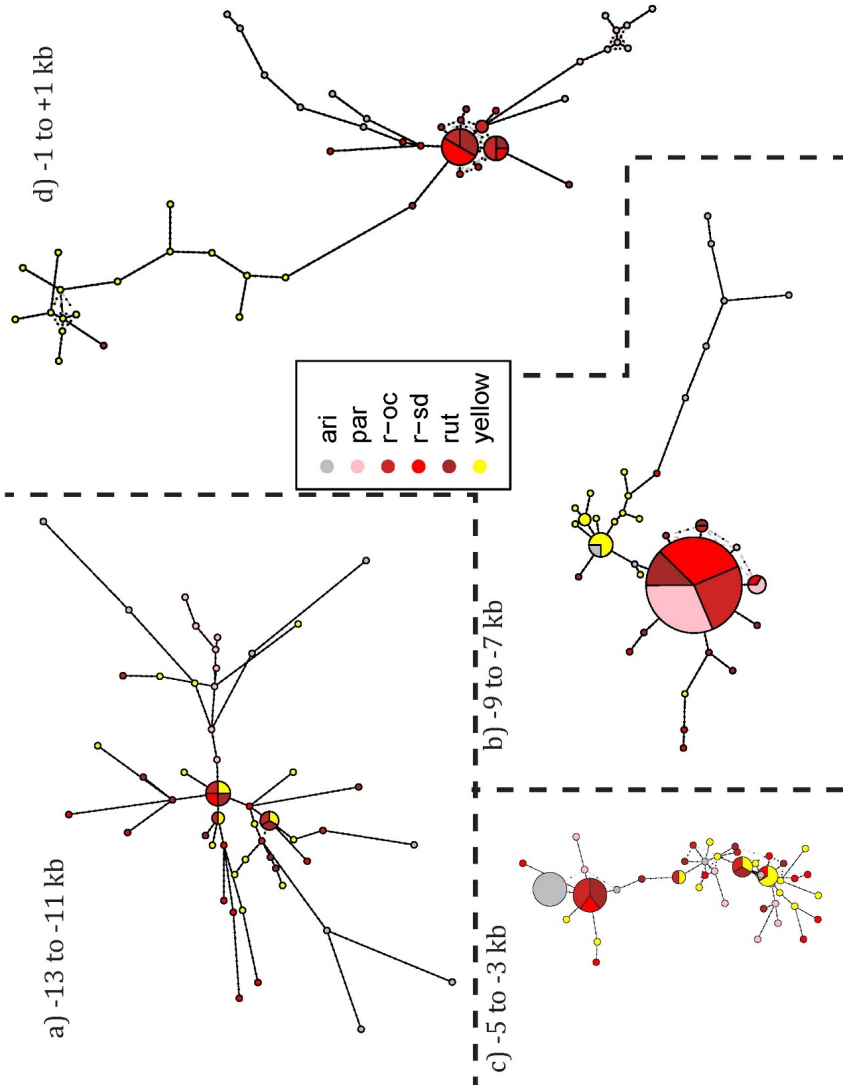


Figure 3. Upstream Haplotype Networks which were selected to be representative of the ranges discussed in the text body.

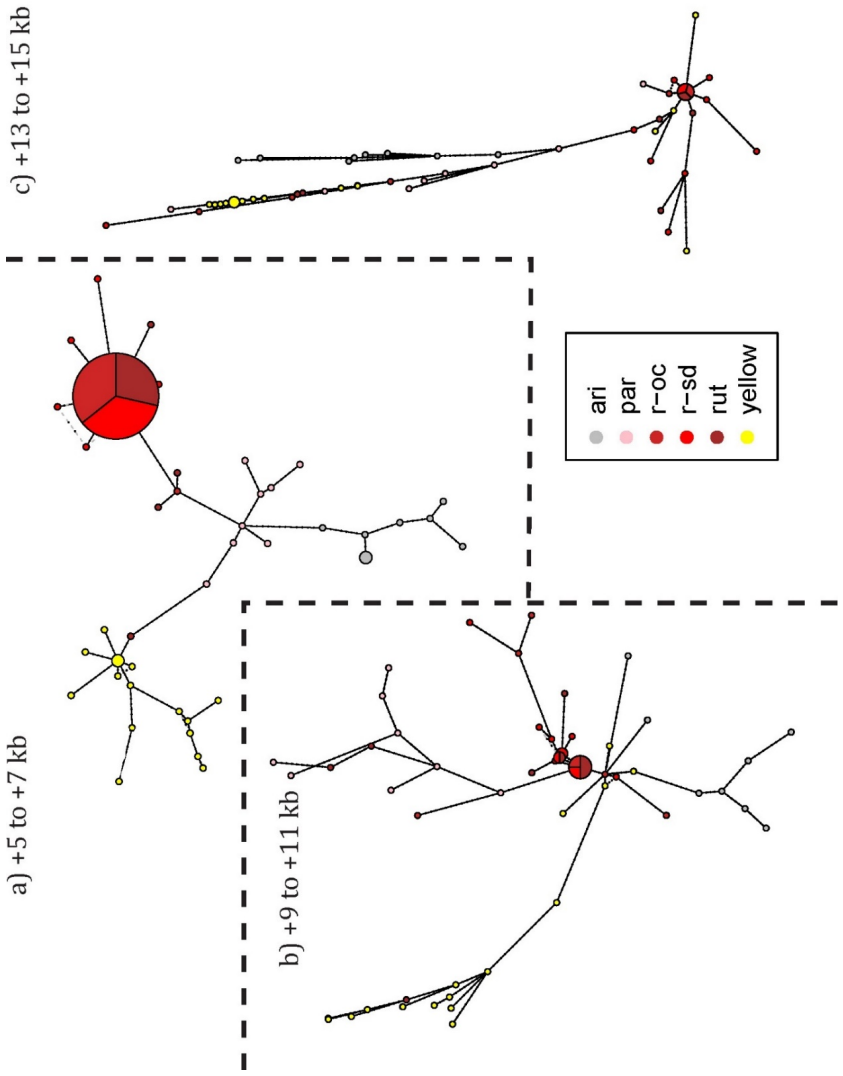


Figure 4. Downstream Haplotype Networks which were selected to be representative of the ranges discussed in the text body .

Results

Haplotype Networks

For all of the analyses in this thesis, the *MaMyb2* gene is centered at position 0, and base pair locations are then given as the relative upstream (-) or downstream (+) from the beginning of the gene. In the -15 kb to -11kb region there is little to no pattern with very little grouping of the taxa (figure 3a), where there should be either the red ecotype grouping with the other red taxa and *M. aridus* or the red ecotype grouping with the yellow ecotype. In the -11 kb to -7kb upstream region there is grouping of the red taxa (red ecotype from both Orange County and San Diego, *M. rutilus*, and *M. parviflorus*). In the -11 kb to -9 kb region, 2 haplotypes are shared between at least 4 individual chromosomes with individuals from *M. rutilus* and the red ecotype from both Orange County and San Diego. *M. parviflorus* groups closest to the other red taxa while the yellow ecotype and *M. aridus* group distantly from the red taxa. In the -9 kb to -7kb upstream there is one haplotype that contains multiple chromosomes from each of the red taxa. Meanwhile the yellow ecotype group together with one chromosome from *M. aridus* and the rest of *M. aridus* groups away from the rest of the samples (figure 3b). The -7kb to -1 kb region returns to having little to no pattern of groupings (figure 3c). In the 2kb region centered around the start of the *MaMyb2* gene, 2 haplotypes are shared between at least 4 individual chromosomes with individuals from *M. rutilus* and the red ecotype from both Orange County and San Diego. In this region *M. aridus* groups slightly closer to the red taxa than *M. parviflorus* with the yellow ecotype grouping furthest from the rest of the taxa (figure 3d). This pattern holds for the +1 kb to +9kb downstream region (figure 4a & 4b), though with fewer single haplotypes being shared

completely. From +9kb to +15kb, most taxa groups still hold together although there is considerable inter-taxa mixing (figure 4c).

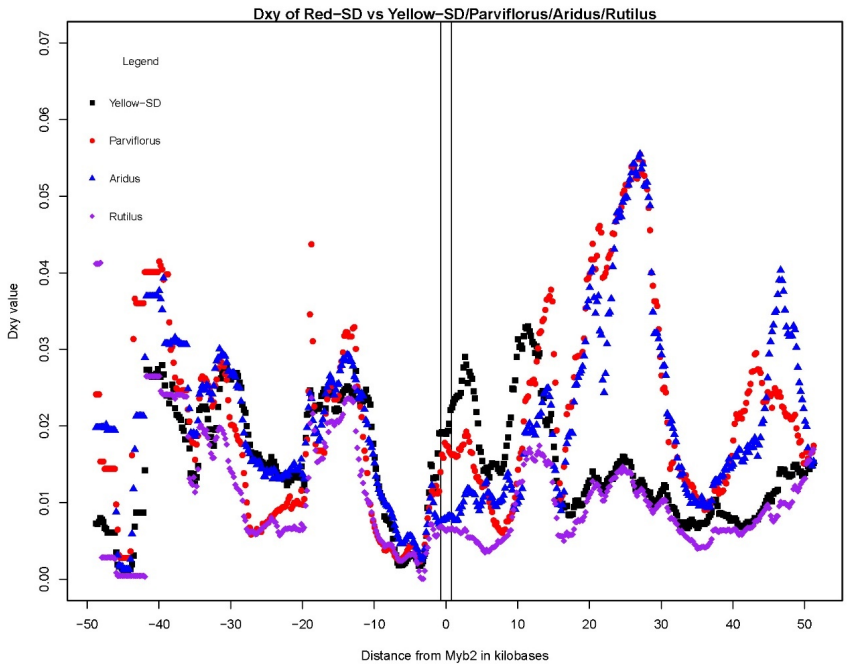


Figure 5. Dxy series for the red ecotype vs the yellow ecotype (black squares), M. parviflorus (red circles), M. aridus (blue triangles) and M. rutilus (purple diamonds).

D_{xy} Calculations

Two D_{xy} plots are included within this thesis. The first plots the D_{xy} values of the red ecotype when paired against the yellow ecotype, *M. parviflorus, M. aridus,* and *M. rutilus.* Upstream of the gene the D_{xy} values follow very similar patterns to one another.

Around -30 kb of *MaMyb2* there is an area where the *M. parviflorus* and *M. rutilus* D_{xy} values become somewhat lower than that of the yellow and *M. aridus* series. The 4 series then reconverge and show little deviance from one another until around -2 kb. Within the *MaMyb2* gene *M. rutilus* shows the lowest D_{xy} with *M. aridus* showing a slightly higher value. *M. parviflorus* is next lowest, and finally the yellow ecotype has the highest D_{xy} in this area. At about +15kb past the gene, the 4 D_{xy} series come back to similar values, after which the *M. parviflorus* and *M. aridus* series both spike significantly, while the yellow ecotype and *M. rutilus* series maintain much lower levels and follow one another closely. This continues to around +35 kb where the 4 converge and then separate once again following the same pattern with a smaller spike.

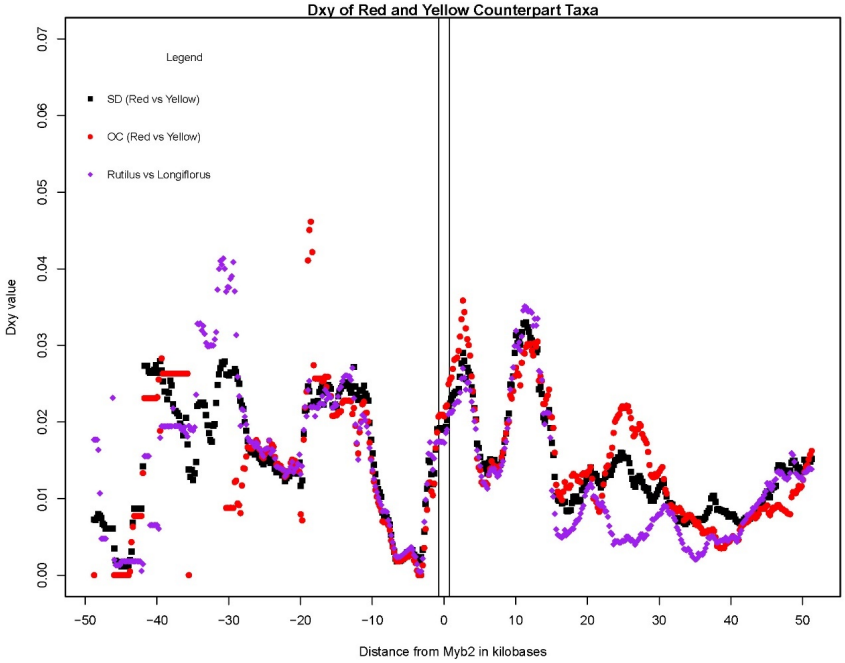
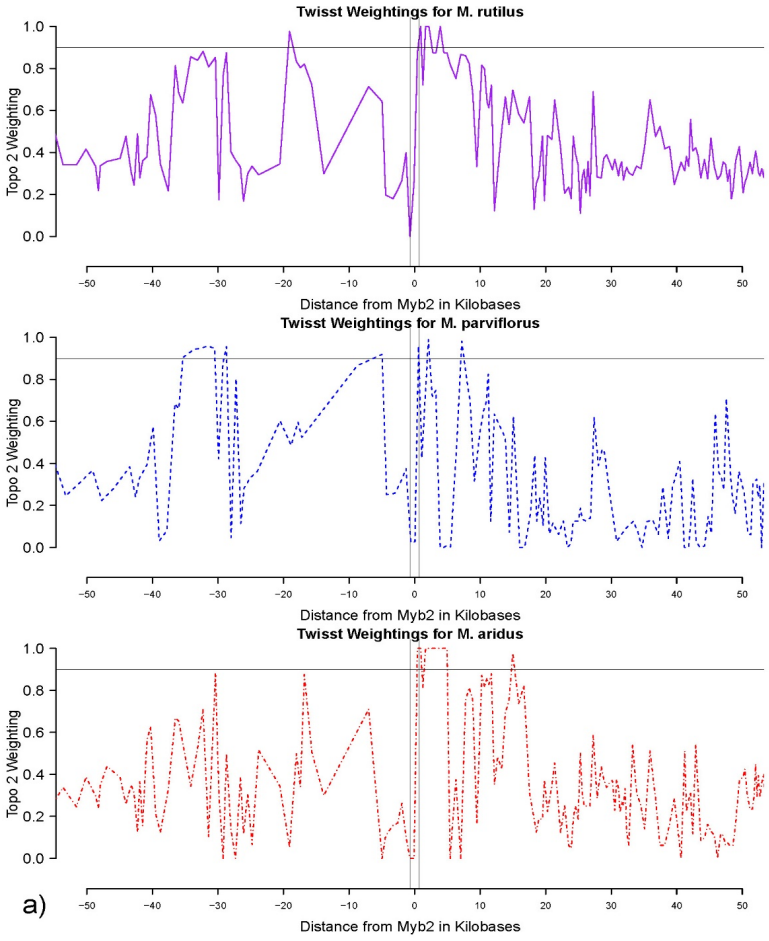


Figure 6. Dxy series of the red and yellow ecotype from San Diego (black squares) and from Orange County (Red Circles) and M. Rutilus vs M. longiflorus (purple diamonds)

The second plot shows the D_{xy} series for some of the red taxa and their yellow counterparts (Red vs yellow *M. puniceus* from both San Diego and Orange County and, *M. rutilus* vs *M. longiflorus*). Aside from some scattered points upstream of *MaMyb2* the three series track each other closely throughout most of the 100 kb region. The one notable place where this break is in the 20 kb to 40 kb region downstream from *MaMyb2. M. rutilus and M. longiflorus* show a dip in D_{xy} while the Orange County *M. puniceus* series spikes and the San Diego *M. puniceus* remains somewhat stable between the two other series.

Twisst Analysis

The Twisst analysis is presented in one figure (figure 7). The plot shows the weightings for topo 2 for each of the pairings. Topo 2 for each pairing is the one that pairs the red ecotype of *M. puniceus* together with the taxa of interest, leaving the yellow ecotype as the next step outside of the pair (Figure 7b). All three series show similar patterns to one another with slight variations. Between -40 and -30 both *M. parviflorus* and *M. rutilus* show points of high confidence (values greater than 0.9 shown by the horizontal line) with *M. parviflorus* having a slightly more consistent peak in this area (figure 7a). All three taxa show a high confidence just after the gene between 0 and +5 kb. *M. aridus* shows the most consistency in this peak, with *M. rutilus* showing similar yet less consistent value, whilst *M. parviflorus* shows a steep decline that reaches 0 confidence around +5kb.



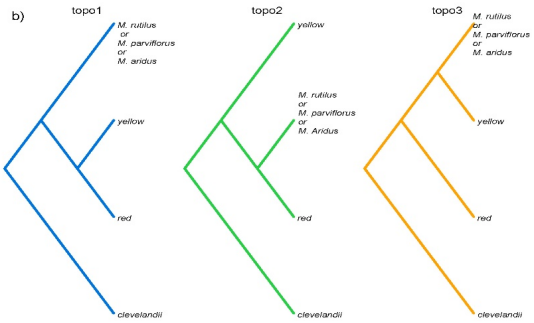


Figure 7. Twisst Results and Topologies: a) Weighting for topo2 of each of the taxa tested. Vertical bars at 0 indicate the MaMyb2 region and the horizontal bar marks the cut off for sites of high confidence (>0.9) b) Phylogentic topology possibilities for the Twisst analysis. Topo2 is the only one graphed in this thesis, however, the sum of the weight values in each window from the 3 topologies equals 1.

Discussion

Limitations of the Data

As can be seen in both the haplotype networks and in the D_{xy} plots, there is a significant amount of obscurity within the upstream region from *MaMyb2*. There is little separation of the taxa within some of the haplotype networks and considerable scattering and 'flat-lining' of data within the D_{xy} plots. This is likely due to issues within the V2 assembly and a high amount of insertion and deletions in the region (Streisfeld, Young and Sobel 2013) that has made genomic analysis challenging. It is important to use a critical eye when assessing the findings in this region to account for the high degree of uncertainty. Another point to consider is that the Twisst analysis was done using the V1 assembly. While both are important for understanding the genomic landscape of *MaMyb2*, comparisons cannot be readily made as the genomes were both sequenced and assembled separately.

Commented [KK1]: I used this reference because it discusses the in-del issue, briefly, not sure if that is good form or not

M. rutilus is more genetically similar to the Red Ecotype than *M. aridus*

Throughout the haplotype networks and the D_{xy} plots, the data signals that *M. rutilus* is more genetically similar to the red ecotype of *M. puniceus* than *M. aridus* is. In the haplotype networks, *M. rutilus* consistently groups closer to or with the red ecotype while *M. aridus* groups further away from the two. This pattern is maintained within the D_{xy} plots as well, where *M. rutilus* maintains relatively low levels of D_{xy} compared to the D_{xy} series that represent the other taxa. Within the gene region *M. rutilus* has the lowest D_{xy} for any of the taxa of interest. When considering the Twisst results, *M. rutilus* had

several points of high confidence but was slightly less consistent with these peaks compared to *M. aridus* which had a longer continuous area of high confidence directly downstream of the gene. It has been previously determined that the introgression at *MaMyb2* was likely sourced from *M. aridus* (Stankowski and Streisfeld 2015). While it is not likely that *M. rutilus* is the source for the *MaMyb2* locus in the red ecotype rather than *M. aridus*, the data suggests that there is a strong relationship between *M. rutilus* and the red ecotype in this area, which is as expected as the two both share red flowers which are made possible by the *MaMyb2* gene.

It is also of interest that the divergence between the red and yellow ecotypes, and the divergence between *M. rutilus* and *M. longiflorus* track so closely together, whilst *M. rutilus* and the red ecotype have the lowest D_{xy} value within the gene. This is a signature of likely introgression within the *M. rutilus MaMyb2* region. While it is not likely that introgression of these two taxa gave rise to the red ecotype's red flowers, it is possible that the two shared a common ancestor in which introgression occurred with *M. aridus*. Alternatively, it is possible that an ancestor of the red ecotype underwent introgression with the ancestor of *M. rutilus.* It has been argued, based on phenotypic data, that the yellow ecotype (called *Diplacus australis* in Tulig and Nesom (2012)) was a result of hybridization between *M. longiflorus* and the red ecotype. This was refuted on the basis of whole genome comparisons that failed to provide evidence for the claim (Chase, Stankowski and Streisfeld 2017). In this same paper, Chase et al (2017) make the recommendation that *M. rutilus* be considered a red-flowered variant of *M. longiflorus,* rather than as its own taxa as it is presented in Tulig and Nesom (2012). Based on the analysis done around the *MaMyb2* locus for this project, the evidence suggests that there

is similarity between the red ecotype and *M. rutilus* in the *MaMyb2* region which relates valuable information to the evolutionary history of the radiation. As more reliable genetic information becomes available about the areas surrounding *MaMyb2,* it will hopefully help in the classification of the taxa with the monkeyflower radiation.

The Genomic Landscape Around *MaMyb2*

From the analyses performed in this thesis, there are some sites of interest that warrant further study to determine whether the site is genetically significant. The first site is the 9 to 7 kb region upstream of *MaMyb2*. In this region all of the red taxa (both San Diego and Orange County of the red *M. puniceus, M. rutilus,* and *M. parviflorus*) all group together under one haplotype shared by many chromosomes. This pattern of all the red taxa being represented by one haplotype holds the potential of representing a functional change that is involved in the anthocyanin pathway that confers the ability to create red flowers. *MaMyb2* itself is a transcription factor that acts upon 3 different enzymes to enable the anthocyanin pathway that creates red pigments within the flowers (Streisfeld, Young and Sobel 2013). Two possible explanations are apparent for why this region has such high grouping. The first is that it is a pattern created from the insertion-deletion errors that exist in this area. The second, which would require further vetting with less contentious data is that the region does hold a functional mutation that acts in the anthocyanin pathway to create red flowers.

There are multiple reasons why the first option is much more likely based on the available data. If the site were to be a functional site that is vital to the creation of red

pigments, it is likely that the yellow flowered taxa would not be as genetically similar in this region as is seen in both the haplotype networks and in the D_{xy} calculations. In the -9 to -7 region the haplotype does show heavy grouping of the red taxa, however, it also shows the yellow ecotype and *M. aridus* grouping very close to the red taxa with about 6 point mutations separating them. This is compared to the approximately 36 point mutations that separate the red taxa and the yellow ecotype in the 2kb region centered around the start of the *MaMyb2* gene. In addition to this questionably close grouping, when looking at this region on the D_{xy} plot, it shows the continuation of a steep decline in the D_{xy} value across all the taxa comparisons which ends in a D_{xy} of almost 0 in the follow region. This strongly suggests that this site's grouping is due to errors within the assembly, rather than to a selected site that is involved with the production of red pigments.

Another site that exhibits notable behavior is in the +20 to +30 kb region downstream of *MaMyb2*. In this region the *M. aridus* and *M. parviflorus* series both spike, following each other closely. In the same area, *M. rutilus* and the yellow ecotype series remain low and track each other as well. One possible explanation for the pattern at this site is that it is a barrier locus. Barrier loci are sites that hold key genetic information to an individual's survival and thus act as a selective pressure against introgression in their located region and lead to further genetic isolation and speciation (Ravinet, et al. 2017). It is important to note that the presence of a barrier locus in this position is mostly speculative, however, there is some evidence to support it. First, as noted above the quick shift in pattern from the *MaMyb2* locus favoring *M. aridus* over the yellow ecotype to the reverse being true in this region with such a large spike in D_{xy}

for *M. aridus* is suspect. The *Mimulus* radiation is relatively young (Marques, Meier and Seehausen 2019) and so introgressed regions are expected to be longer and less broken up (Liang and Nielsen 2014). Additionally, the *M. rutilus vs M. longiflorus* series shows a lowered D_{xy} in this region that reflects a slight raise for the red ecotype vs *M. rutilus* series. While the presence of a barrier locus in this location remains mostly speculative, it is necessary that barrier loci will form across the genome of all the taxa within the *Mimulus* radiation as speciation progresses. In Heliconius butterflies it has been shown that barrier loci contribute significantly pre- and post- zygotic barriers of reproduction between closely related taxa (Martin, Davey, et al. 2019). For instance, as *M. parviflorus* has a range restricted only to the islands off the coast of California that it shares with *M. longiflorus* (Chase, Stankowski and Streisfeld 2017) it is almost certain that barrier loci have begun to form that restrict the gene flow between the two taxa.

The Bounds of Introgression

The bounds of introgression are difficult to discern from the results received. Upstream of the gene the issues with the assembly show little pattern, let alone a pattern of the yellow ecotype becoming more similar to the red ecotype compared to *M. aridus* that would indicate the upstream bounds of the introgressed region. The downstream region shows a bit more of a pattern with the yellow ecotype showing lower values of D_{xy} compared to the *M. aridus* series at around 14 kb. The peak of D_{xy} for the yellow and red ecotypes occurs just upstream of that at around 12 kb. This pattern is also demonstrated in the haplotype networks, which show that around this region samples

from the yellow ecotype begin to group closer to the red taxa and the red taxa begin to have samples that fall outside of the main cluster.

Interestingly, the series comparing the red ecotype and *M. rutilus* maintains a lower D_{xy} value compared to the red and yellow ecotype series all the way to the end of the region calculated. As stated before, this follows the pattern of *M. rutilus* being more genetically similar to the red ecotype within this region. It is possible that this could indicate a larger tract with introgression between the red ecotype and *M. rutilus*, however, more analysis would need to be completed to accurately assess the length of this stretch and how it compares to the tracts of the other taxa of interest. As more genomic analysis has been performed, better models have been created to assess things such as this. One method presented by Liang and Nielsen (2014) takes into account many of the classical models such as the hidden Markov model and the Wright-Fisher model. This method is better able to account for recent events and is able to predict multiple events that have occurred. Gaining a better understanding of the length of these tracts would increase the understanding of the evolutionary history of these taxa and help to demystify the challenges faced when performing genomic analysis is this region.

In conclusion, there is significant issue in the region upstream of the *MaMyb2* locus. While there is evidence to suggest that there may be an additional functional mutation in the 9 to 5 kb upstream area, there would need to be significant further study to show that this is not a false signal due to issues with the assembly. There is also an overall pattern throughout the haplotype networks, the D_{xy} plots and the twistt analysis, that *M. rutilus* is the more genetically similar to the red ecotype in the *MaMyb2* region than *M. aridus.* This is likely due to both the red ecotype and *M. rutilus* maintaining

their yellow flowers while *M. aridus* now has yellow flowers. Overall, these questions deserve to be addressed through further genomic sampling of populations in southern California and through controlled genetic experiments of the region.

Bibliography

Arnold, Brian J, Brett Lahner, Jeffrey M DaCosta, Caroline M Weisman, Jesse D Hollister, David E Salt, Kirsten Bomblies, and Levi Yant. 2016. "Borrowed alleles and convergence in serpentine adaption." *PNAS* 8320-8325. doi:doi.org/10.1073/pnas.1600405113.

Barton, N.H., A.M. Etheridge, J. Kelleher, and A. Veber. 2013. "Genetic hitchhiking in spatially extended populations." *Theoretical Population Biology* 87: 75-89. doi:doi.org/10.1016/j.tpb.2012.12.001.

Burgarella, Concetta, Adeline Barnaud, Ndjido Ardo Kane, Frederique Jankowski, Nora Scarcelli, Claire Billot, Yves Vigouroux, and Cecile Berthouly-Salazar. 2019. "Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation." *Frontiers Plant Science.* doi:doi.org/10.3389/fpls.2019.00004.

Chase, Madeline A, Sean Stankowski, and Matthew A Streisfeld. 2017. "Genomewide variation provides insight into evolutionary relationships in a monkeyflower species complex (Mimulus sect. Diplacus)." *American Journal of Botany* 104 (10): 1510-1521. doi:doi.org/10.3732/ajb.1700234.

Fraga, Naomi S. 2018. "Diplacus aridus, in Jepson Flora Project (eds.)." *Jepson eFlora, Revision 6.* Accessed May 24, 2021. https://ucjeps.berkeley.edu/eflora/eflora_display.php?tid=23076.

Fraga, Naomi S. 2018. "Diplacus parviflorus, in Jepson Flora Project (eds.)." *Jepson eFlora, Revision 6.* Accessed May 23, 2021. https://ucjeps.berkeley.edu/eflora/eflora_display.php?tid=23092.

Fraga, Naomi S. 2018. "Diplacus Longiflorus, in Jepson Flora Project." *Jepson eFlora, Revision 6.* Accessed May 24, 2021. https://ucjeps.berkeley.edu/eflora/eflora_display.php?tid=23090.

Harrison, Richard G, and Erica L Larson. 2014. "Hybridization, Introgression, and the Nature of Species Boundaries." *Journal of Heredity* 105 (Special Issue): 795-809. doi:10.1093/jhered/esu033.

Lamichhaney, Sangeet, Jonas Berglund, Markus Sallman Almen, Khurram Maqbool, Manfred Grabherr, Alvaro Martinez-Barrio, Marta Promerova, et al. 2015. "Evolution of Darwin's finches and their beaks revealed by genome sequencing." *Nature* (518): 371-374. doi:doi.org/10.1038/nature14181.

Liang, Mason, and Rasmus Nielsen. 2014. "The Lengths of Admixture Tracts." *Genetics* 197 (3): 953-967. doi:doi.org/10.1534/genetics.114.162362.

Losos, Jonathan B. 2010. "Adaptive Radiation, Ecological Opportunity, and Evolutionary Determinism: American Society of Naturalists E. O. Wilson Award Address." *The American Naturalist* (The University of Chicago Press for The American Society of Naturalists) 175 (6): 623-639.

Marques, David A, Joana I Meier, and Ole Seehausen. 2019. "A Combinatorial View on Speciation and Adaptive Radiation." *Trends in Ecology & Evolution* 34 (6): 531-544. doi:doi.org/10.1016/j.tree.2019.02.008.

Martin, Simon H, and Chris D Jiggins. 2017. "Interpreting the genomic landscape of introgression." *Current Opinion in Genetics & Development* 47: 60-74. doi:doi.org/10.1016/j.gde.2017.08.007.

Martin, Simon H, and Steven M Van Belleghem. 2017. "Exploring Evolutionary Relationships Across the Genome Using Topology Weighting." *Genetics* 206 (1): 429-438. doi:doi.org/10.1534/genetics.116.194720.

Martin, Simon H, John W Davey, Camilo Salazar, and Chris D Jiggins. 2019. "Recombination rate variation shapes barriers to introgression across butterfly genomes." *PLOS Biology* 17 (2): 1-28.

Nelson, Thomas C, and William A Cresko. 2018. "Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations." *Evolution Letters* 2 (1): 9-21. doi:doi.org/10.1002/evl3.37.

Paradis, E. 2010. "pegas: an R package for population genetis with an integrated-modular approach." *Bioinformatics* 419-420. doi:doi:10.1093/bioinformatics/btp696.

Ravinet, M, R Faria, R K Butlin, J Galindo, N Bierne, M Rafajlovic, M A F Noor, B Mehlig, and A M Westram. 2017. "Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow." *Journal of Evolutionary Biology* 1450-1477. doi:doi.org/10.1111/jeb.13047.

Stankowski, S, and MA Streisfeld. 2015. "Introgressive hybridization facilitates adaptive divergence during a recent radiation of monkeyflowers." *Proceedings of the Royal Society of London B: Biological Sciences* 2015. doi:10.1098/rspb.2015.1666.

Stephan, Wolfgang. 2010. "Genetic hitchhiking versus background selection: the controversy and its implications." *Philosophical Transactions of The Royal Society* 1245-1253.

Streisfeld, MA, and JR Kohn. 2007. "Environment and pollinator-mediated selection on parapatric floral races of Mimulus aurantiacus." *Journal of Evolutionary Biology* 20: 122-132.

Streisfeld, Matthew A, Wambui N Young, and James M. Sobel. 2013. "Divergent Selection Drives Genetic Differentiation in an R2R3-MYB Transcription Factor That Contributes to Incipient Speciation in Mimulus aurantiacus." *PLOS Genetics* 9 (3). doi:doi.org/10.1371/journal.pgen.1003385.

Suarex-Gonzalez, Adriana, Christian Lexer, and Quentin CB Cronk. 2018. "Adaptive Introgression: A Plant Perspective." *Biology Letters.* doi:10.1098/rsbl.2017.0688.

Tulig, Melissa C, and Guy L Nesom. 2012. "Taxonomic Overview of Diplacus Sect. Diplacus (Phrymaceae)." *Phytoneuron* 2012-2045.

Wang, Xutong, Liyang Chen, and Jianxin Ma. 2019. "Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication." *Genome Biology.* doi:10.1186/s13059-019-1631-5.