NATURAL HYBRIDIZATION AND INTROGRESSION BETWEEN *BERBERIS TRIFOLIOLATA* AND *BERBERIS SWASEYI* IN THE EDWARDS PLATEAU

by

Chloe Nicole Reeves, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major Population and Conservation Biology
May 2021

Committee Members:

Noland H. Martin, Chair

Chris Nice

Lauren Fuess

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

This work is dedicated to my loving father, Michael Reeves, who has always put his children first and has given me absolutely everything I could have ever needed and more. Thank you, Dad. I love you.

ACKNOWLEDGEMENTS

I would like to first thank my thesis committee for supporting me, encouraging me and empowering me throughout my time here at Texas State. I am eternally grateful for the employment I was given by the Texas State University Biology Department and the funding graciously given by the Bonner lab. I thank all members of the Nice, Martin, & Ott lab who have provided comradery and a sense of community. Others should also be mentioned, Alex Zalmat, Alex Sotola, David Suggs, Joan-Sayah Dixon, and Avery Mottet for help with sampling, computing, and lab work.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

ABSTRACT

Hybridization and hybrid zones can give us a means by which to discover the ecological and genomic interactions that occur between closely related species, providing an understanding of the mechanisms by which biodiversity arises and is maintained through reproductive isolation even in the face of introgression. Using next generation DNA sequencing techniques, we are now able to ask such questions at a genomic level. Here I used two species in the genus Berberis, Berberis trifoliolata and Berberis swaseyi to determine if hybridization is occurring, identify hybrid classes, examine the genomic architecture of introgression, and examine the asymmetry of introgression if any. Both species are native to the Edwards Plateau region of central Texas, and hybridization between the two species has been hypothesized because of morphological intermediacy. Admixture proportions were calculated in order to determine the degree to which hybridization is occurring as well as to describe substructure in the more widespread and common B. trifoliolata. Using the Admixture Class model in Entropy, it was found that the hybrids are either early hybrids (F1 and F2) or late generation hybrids that appear to have selfed instead of crossing with others. The Bayesian Genomic Cline model was used to quantify variation in introgression in hybrid individuals. The exceptional loci revealed that introgression tended to more readily occur from B. trifoliolata towards B. swaseyi with 305 exceptional loci introgressing towards B. swaseyi, and 229 introgressing towards B. trifoliolata. The results of all tests showed evidence of hybridization occurring with later generation hybrids largely absent, indicating that perhaps they were less fit.

Small amounts of introgression were detected such that loci from B. swaseyi largely appeared to have a selective advantage over those of B. trifoliolata. This study provides a starting point to ask more questions about this system and the effects of reproductive isolation in closely related plant species.

I. INTRODUCTION

The study of hybridization can provide insight into the degree to which genomic material can be exchanged between species as well as provide a richer understanding of the maintenance of reproductive isolation (Abbott et al. 2013; Harrison 1993; Payseur & Rieseberg 2016; Soltis & Soltis 2009; Whitham et al. 1999), thereby enabling further insight into how biodiversity arises and is maintained. The directionality of gene exchange and introgression between species (i.e. from one species to the other species) has been found in some cases to be influenced by the relative abundance of the parental species (Kron et Al 1993; Guichoux et al. 2013; Lepais et al. 2009; Sotola et al. 2019; Field et al. 2011). This can occur when species ranges shift based on environmental changes brought upon by natural or anthropogenic causes (Marie, Bernatchez, & Garant 2012; Yau & Taylor 2013). Such disproportionate introgression can influence the evolutionary path of the taxa involved in a number of ways including the creation of new hybrid species, reinforcement of reproductive isolation among the parental species, or eradication of one or both parental species through genetic swamping (Grant 1981). In addition, the phenomenon known as "introgressive hybridization" can occur where hybridization moves genetic material between parental populations while both populations remain on separate evolutionary trajectories throughout much of the genome. Such introgression is often largely neutral however, it is possible that some of the introgressed genetic material is selectively advantageous (Arnold & Martin 2009; Arnold et al. 2010; Martin et al. 2006). This potentially adaptive introgression can give rise to new, potentially selectively advantageous, genetic variation in one or both parental species (Anderson & Hubricht, 1938; Heiser 1951). Hybrid zones allow us to determine

the impact of hybridization by giving us examples of species potentially diverging in real time.

Hybrid zones occur first when two species meet and create F1 hybrids. If hybrids are able to be created, the fate of the hybrid zone is determined by the success of the F1 hybrids. If F1s are not able to be created, this is caused by reproductive isolation. Reproductive isolation results when two populations show reduced amounts of gene flow due to genomic divergence. This can also occur due to prezygotic factors such as phenology times not lining up preventing reproduction entirely. There are usually multiple reproductive barriers between species that affect total isolation (Dobzhansky 1937; Mayr 1942, 1947; Coyne 1992; Schluter 2001; Ramsey et al. 2003; Husband & Sabara 2004; Kay 2006; Martin & Willis 2007). In this study I am examining two species in the *Berberis* genus: specifically, *Berberis trifoliolata* and *Berberis swaseyi.* The amount of research that has been conducted with respect to hybridization and reproductive isolation in this study system, is very limited (Greathouse & Watkins 1938; Harms 2007) with no genetic research with respect to interspecific hybridization having been conducted to date. While putatively intermediate hybrid individuals have been identified in nature (Harms 2007), it is unknown the degree to which such hybrids are influencing gene flow between the two species where they co-occur in sympatry. For instance, while F1 hybrid production may occur at some appreciable frequency, if such early-generation hybrids are completely sterile, then little genomic introgression would be expected. However, if F1 hybrids are fertile and capable of producing later-generation hybrids, then a number of prezygotic and postzygotic barriers may be preventing the ultimate amalgamation of these two species and facilitate coexistence in sympatry.

Berberis trifoliolata ranges from Texas down into Mexico (Angulo et al. 2017), is palmately trifoliate (Figure 1), flowers in February and March, and fruits from May to July. *Berberis swaseyi* has a comparatively much smaller range endemic to the Edwards plateau of central Texas. *Berberis swaseyi* is pinnately compound with 2-5 pairs of leaflets and one terminal leaflet with the lowest leaflet pair being greatly reduced in size as compared to the other pairs (Figure 2). This species flowers on average later than *B. trifoliolata* in late February to April and fruits from May to July. Harms (2007) believed this to be the primary isolating barrier to hybridization. Putative hybrids of the two species are found to have 2 to 3 pairs of leaflets with a terminal leaflet; differing from *Berberis swaseyi*, the lowest leaflet pair is not smaller than the others (Figure 3). It is the case that in every location *Berberis swaseyi* is found, *B. trifoliolata* is found as well, often with putative hybrids. Hybridization between the two species has been hypothesized based on the morphology described above, however the extent of hybridization, geographically and genomically, is unknown.

Studies involving natural hybrids can reveal the extent to which taxa are reproductively isolated from one another, which genomic regions are favored, and which regions are resistant to introgression. Recent studies across a broad range of taxa reveal that many genomes are quite "porous," with much of the genome being susceptible to introgression through either neutral or adaptive processes, and only a few loci contributing to the reproductive isolation observed between hybridizing taxa (Wu 2001; Sung et al. 2018). While the genomic architecture of isolation and introgression has been thoroughly studied for a variety of taxa in a lab setting (Martin et al. 2005; 2006; 2007; 2008, Jiggins et al. 2008, Scriber 2017, Wu et al. 2020), fewer studies have used naturally

occurring hybrid zones for long-lived species (Fogelqvist et al. 2015, Sung et al. 2018). Naturally occurring hybrid zones are beneficial in that you do not have to prepare crosses and can simply take samples of existing individuals (Hewitt 1988; Buerkle & Lexer 2008).

Here, I use genotyping-by-sequencing (GBS) techniques on naturally occurring *B. trifoliolata, B. swaseyi,* and putatively hybrid individuals to examine the degree to which hybridization is occurring between these taxa and to understand the genomic relationship to reproductive isolation. A varying number of SNPs were generated for different data sets and used to answer the following questions: (i) is hybridization occurring between the two species?, (ii) what is the makeup of the hybrid classes (i.e, early-generation hybrids, or are hybrid zones comprised of multiple backcrossed individuals)?, and (iii) what is the genomic extent of introgression, and is there asymmetry associated with such introgression? To answer these questions, I analyze the admixture proportions of each individual to discover if hybridization is occurring and what admixture class each individual belongs to. Then, I will examine the rates of introgression of alleles across the genome to determine the extent to which introgression is occurring between the two species.

II. METHODS

Sampling, assembly, and sequence coverage

A total of 43 localities were sampled across the Edwards Plateau region of Central Texas (Figure 4; Table S1). Young leaves were collected from 735 individuals and immediately placed in clay kitty litter which served as a desiccant (Sung et al. 2018). Of the 735 individuals collected, 635 were morphologically identified in the field as *Berberis trifoliolata*, 67 *Berberis swaseyi,* and 31 were identified a-priori as putative hybrids.

DNA extraction and assembly

Genomic DNA was initially extracted from the young leaves of 735 individuals using a standard cetyltrimethyl ammonium bromide (CTAB) DNA extraction protocol (Doyle 1991). This resulting extracted product was dark in color and likely rich in plant secondary metabolites, specifically berberine. Berberine has been shown to bind to and damage DNA (Gu et al. 2015, Hou et al. 2017). PCR was inhibited in preliminary experiments. To remove the metabolites from the DNA, Qiagen QIAquik Gel Extraction Kits (Qiagen) were used to remove the plant secondary metabolites using standard protocols excluding steps 1-4 and 9. The purpose of the removed steps is to remove agarose gel from DNA samples if it is present. Given that there was no gel present in the samples we did not perform those steps. Once the DNA samples were further cleaned using the extraction kits, a reduced-complexity genomic library was generated (Meyer & Kircher, 2010; Gompert et al., 2012; Parchman et al., 2012; Mandeville, Parchman, McDonald, & Buerkle, 2015). Six μL of each individual's cleaned DNA was digested using restriction enzymes EcoRI and MseI. The fragments were labeled by ligating the

ILLUMINA adapter and unique individual 8-10 base pair adapter oligonucleotides or barcodes for later identification of sequences. PCR was run in two separate rounds on the restriction-ligation products using ILLUMINA primers. Then, the PCR products were pooled and sent to the University of Texas Genomic Sequencing and Analysis Facility (Austin, TX, USA) and size selection between 250 and 350 base pairs was performed via a Blue Pippen Prep quantitative electrophoresis unit (Sage Science, Beverly, MA). Once size selection was completed, the sample was sequenced over two lanes on an Illumina HiSeq 2500 SR 100 platform.

PhiX sequences were added by the sequencing site for quality control purposes. To filter out PhiX control sequences, all raw reads that assembled to the PhiX genome were removed. This was done using Bowtie-align-s version 1.2.2 (Langmead & Salzberg 2012). Custom Perl scripts were used to remove the reads that included the MseI adapters and barcodes from the single-end 100bp sequence reads. This script also matched the sample IDs with unique barcodes and corrected up to two single base sequencing mutations in those barcodes.

The library prep resulted in a number of failed reactions for individual samples, which in turn resulted in a large number of individuals with very few reads. This is likely due to the nature of this plant family to produce berberine, a secondary metabolite possibly used as anti-herbivore defense mechanism (Neag et al. 2018), and the inability of Qiagen gel cleanup kits to completely remove these secondary metabolites, which I found interfered with PCR reactions. Although it was believed that the samples were cleared of the majority of those compounds, it is possible that some still remained and inhibited parts of the PCR in the library prep or parts of the bridge amplification process.

This would cause that individual to have low coverage because of failed or inefficient reactions. To account for this, the reads were filtered to only include individuals who had 100,000 reads or more. After removing individuals with low read numbers, the total number of individuals in the data set was 186. The samples were distributed across all but four of the sampling sites as seen in supplemental Table 1.

In the absence of a reference genome for either *B. trifoliolata* or *B. swaseyi*, a *de novo* assembly was performed adapting part of the dDocent variant calling wrapper (Puritz, Hollenbeck, & Gold 2014). Parts of the wrapper that are for single end read data were used. Filtered reads were assembled using CD-hit (Li & Godzik 2006; Fu et al. 2012) with a threshold of 80% similarity. Consensus sequences from the *de novo* assembly formed the basis of a reference-based assembly of all reads from the 186 individuals. This was done using the aln and samse algorithms from BWA 0.7.13-r1126 (Li & Durbin 2009). BCFtools version 1.9 was used to call variant SNP sites. The -d parameter, maximum read depth per individual, was set to 8000 to reduce memory use. For subsequent analysis, one SNP was randomly chosen from contigs that contained more than one SNP. Custom PERL scripts were used in two rounds of filtering. Minimum coverage was set to 320, 2X the amount of individuals. Loci with minor allele frequencies less than 0.05 were excluded. The maximum number of individuals permitted to be missing data for any locus in order for the SNP to be retained, was set at 35. For the second round of filtering, all of the parameters were kept the same with the addition of a maximum coverage, which was set at 8344. This was calculated by taking the mean coverage and adding 2 standard deviations. Each individual was given a genotype likelihood estimate for each variable site instead of the genotypes being “called”. These

genotype likelihood estimates were used to calculate population allele frequencies. A total of 62,288 SNPs were identified across 186 individuals and used for population genomic analysis.

Population structure

To examine the genetic structure within and between the phenotypically divergent taxa *B. trifoliolata* and *B. swaseyi*, the program Entropy was used to estimate population genetic parameters (Gompert et al. 2014; Mandeville et al. 2015). Entropy is a hierarchical model in which an individual's admixture proportions of each of a given number of populations is estimated using a Bayesian framework. Although the output is visually and interpretatively similar to Structure (Pritchard, Stephens, & Donnelly 2000; Falush, Stephens, & Pritchard 2003), Entropy is able to account for sequence alignment errors, genotyping errors, variation in sequence coverage, and produces posterior genotype probability distributions by using prior probabilities from cluster allele frequencies (Gompert et al. 2014). Multiple models with different numbers of populations (k) ranging from 2 through 4 were compared, as well as different subsets of the data that will be described later. The process of finding a "best" k was not used, but I rather chose to show all results of k = 2- 4, as all models can hold pertinent information about population structure (Meirmans 2015; Sotola et al. 2019; Shastry et al. 2020). I did not show results past k4 because the model was no longer giving a biological explanation of variation. Two combined Markov Chain Monte Carlo (MCMC) simulations with 100,000 iterations sampling every 10^{th} iteration were used to calculate the posterior distributions of genotypes and admixture proportions. The first 5,000 iterations were discarded. To check the chain convergence, calculations of the Gelman-Rubin diagnostic

statistics and effective sample sizes were performed. Genotype probabilities and admixture proportions were combined across both chains of each model. The posterior distributions for parameters were summarized as means, medians, and 95% credible intervals. Different subsets of the samples were separately analyzed, restarting at variant calling, to answer further questions about the population structure of *B. trifoliolata*. The primary dataset analyzed included the entire sample set of 186 individuals (Figure 5), and an additional subset of the data that included only the "pure" *B. trifoliolata* individuals, N=135 (Figure 6) was analyzed to further examine fine-scale population structure within this species. A second subset removed all individuals from western localities, where only *B. trifoliolata* was found, in order to further examine structure across the phenotypically distinguishable species and explore interspecific hybridization, N=156 (Figure 7).

To additionally summarize the distribution of genomic variation in the samples, a principal components analysis (PCA) was performed using genotype probabilities that were estimated in Entropy. These estimates were used to generate a genetic covariance matrix, and this matrix was then used to create principal components scores which were plotted to illustrate relationships among individuals (Figure 6,9).

Identifying specific hybrid classes can provide information about the stability of a hybrid zone and the extent to which hybridization results in introgression via backcrossing. This was found by extracting both *q*1, which is equivalent to admixture proportion (or hybrid index) from the k2 run of 156 individuals, and Q12 which is the interspecific ancestry coefficient, or the proportion of the genome that is heterozygous for ancestry from each of the parental species. Both summary statistics for each individual were calculated in Entropy and individuals were plotted on a scatter plot. Both *q*1 and

Q12 measures range between 0 and 1. A *q*1 of 0 would be indicative of that individual being a pure of one species, in this case *B. swaseyi*, while a *q*1 of 1 would mean that individual is a pure of the other species, in this case *B. trifoliolata*. A Q12 of 0 would mean that such an individual is completely homozygous for ancestry, while a Q12 of 1 would mean they are completely heterozygous for ancestry. For example, with *q*1 as the x-axis and Q12 as the y-axis, F1 hybrids will be near (0.5,1.0) for *q*1 and Q12 respectively) and F2 hybrids will be near (0.5,0.5). If backcrossing occurs, the individual's *q1* would fall in between 0 and 0.5 or 0.5 and 1, but not on either 0, 0.5, or 1. If backcrossing has been occurring long-term, hybrid individuals would be found scattered along the triangles border of the plot as seen in (Gompert et al. 2014; Sotola et al. 2019). Individuals from eastern localities, where both species persist, were analyzed resulting in 156 individuals. The ancestry of each individual was determined with an admixture class model in Entropy (Q12; Gompert et al. 2014). Two independent MCMC analyses were run with 15000 iterations, sampled every 5th iteration after a burn-in of 5000 iterations.

Genomic variation of introgression

The Bayesian genomic cline model (BGC; Gompert & Buerkle 2011; 2012) is hierarchical and examines the probability of ancestry for all loci as a function of the distribution of the hybrid index of individuals. Ancestry probability ranges from 0 to 1 and were estimated using the maximum likelihood of each genotype at each locus for inheriting alleles from each parent species. An individual with an *h*=0 would only have ancestry from parent species 0 and an individual with *h*=1 would only have ancestry from parent species 1 (Buerkle 2005). The two locus specific parameters that were estimated

were α and β. The α parameter represents either an increase (+α) or decrease (-α) in the probability that a locus has increased ancestry from one group relative to their hybrid index. In this data set, positive measures of α represent cases in which there is excess ancestry of *B. swaseyi* at a locus, while negative values represent cases in which excess ancestry is observed for *B. trifoliolata.* Significant positive and negative α values could result from selection if specific alleles are advantageous in hybrids or even in the alternative genomic background. The β parameter shows an increase (+β) or decrease (-β) in the rate of change of the cline. Positive values indicate steeper clines and limited rates of introgression between the two species at that locus, and these loci have traditionally been interpreted as affecting reproductive isolation for their reduced rates of introgression (Sung et al., 2018). Negative values indicate wider clines with increased rates of bi-directional introgression, and these loci are interpreted as heterospecific loci that are more fit in both of the pure genomic backgrounds (Sung et al., 2018). New SNPs were called for this data set which resulted in 81,823 SNPs. The BGC model was used to quantify introgressive variation across the genome among individuals that were previously identified as admixed using the *q*1 parameter output from Entropy k2, with all individuals revealing any level of admixture above 0.01 being included (N= 20). The marginal posterior probability distributions for α and β were estimated with two independent MCMC simulations. Each was performed with 50,000 iterations, after a 25,000 iteration burn in and sampled every 5^{th} iteration. The convergence of the chains to the same stationary distributions was determined using the coda package in R, then the output from each chain was combined. Both the medians and 95% credible intervals were reported for α and β. Those loci whose 95% CI do not intersect zero were denoted as

exceptional loci and a binomial sign test was run with R package BSDA to ascertain whether the number of negative and positive α values differed significantly from each other, and thus whether asymmetric introgression was favored in either species.

III. RESULTS

Sampling, assembly, and sequence coverage

A total of 43 localities were sampled and 735 individuals were collected across all localities and species (Figure 4). Library preparation was unsuccessful for 549 individuals, revealed by the fact that those individuals had sequence coverage of less than 100,000 reads per individual. These individuals were removed from the dataset and not included in further analyses (Table S1). In all, library preparation was successful on 186 individuals: 135 previously identified, *B. trifoliolata,* 34 *B. swaseyi,* and 17 purported hybrid individuals with intermediate leaf morphologies. A total of 62,288 SNPs were found for this dataset. Based on preliminary results, 4 individuals were mis-identified in the field. This could have occurred as a simple labeling error or misplacement of the sample by the collector, and thus were removed from the data set. Substructure was found in *Berberis trifoliolata*. Western localities were found to be genetically different from eastern localities. *B. swaseyi* is not found in the western localities, so to both balance the sample sizes and generate a finer scale look at the picture of the hybrids, the western localities were removed from the data set. In the resulting 156 individuals, 50,925 SNPs were found.

Population structure

The Entropy program was run for the entire dataset that included all 186 individuals k2-4 and the posterior medians for each individual's *q*1 (admixture proportion) were summarized (Figure 5). The medians were plotted on separate bar graphs, one for each k, and individuals are organized by a-priori field identifications of species call and then by their admixture proportions. Each of these bar graphs give

information about the genetic structure of the populations. K2 shows a deep split between the two species *B. trifoliolata* in red and *B. swaseyi* in blue, as well as shows many potential hybrids (i.e., admixed individuals), many of which were pre-identified as such, at varying levels of admixture. These groupings in K2 align with a-priori species designations. K3 reveals further geographic substructure within *B. trifoliolata*, with eastern and western substructure appearing. This within-species sub-structure is explored further below. K4 does not separate out a new group, but rather multiple individuals show a small proportion of their assignment probability to a group that makes little biological sense. This suggests that no further structure exists.

A PCA utilizing the genotype probabilities of all 186 individuals was performed and plotted (Figure 6). PC1 explained 54.7% of the variation and split the individuals into three distinct groups, *Berberis trifoliolata*, *Berberis swaseyi,* and hybrids. The PC2 explained only 6.8% of the variation and this PC and additional PC axes with less explanatory variation while displaying variation in each species, did not result in further clustering among individuals. In this PCA as well as the previous set of Ks there are 4 individuals that appear to be misidentified potentially due to miss-labeling in the field. Those individuals were removed from the data set for subsequent Entropy analysis, resulting in 182 individuals.

Further sub-structure within *B. trifoliolata* individuals was explored by running Entropy on only the 135 pure *B. trifoliolata* individuals that were identified in the previous Entropy run included 93,974 SNPs. This revealed that there is indeed sub-structure with an eastern - western split (Figure 7). The admixture proportion for all individuals in each of their respective localities was averaged and plotted on a map. This

revealed that this subgroup was found in the western populations where *B. swaseyi* and hybrids between *B. trifoliolata* and *B. swaseyi* are not found (Figure 7).

To further explore interspecific hybridization between *B. trifoliolata* and *B. swaseyi*, Entropy was run on only 156 individuals from eastern localities where hybrid individuals were detected: 108 *B. trifoliolata*, 31 *B. swaseyi*,17 putative hybrids. K2 (Figure 8) reveals definitively that there are 8 hybrids with approximately 50% assignment probabilities to each species, which comports with the individuals being potentially early generation F1 and F2 hybrids. This was interpreted as such because there are a number of individuals in this figure that show equal admixture proportion to both species, as well as high inter-source ancestry. Twelve later-generation hybrids with hybrid index values of <0.2 and >0.8 were also identified, but no individuals with assignment probabilities that would bridge between the early and late-generation hybrids. K3 (Figure 8) revealed more sub-structure within *B. trifoliolata*. This appears to be a new split of genetically differentiated individuals. K4 (Figure 8) supports the presence of additional structure within *B. trifoliolata.*

A PCA (Figure 9) was plotted of the genotype probabilities of the 156 individuals, this included only individuals from the eastern localities. PC1 explained 83% of the variation and split the individuals into three distinct groups, *Berberis trifoliolata*, *Berberis swaseyi*, and hybrids. The PC2 explained only 1% of the variation and did not result in further clustering among individuals.

To determine hybrid classes of the individuals with mixed ancestry, we utilized the same 156 eastern individuals described above to be analyzed with the Admixture Class model in Entropy. The resulting figure (Figure 10) shows a cluster of hybrid

individuals at about (0.5, 0.8, *q*1 and Q12 respectively) and a single hybrid individual at (0.5, 0.4, *q*1 and Q12, respectively). I determined these to be F1s and a single F2 respectively. Conventionally an F1 individual would have *q* of 0.5 and a Q12 of 1. F2 individuals would also have *q*1 of 0.5, but a Q12 of 0.5, however variation of Q12 does occur (Shastry et al. 2020). These specific individuals not matching the Q12 expectation could be caused by the two species being closely related, and thus could have some overlap in their genetic makeup. There are also individuals on the x-axis, meaning there is no heterozygosity for ancestry, but they also do not assign completely to either species. This pattern could be explained by early generation hybrids repeatedly selfing until heterozygosity for ancestry is largely lost.

Genomic patterns of introgression

The Bayesian Genomic Cline Model was run on 81,823 loci and 20 individuals with intermediate assignment probabilities and all pure individuals of both species. All individuals with intermediate assignment probabilities were included. The hybrid index of the hybrids in this system ranged from 0.00003 to 0.99987 (Figure 11). The β parameter was less variable and ranged from only -0.25604 to 0.25201 (Figures 12 &13). The α parameter was variable and ranged among loci from -1.602 to 1.566 (Figures 14 &15). The model identified 534 (0.65%) loci with exceptional α values (95% CI did not include zero), and no loci with exceptional β values (Figure 12 & 13). Of the α loci, 229 showed patterns of positive introgression from *B. swaseyi* (+) to *B. trifoliolata* (-) (Figure 14) and 305 revealed patterns of positive introgression from *Berberis trifoliolata* (-) into *Berberis swaseyi* (+) genomic backgrounds (Figure 15). A binomial sign test found the number of positive α values was significantly larger than the number of negative α values

(binomial sign test, $p = 0.0011$) indicating that generally *B. trifoliolata* alleles are more favored than *B. swaseyi* alleles.

IV. DISCUSSION

Population structure

I examined population structure in the two native Texas evergreen shrubs *Berberis trifoliolata* and *Berberis swaseyi* to determine if hybridization is occurring, examined the population substructuring of *B. trifoliolata,* explored the genomic architecture of isolation and introgression, and determined whether asymmetric introgression was observed. Hybridization was found to be occurring between the two species as shown by the output from Entropy. It also revealed substructure within *B. trifoliolata*. This is expected as its range is more extensive than *B. swaseyi's* extending from Texas down into Mexico, while *B. swaseyi* is only found on the Edwards Plateau. This substructure in *B. trifoliolata* was also found to be based on geography as one group was found solely in the east and one, solely in the west with some admixture in the middle of the study area. The eastern group of *B. trifoliolata* was the only group that was found to be hybridizing with *B. swaseyi*, despite the fact that there was geographic overlap with the western group (Figure 7). There being geographical based population structure within *B. trifoliolata* over its broader range was previously confirmed by Angulo et al. (2017).

Of the hybrids identified with 50:50 assignment probabilities between both species, seven F1 individuals and a likely F2 individual were found. However, no early-generation backcrossed individuals were found, but some individuals who had undergone many generations of selfing as evidenced by the hybrid individuals largely found along 0.5 vertical line or along the x-axis (Figure 10). This could be due to the flowers not being fertilized by insects and subsequently the flowers self-fertilizing through the

process of autogamy (Eckert 2001). This can occur while a single flower withers, or the anthers of a single flower being triggered by something other than an insect and the pollen consequently falling into the flower (Angulo et al. 2014). Self-fertilization can also take place between different flowers on the same plant. This is known as geitonogamy (Eckert 2001). This can occur through wind dispersal of pollen to nearby flowers, or through a pollinator landing on multiple flowers from the same plant. This has been known to occur on *B. trifoliolata* with honeybees (Angulo et al. 2014). The lack of early-generation backcrosses is stark and is indicative of reproductive isolating barriers that are preventing F1 matings with pure-species parents. One potential reproductive isolating barrier preventing further backcrossing is the distinct phenological differences that exist between the species. However, Harms (2007) found the flowering time of the hybrids to be an intermediate of the two species, meaning their phenology should overlap both parent species. Overall, the admixture class model shows us that the hybrid zone is well established because the amount of time it would take for an individual to get its heterozygosity to zero by selfing alone, would take generations. Furthermore, when we include the knowledge that in every location *B. swaseyi* is found, so are *B. trifoliolata* and hybrids, it appears the two species do hybridize readily. Although hybrids appear to persist and are then fit enough to survive, the hybrids do not appear fit enough to continue reproducing with individuals of either pure parent species. This can occur when the parent species chromosomal complements combine to make a poorly fertile hybrid (Charlesworth 1995). There is evidence for this taking place with sunflower species, which allowed for hybrid speciation to occur (Rieseberg et al. 1995). Finding the mechanisms of the reproductive isolation of the hybrids would be a next step for this

system.

Genomic introgression

Five hundred and thirty-four exceptional α loci were identified. Of these loci, a significant amount were such that *B. trifoliolata* loci tended to introgress into *B.swaseyi* genomic backgrounds more than the reciprocal. This directional introgression could be because *B. swaseyi* is endemic to the Edwards Plateau region and therefore more specialized to this particular environment than the more generalist *B. trifoliolata.* This supports the idea of genetic swamping, where a species with more individuals or who is a generalist can genetically overtake the species with fewer individuals (Coleman 2014; Gibson 2019).

This hybrid zone is well established and likely went through times of increased hybridization rates and lower hybridization rates. This is evidenced by the fact that there were very late generation back-crossed and then selfed individuals found. However, there were no early generation back-crossed individuals found. This is likely caused by effective reproductive isolation possibly in the form of hybrid infertility. Because of this, the introgression found is likely not a conservation concern, though the rarity of *B. swaseyi* should still be considered for conservation decision-making.

This work will give a good starting place for more questions to be asked about this system. A better way of extracting DNA from these individuals can be developed and used to gain a larger sample size, and the genetic data found here can be combined with morphological data to have a greater understanding of how hybrid zones are formed and sustained, and if any of the genes that are shown to be introgressing are tied to any morphological characteristics. Since it was found that there is a subset of *B. trifoliolata*

that is separated out by Entropy early on, then a greater sample size and a specific of that species could allow for greater precision in data and ability to answer more early-stage questions about speciation. A closer look into the mechanisms of their reproductive isolation would be beneficial as well.

FIGURES

Figure 1: Example of palmately compound "pure" *Berberis trifoliolata* leaf. Photo by Avery Mottet.

Figure 2: Example of pinnately compound “pure” *Berberis swaseyi* leaf. Photo by Avery Mottet.

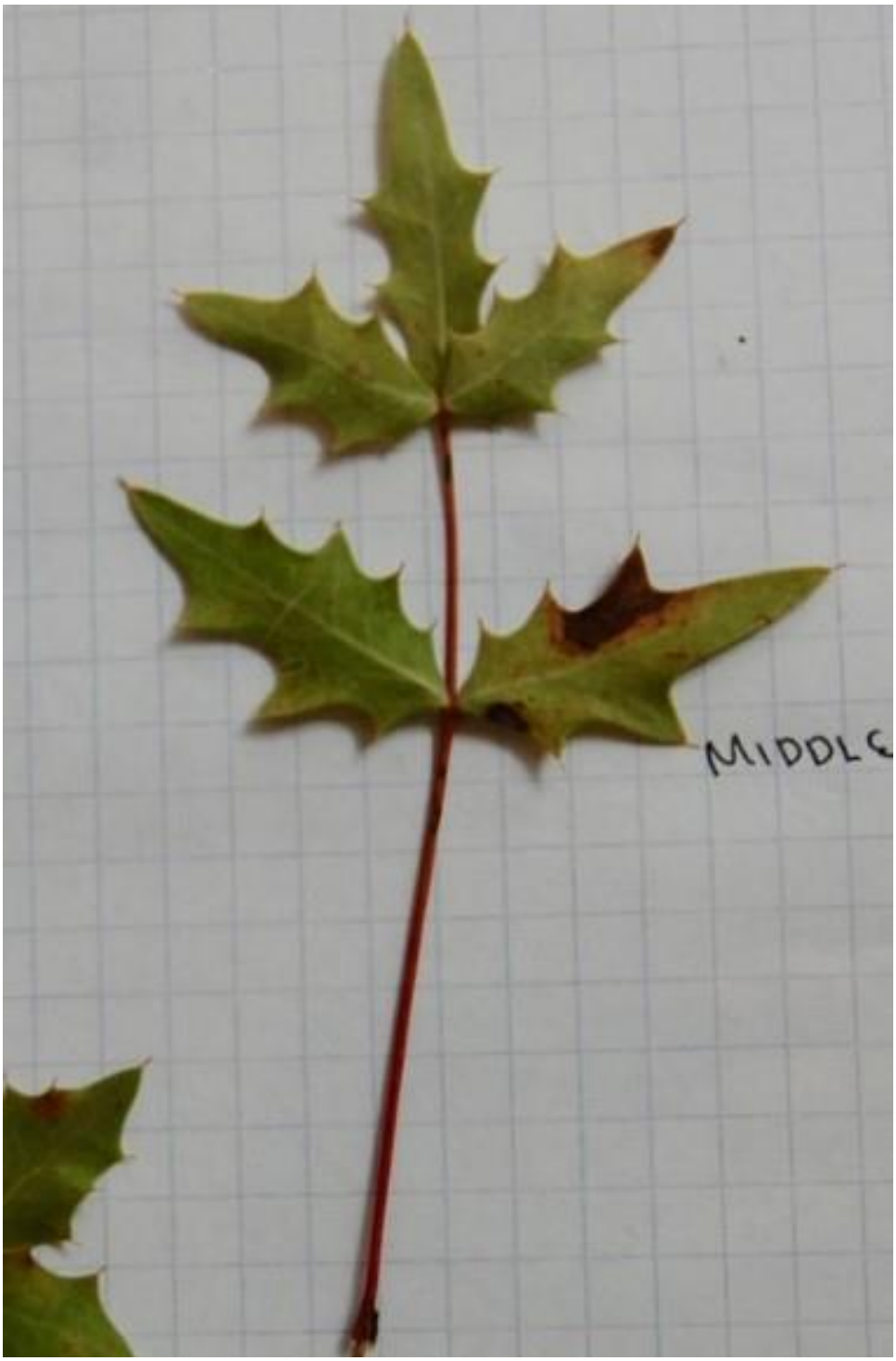

Figure 3: Example of putative *Berberis trifoliolata* x *Berberis swaseyi* hybrid leaf. Photo by Avery Mottet.

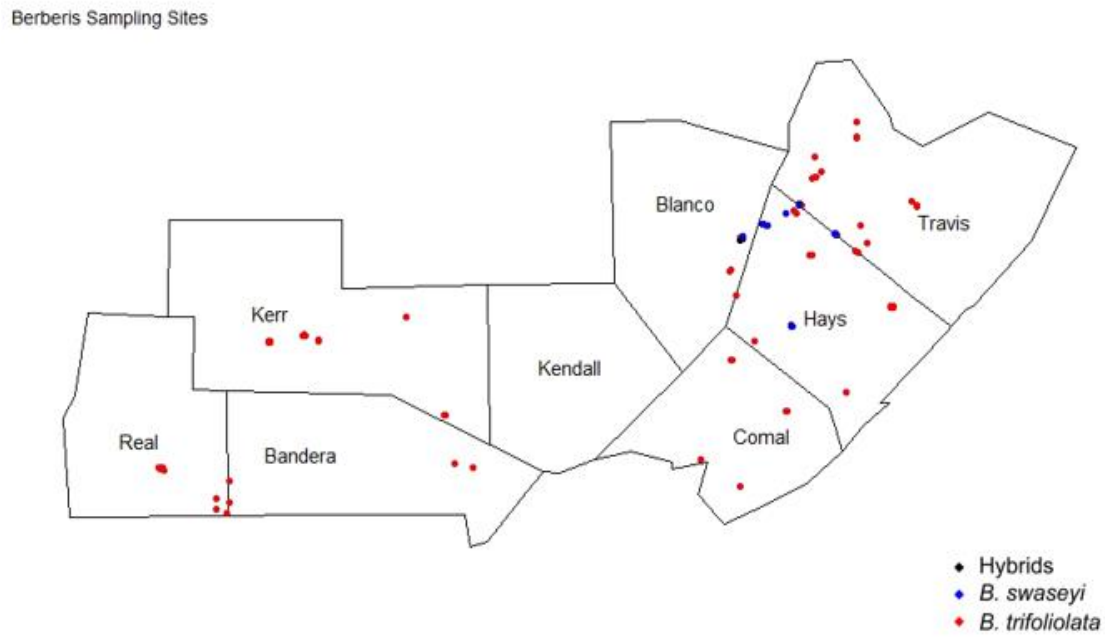

Figure 4: Sampling map for all 182 individuals of both *Berberis trifoliolata* and *Berberis swaseyi.* Blue represents *B. trifoliolata*, red represents *B. swaseyi*, and black represents hybrids.

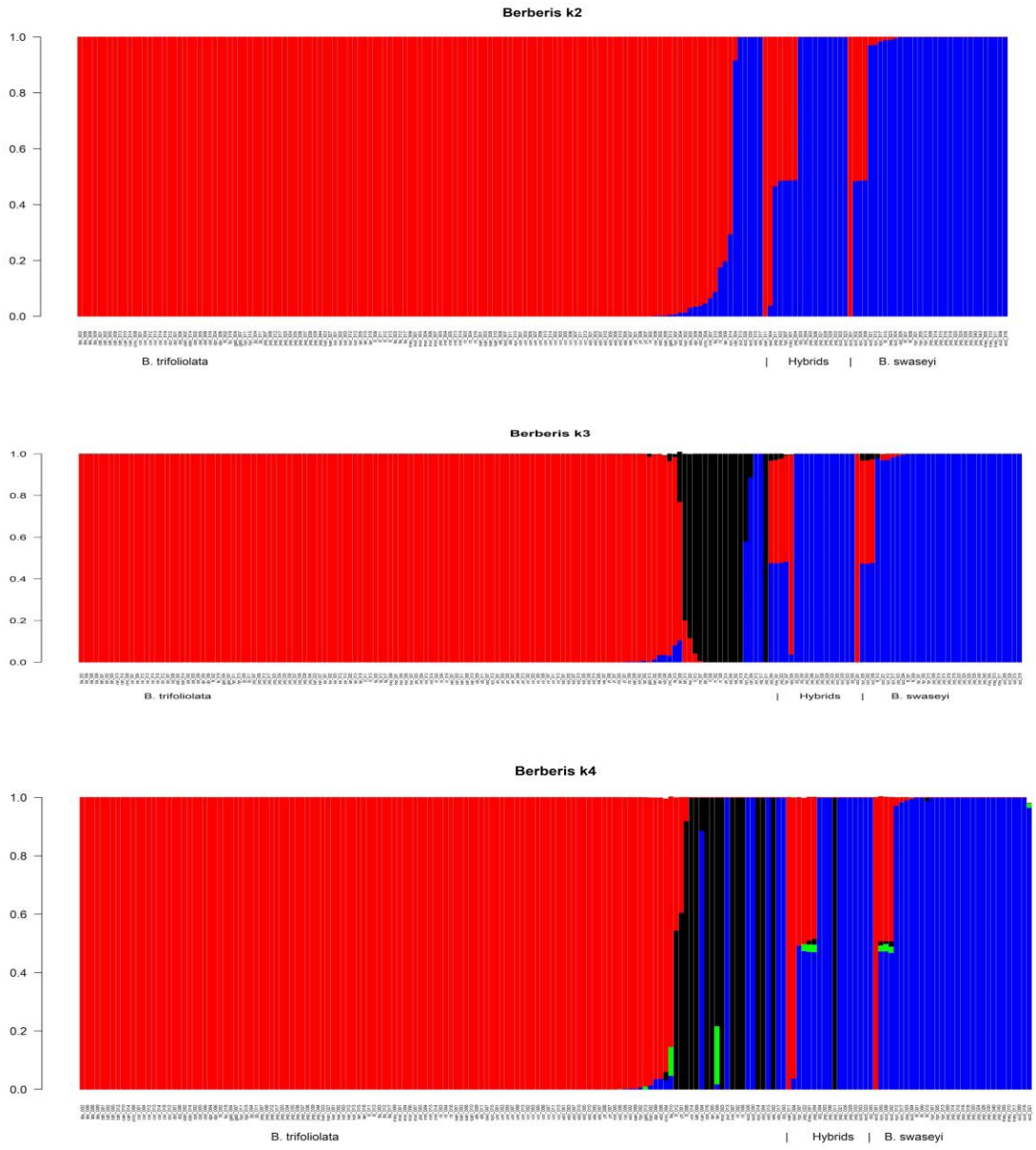

Figure 5: Entropy plots for k2-4 for the entire dataset that included all 186 *Berberis trifoliolata, Berberis swaseyi,* and purported hybrid individuals. Red bars represent *B. trifoliolata,* and Blue *B. swaseyi.* Black shows a subgroup of *B. trifoliolata* that appears in K = 3. The additional "green" group at K = 4 shows that Entropy has no more explanatory power. Ordered by species call, then q values. Hybrids are confirmed.

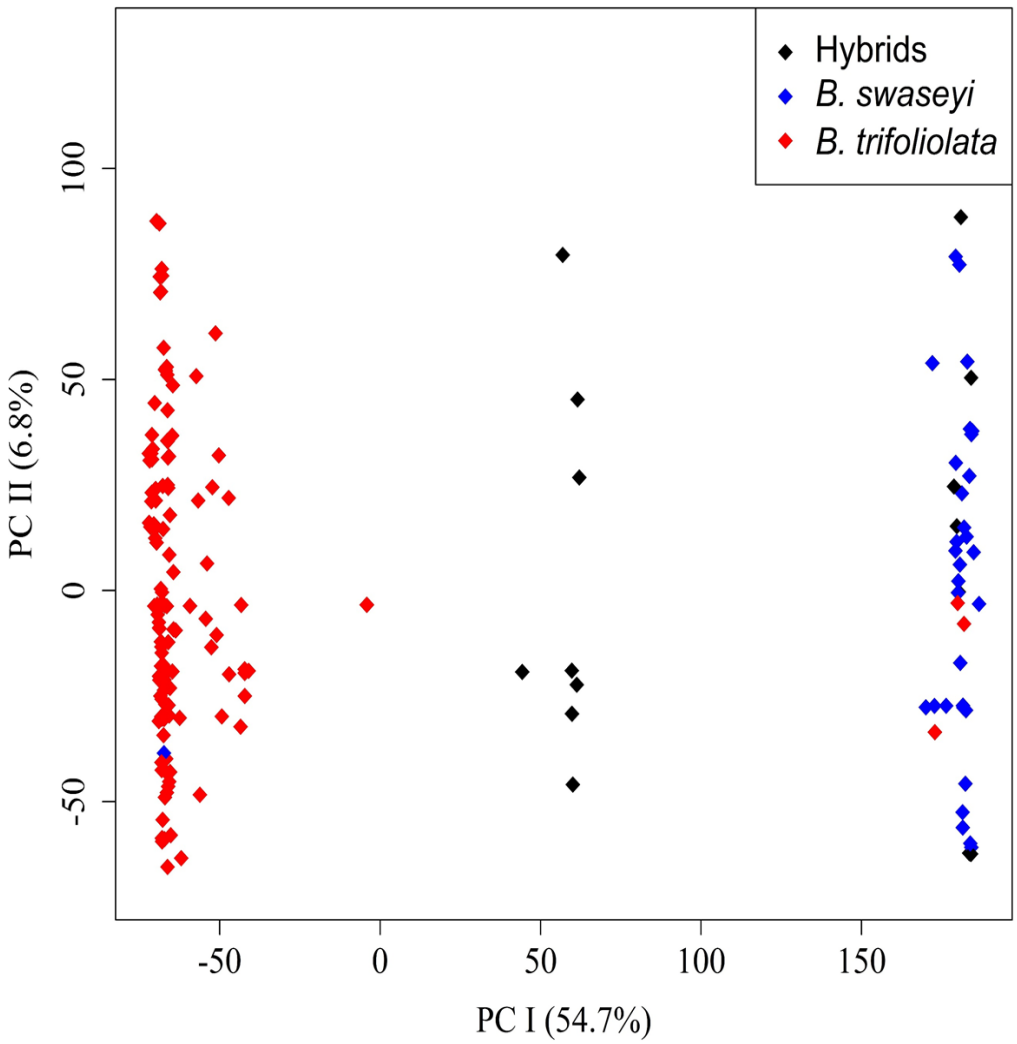

Figure 6: Principal components analysis (PCA) of genotype probabilities as calculated by Entropy. (N= 186) PC1 explains 54.7% of the variation in this dataset. PC2 explains 6.8% of the variation in this dataset.

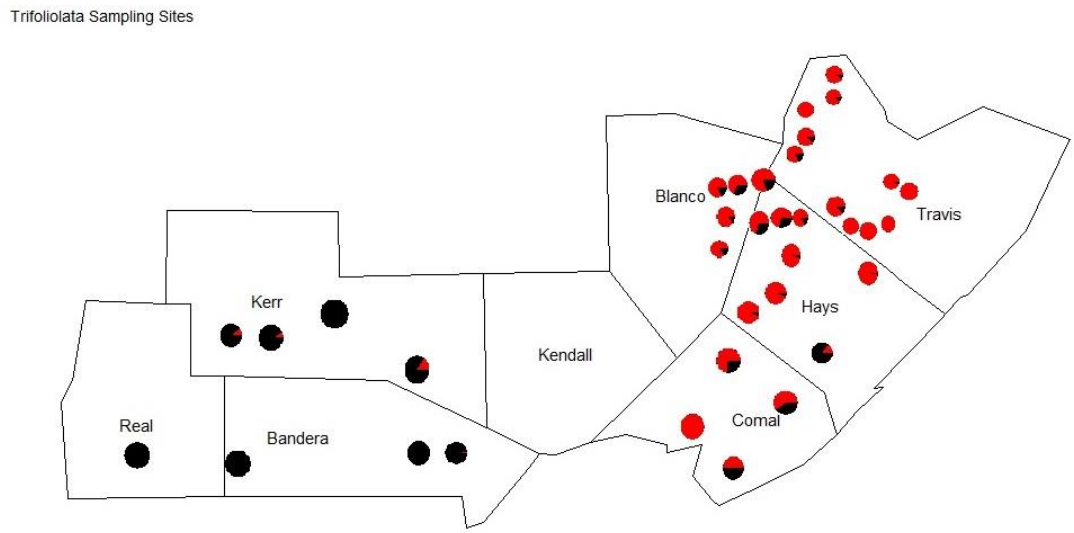

Figure 7: Entropy data for each *B. trifoliolata* individual at each locality was combined to make pie charts. Those pie charts were plotted on a map. This showed that the subgroup, black, was found in the west. Because no *B. swaseyi* were found in these sampling sites and further programs assumed input data only included two lineages, individuals with assignment probability of .2 or greater to the black group were removed.

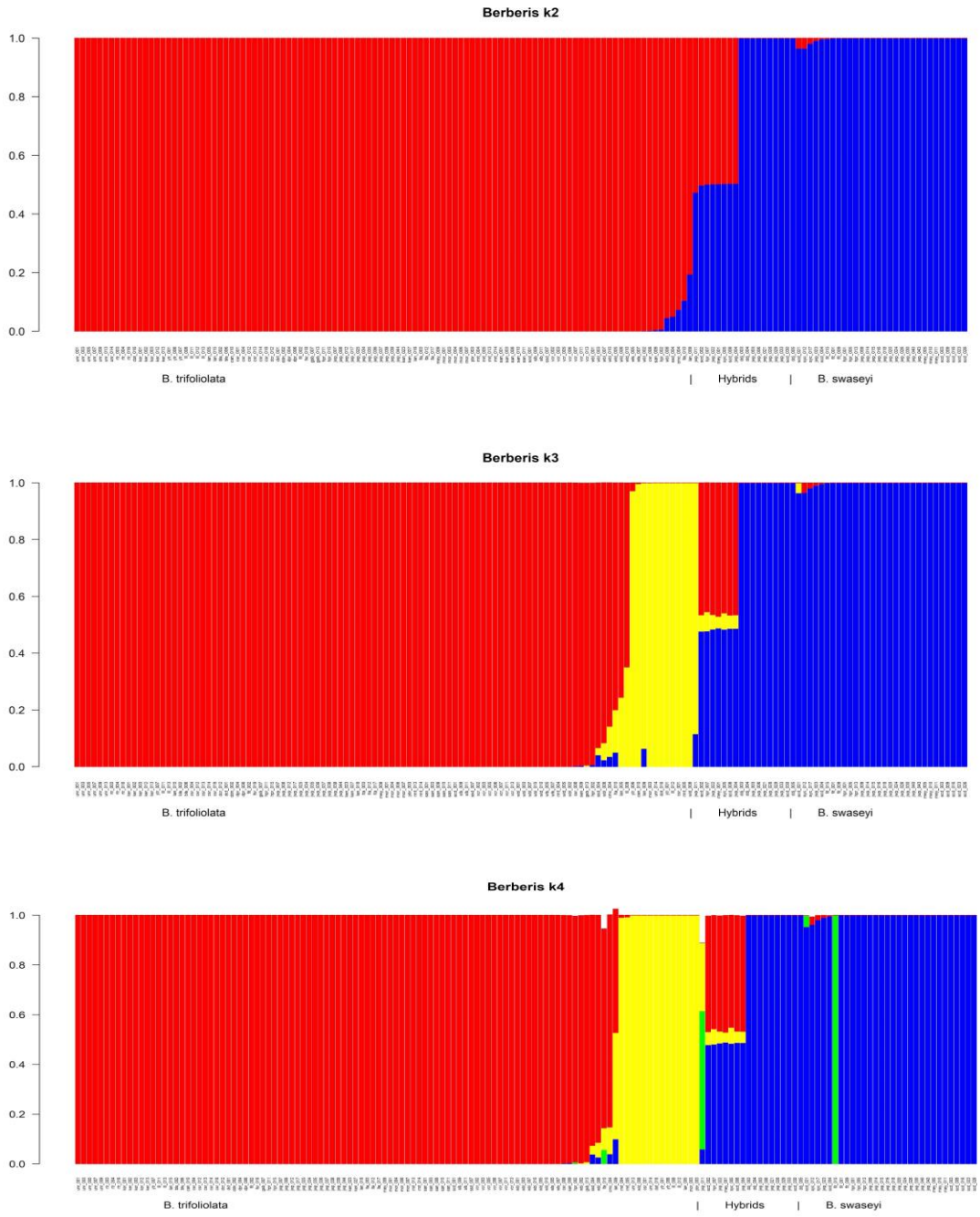

Figure 8: Entropy plots for 156 *Berberis trifoliolata, Berberis swaseyi,* and hybrid individuals for k2-4. Blue represents ancestry for *B. swaseyi* and red represents ancestry for *B. trifoliolata.* Yellow shows a new subset of *B. trifoliolata.* The model appears to break down in k4 based on the few individuals not summing to one. Arranged by species call, then q values.

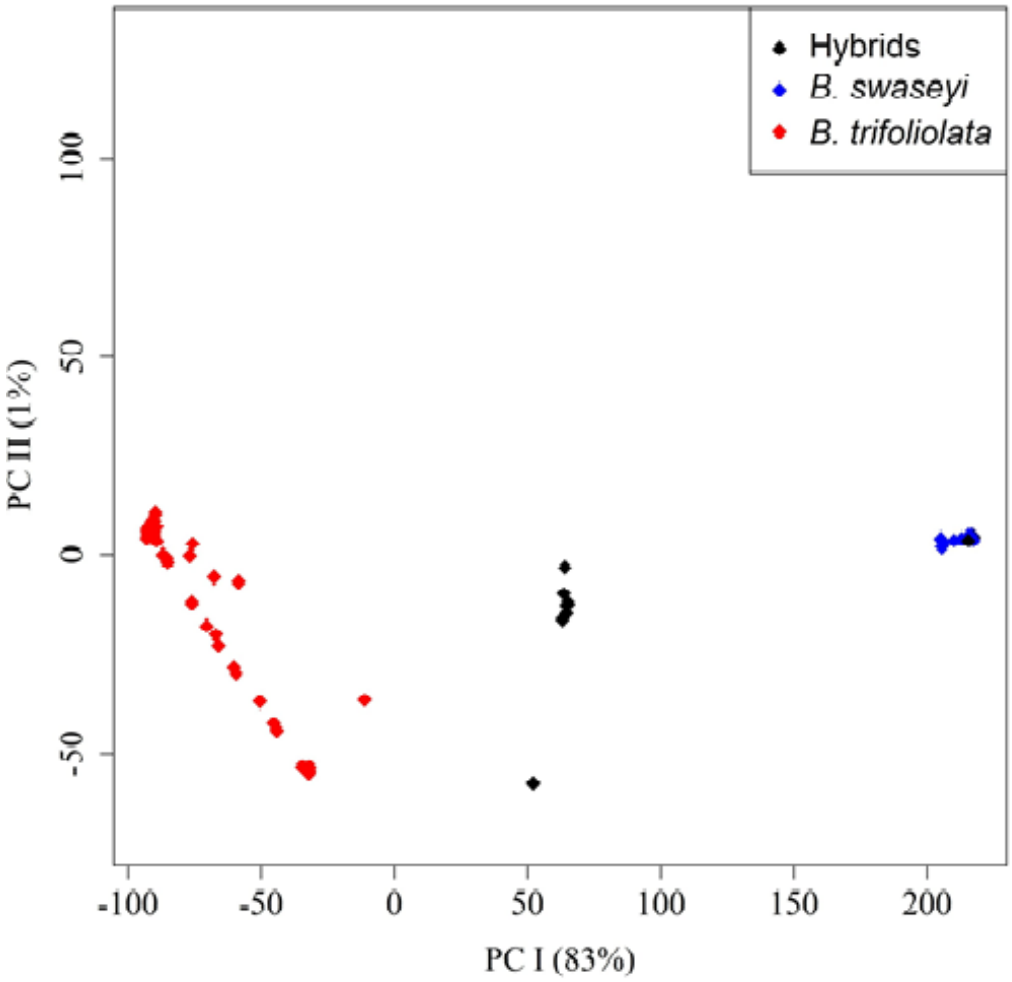

Figure 9: Principal components analysis (PCA) of genotype probabilities as calculated by Entropy. PC1 explains 83% of the variation in this dataset. PC2 explains 1% of the variation in this dataset. (N=156)

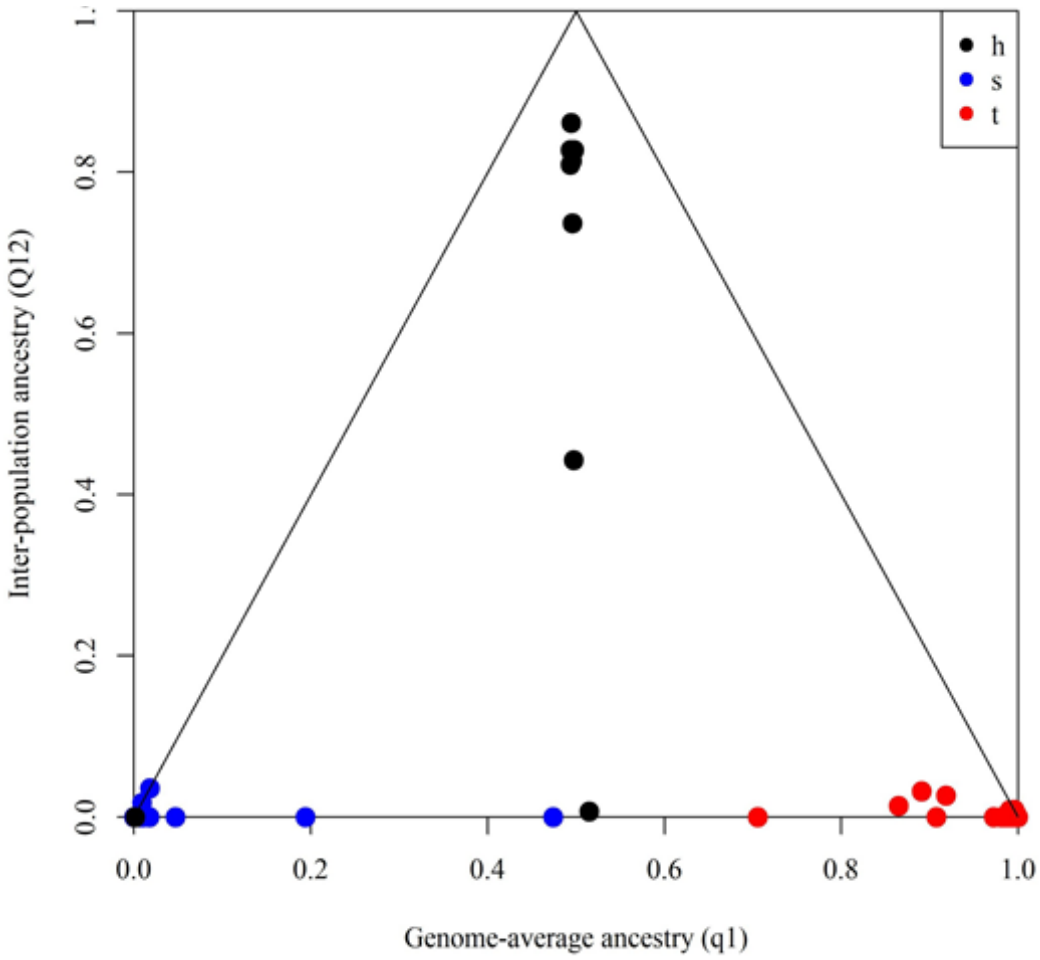

Figure 10: Admixture class model on 156 *B. trifoliolata*, *B. swaseyi* and hybrids. 0,0 is pure *B. swaseyi* and 1.0,0 is pure *B. trifoliolata.* The individuals in black in the middle are hybrids, the black individual at (0,0) is miss-identified. They appear to be F1s, with one F2. The individuals on the x-axis are potentially backcrossed and then selfed individuals.

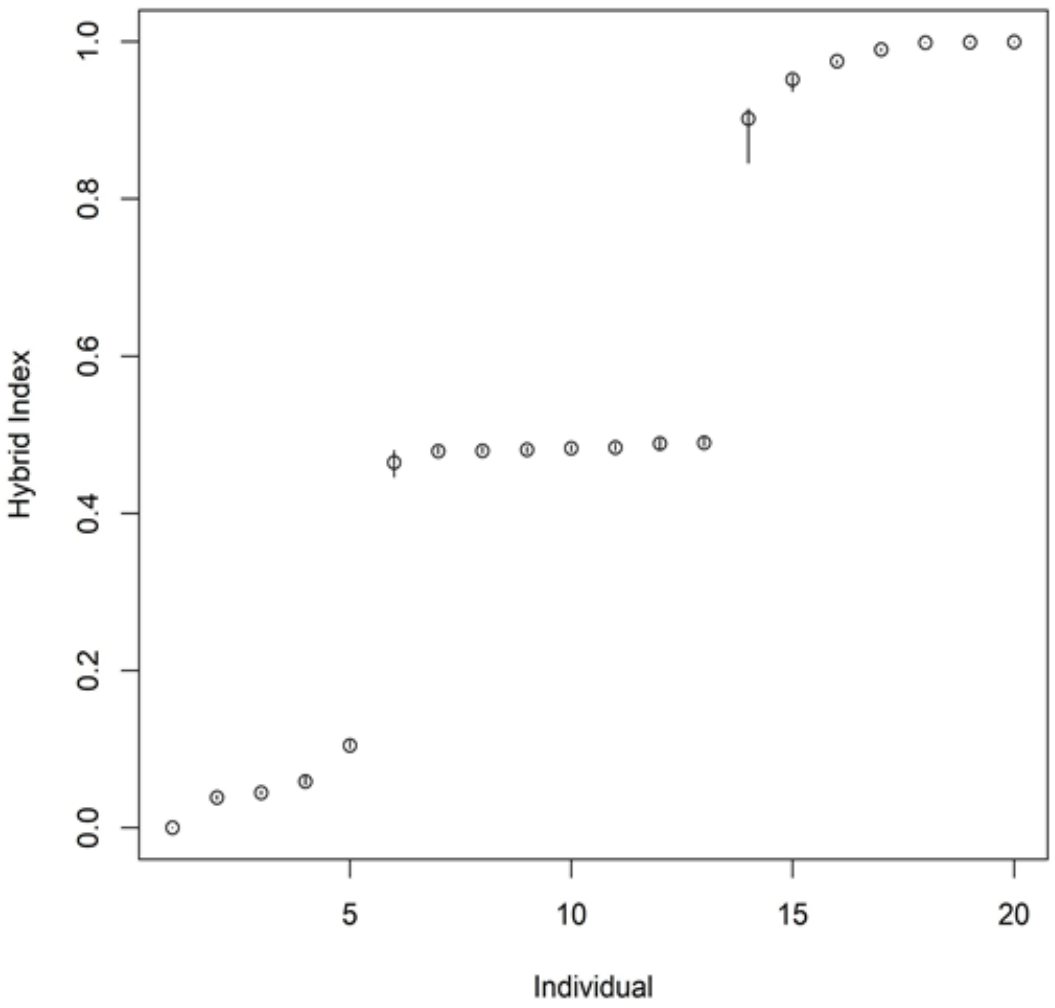

Figure 11: Hybrid index as calculated by Bayesian Genomic cline model for 20 hybrid individuals with 95% credible intervals plotted.

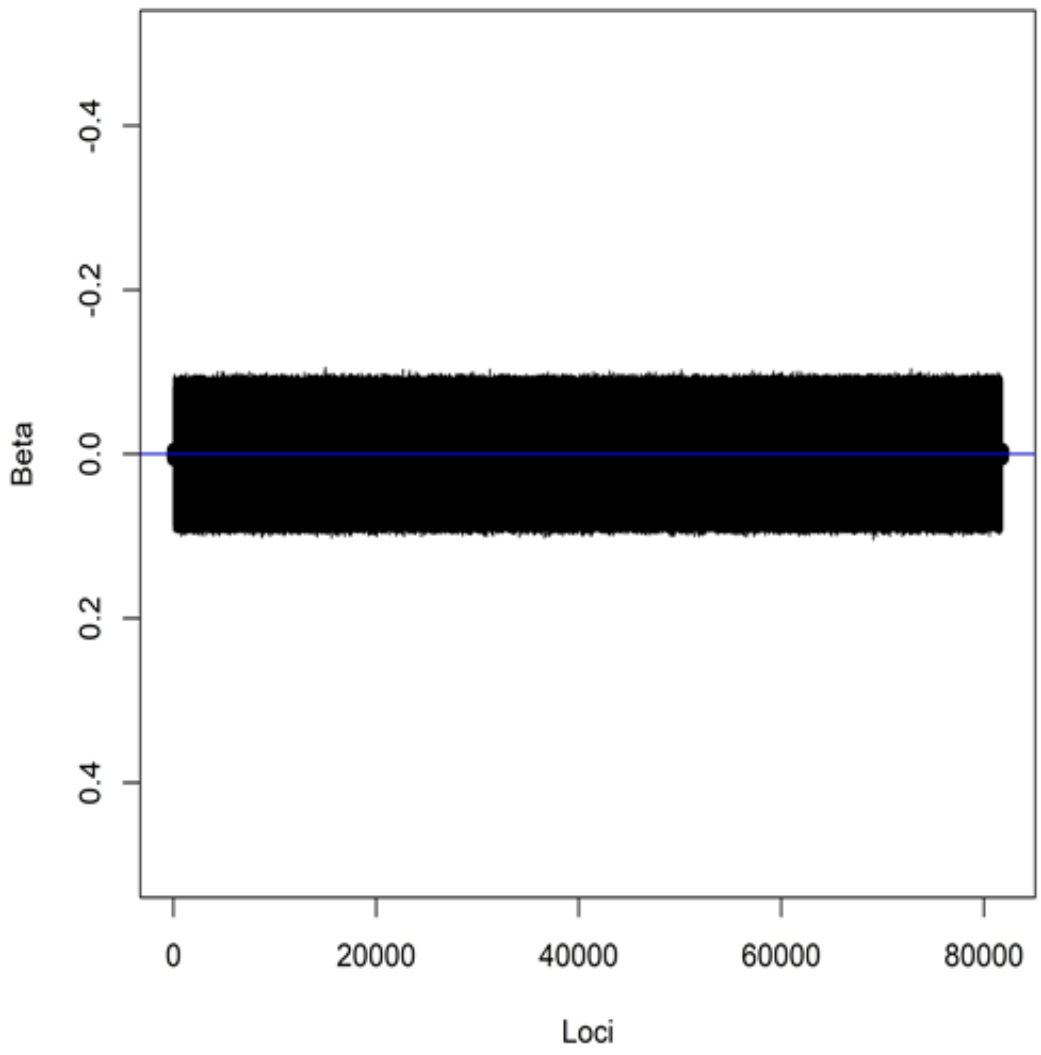

Figure 12: Distribution of betas across loci, sorted by upper 95% CI. None were considered exceptionally high.

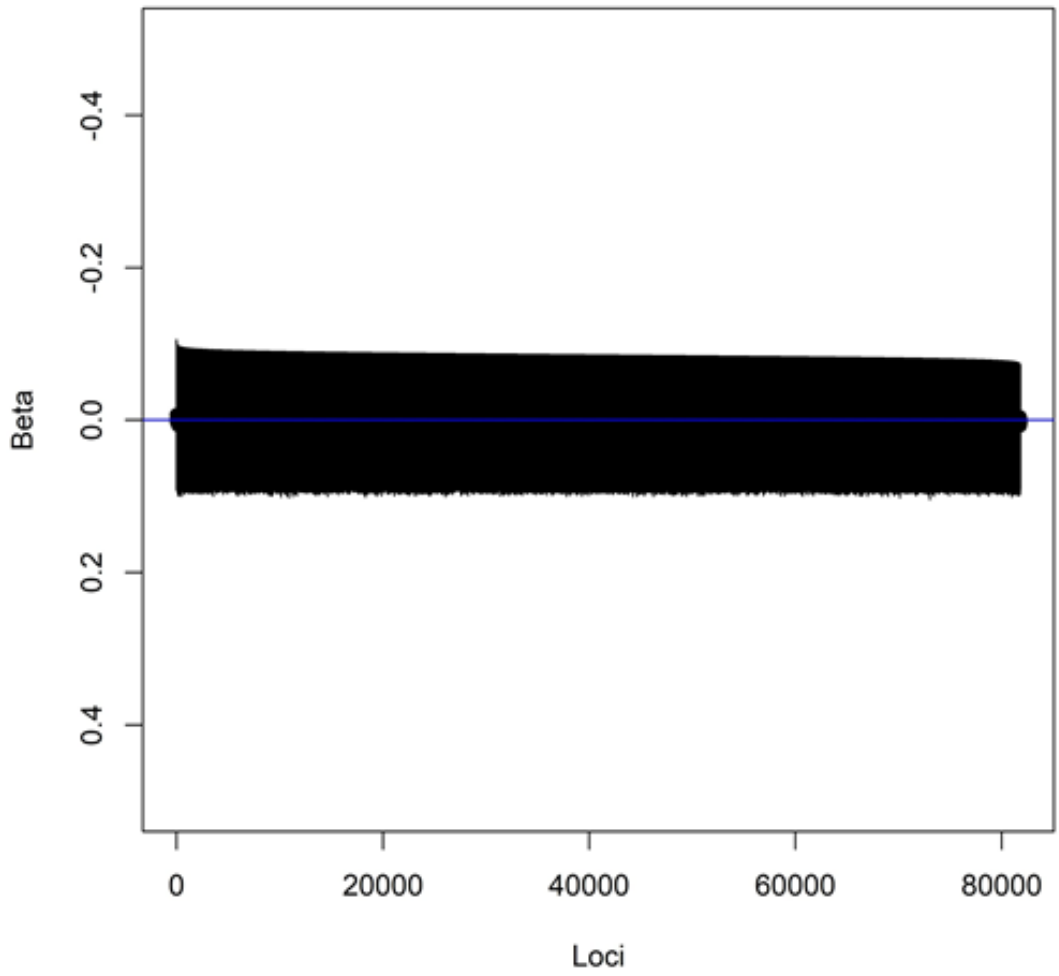

Figure 13: Distribution of betas across loci, sorted by lower 95% CI. None were considered exceptionally low.

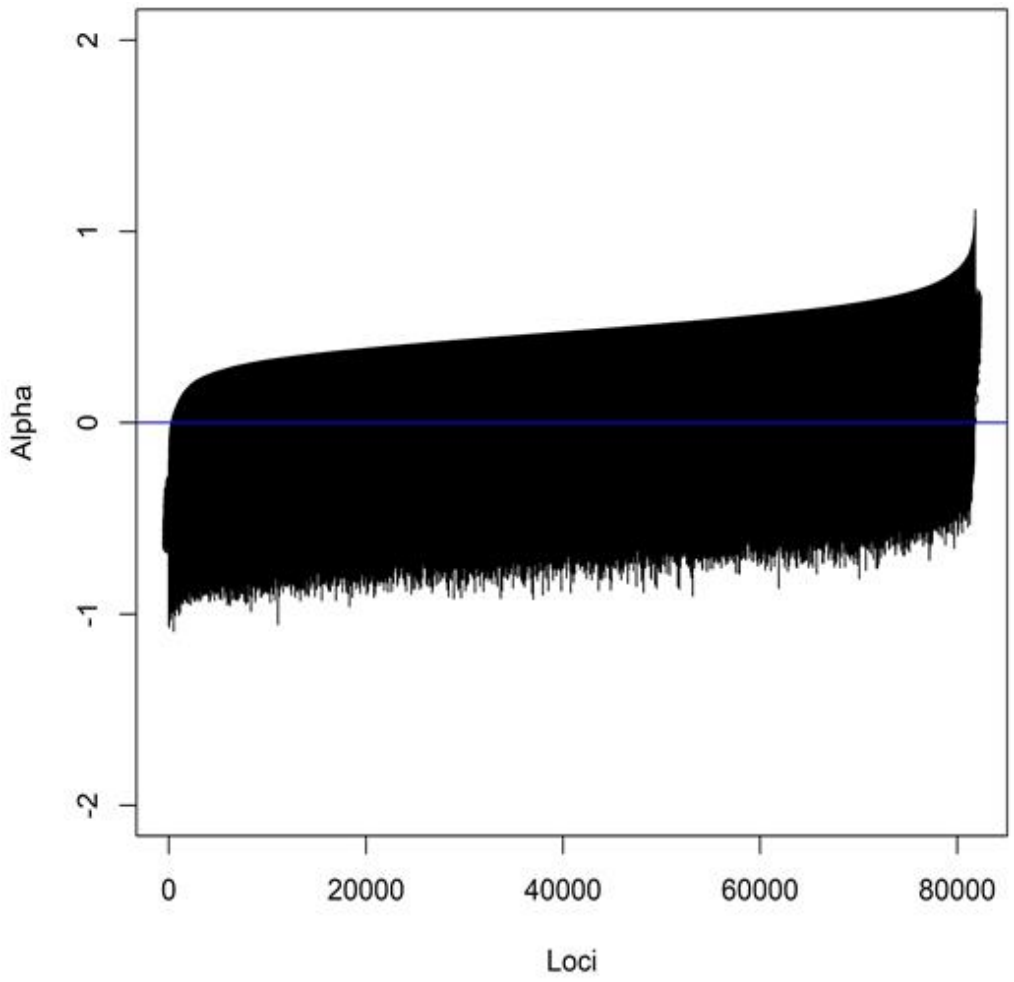

Figure 14: Distribution of α measures across loci, sorted by upper 95% CI. This shows the proportion that were exceptionally low (95% CI did not intersect zero). 229 loci were found to be exceptionally low. *B. swaseyi* to *B. trifoliolata.*

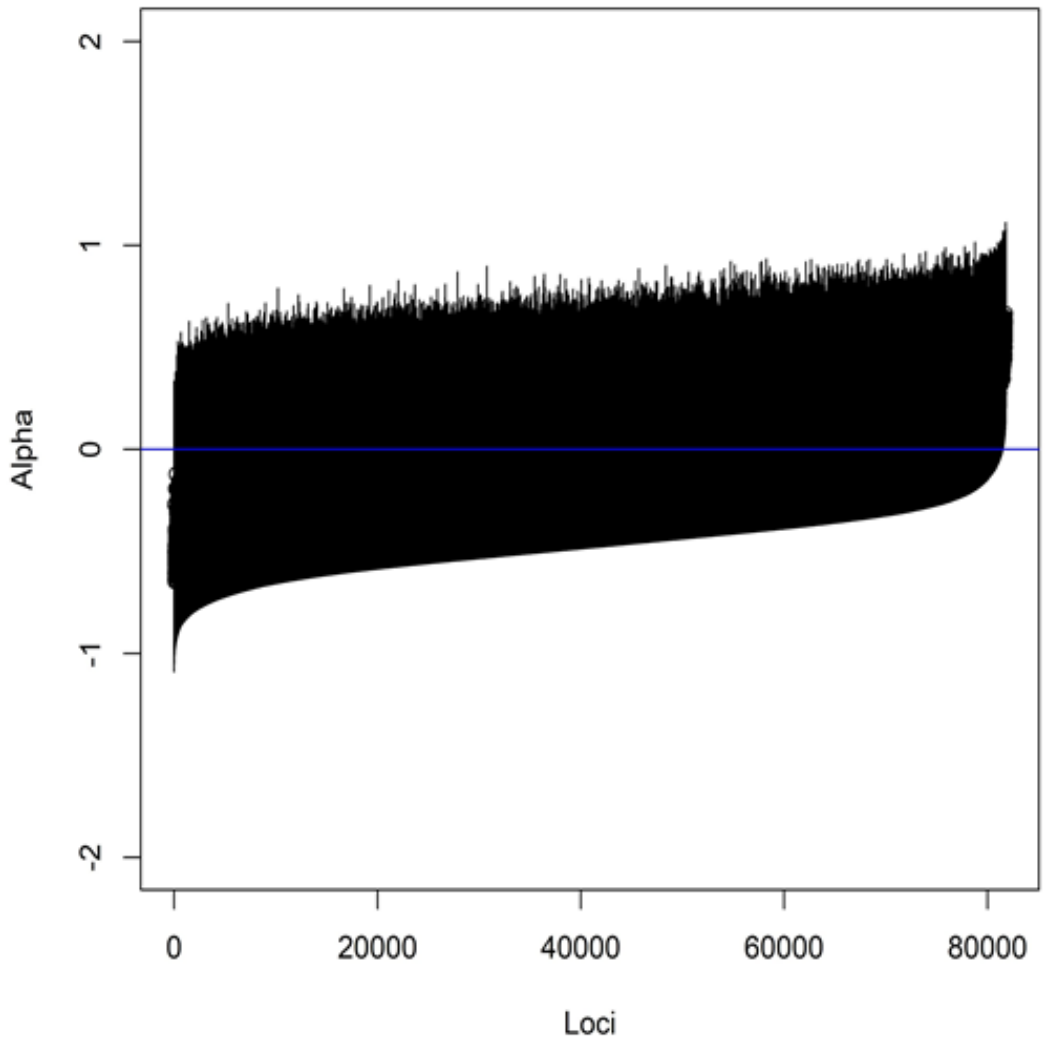

Figure 15: Distribution of α values across loci, sorted by lower 95% CI. This shows the proportion of loci that were exceptionally high (95% CI did not intercept zero). 305 loci were found to be exceptionally high. *B. trifoliolata* to *B. swaseyi.*

APPENDIX SECTION

Table S1: Listed below are each of the sampling localities along with the number of individuals sampled at each site. Also included is the number of individuals from each locality included in the analysis after selection for number of reads, 100,000.

Sampling locality	Number of individuals sampled	Number of Individuals with over 100,000 reads
Upper Purgatory	30	1
Lower Purgatory	30	0
Fisher Side Road	30	1
Cliffside Road	30	0
Cole Spring Road	23	7
East HWY 290	30	3
West Lakeshore	17	4
Grand Summit Blvd	13	3
Fitzurgh Road	16	4
Divide Pass Road	9	3
Old Burnet Road	30	4
Sycamore Creek Dr.	25	5
Myers Creek Road	11	5
Sycamore Creek Dr. 2	30	11
Heart Springs Road	7	0
Davey Crockett Dr.	14	1

Dorchester Dr.	7	4
Hamilton Pool Road	15	7
Siesta Shores Dr	9	2
Pace Bend Road	7	1
Davy Crockett Dr 2	2	1
Bertram	10	4
Vance Circle Road	13	7
Wild Bason Ledge	11	6
Sandstone and Ramble	13	6
Morgan Road	13	5
Yorktown Blvd	14	3
HWY 31	15	5
Diez Osos Trail	17	4
Unnamed Road	14	7
Rocky Top Road	16	5
Cranes Mill Road	14	2
Dedek Dr	15	2
Sherri Lea	14	2
Westlake Dr	15	5
RM 2325	15	5
Moonview	15	3
Unknown	15	2

Lantana Lane	20	4
Linda Lou Lane	15	4
Blackjack Hallow	15	5
Jacobs Well	45	33
Tires Made Easy	15	0

LITERATURE CITED

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., ... & Butlin, R. K. (2013). Hybridization and speciation. *Journal of evolutionary biology*, 26(2), 229-246.

Anderson E (1948) Hybridization of the habitat. *Evolution* 2: 1–9.

Anderson E, Hubricht L (1938) Hybridization in Tradescantia. III. The evidence for introgressive hybridization. Am J Bot 25: 396–402

Angulo, D. F., Sosa, V., & García-Franco, J. G. (2014). Floral movements: stamen motion in Berberis trifoliolata. *Botanical Sciences*, 92(1), 141-144.

Angulo, D. F., Amarilla, L. D., Anton, A. M., & Sosa, V. (2017). Colonization in North American arid lands: The journey of agarito (Berberis trifoliolata) revealed by multilocus molecular data and packrat midden fossil remains. *PloS one*, 12(2), e0168933.

Arnold, M. L., & Martin, N. H. (2009). Adaptation by introgression. *Journal of biology*, 8(9), 1-3.

Arnold, M. L., Tang, S., Knapp, S. J., & Martin, N. H. (2010). Asymmetric introgressive hybridization among Louisiana Iris species. *Genes*, *1*(1), 9-22.

Buerkle, C. A. (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, *5*(3), 684-687.

Buerkle, C. A., & Lexer, C. (2008). Admixture as the basis for genetic mapping. *Trends in ecology & evolution*, 23(12), 686-694.

Charlesworth, D. (1995). Hybrid speciation: Evolution under the microscope. *Current Biology*, 5(8), 835-836.

Coleman, R. R., Gaither, M. R., Kimokeo, B., Stanton, F. G., Bowen, B. W., & Toonen, R. J. (2014). Large-scale introduction of the I ndo-P acific damselfish Abudefduf vaigiensis into H awai'i promotes genetic swamping of the endemic congener A. abdominalis. *Molecular Ecology*, 23(22), 5552-5565.

Coyne, J. A. (1992). Genetics and speciation. *Nature* 355:511–515.

Dobzhansky, T. H. 1937. Genetics and the origin of species. Columbia Univ. Press, New York.

Doyle, J. (1991). DNA protocols for plants. *Molecular techniques in taxonomy* (pp. 283–293). Berlin Heidelberg: Springer.

Durand, H. K. (1972). Texas Mahonia—A neglected economic plant. *Economic Botany*, 26(4), 319-325.

Eckert, C. G. (2000). Contributions of autogamy and geitonogamy to self-fertilization in a mass-flowering, clonal plant. *Ecology*, 81(2), 532-542.

Falush, D., Stephens, M., & Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567-1587.

Field, D. L., Ayre, D. J., Whelan, R. J., & Young, A. G. (2011). Patterns of hybridization and asymmetrical gene flow in hybrid zones of the rare Eucalyptus aggregata and common E. rubida. *Heredity*, *106*(5), 841.

Fogelqvist, J., Verkhozina, A. V., Katyshev, A. I., Pucholt, P., Dixelius, C., Rönnberg-Wästljung, A. C., ... & Berlin, S. (2015). Genetic and morphological evidence for introgression between three species of willows. *BMC evolutionary biology*, 15(1), 1-10.

Gibson, I., Welsh, A. B., Welsh, S. A., & Cincotta, D. A. (2019). Genetic swamping and possible species collapse: tracking introgression between the native Candy Darter and introduced Variegate Darter. *Conservation Genetics*, 20(2), 287-298.

Gompert, Z., & Buerkle, C.A. (2011). Bayesian estimation of genomic clines. *Molecular Ecology*, 20, 2111-2127.

Gompert, Z., & Buerkle, C.A. (2012). Bgc: software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, 12, 1168-1176.

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, 23(18), 4555-4573.

Gompert, Z., Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., & Buerkle, C. A. (2012). Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, 66(7), 2167-2181.

Grant, V. (1981). *Plant speciation* (No. 581.38 G7671p Ej. 1). Columbia University Press,.

Greathouse, G. A., & Watkins, G. M. (1938). Berberine as a factor in the resistance of Mahonia trifoliolata and M. swaseyi to Phymatotrichum root rot. *American Journal of Botany*, 25(10), 743-748.

Gu, M., Xu, J., Han, C., Kang, Y., Liu, T., He, Y., ... & Liu, C. (2015). Effects of berberine on cell cycle, DNA, reactive oxygen species, and apoptosis in L929 murine fibroblast cells. *Evidence-Based Complementary and Alternative Medicine*, 2015.

Guichoux, E., Garnier-Géré, P., Lagache, L., Lang, T., Boury, C., & Petit, R. J. (2013). Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, 22(2), 450-462.

Harms, R. T. (2007). A field study of hybridization between Berberis swaseyi and B. trifoliolata (Berberidaceae) in Hays County, Texas. *Lundellia*, *2007*(10), 18-32.

Harrison, R. G., & Harrison, R. G. (Eds.). (1993). Hybrid zones and the evolutionary process. Oxford University Press on Demand.

Heiser Jr, C. B. (1951). Hybridization in the annual sunflowers: Helianthus annuus× H. debilis var. cucumerifolius. *Evolution*, 42-51.

Hewitt, G. M. (1988). Hybrid zones-natural laboratories for evolutionary studies. *Trends in ecology & evolution*, 3(7), 158-167.

Hou, D., Xu, G., Zhang, C., Li, B., Qin, J., Hao, X., ... & Shao, C. (2017). Berberine induces oxidative DNA damage and impairs homologous recombination repair in ovarian cancer cells to confer increased sensitivity to PARP inhibition. *Cell death & disease*, 8(10), e3070-e3070.

Husband, B. C., and H. A. Sabara. (2004). Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, Chamerionangustifolium (Onagraceae). *New Phytol.* 161:703–713.

Jacquet, S., B. Delesalle, J.P. Torreton, & J. Blanchot. 2006. Response of phytoplankton communities to increased anthropogenic influences (southwestern lagoon, New Caledonia). *Marine Ecology Progress Series*, 320: 65-78.

Jiggins, C. D., Salazar, C., Linares, M., & Mavarez, J. (2008). Hybrid trait speciation and Heliconius butterflies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506), 3047-3054.

Kay, K. M. (2006). Reproductive isolation between two closely related hummingbird-pollinated neotropical gingers. *Evolution* 60:538–552.

Kron, K. A., Gawen, L. M., & Chase, M. W. (1993). Evidence for introgression in azaleas (Rhododendron; Ericaceae): Chloroplast DNA and morphological variation in a hybrid swarm on Stone Mountain, Georgia. *American Journal of Botany*, 80(9), 1095-1099.

Langmead, B., & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359.

Lean, J.L. & D.H. Rind. 2008. How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophysical Research Letters*, 35: 1:6.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27, 2987–2993.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658-1659.

Mandeville, E.G., Parchman, T.L., McDonald, D.B., & Buerkle, C.A. 2015. Highly variable reproductive isolation among pairs of Catostomus species. *Molecular Ecology*, 24, 1856- 1872

Martin, N. H., Bouck, A. C., & Arnold, M. L. (2006). Detecting adaptive trait introgression between Iris fulva and I. brevicaulis in highly selective field conditions. *Genetics*, *172*(4), 2481-2489.

Martin, N. H., & Willis, J. H. (2007). Ecological divergence associated with mating system causes nearly complete reproductive isolation between sympatric Mimulus species. *Evolution*, 61(1), 68-82.

Mayr, E. (1942). Systematics and the origin of species. Columbia Univ. Press, New York.

Mayr, E. (1947). Ecological factors in speciation. *Evolution* 1: 263-288.. 1963. Animal species and evolution. Harvard University Press, Cambridge, Mass.

Meirmans, P.G., 2015. Seven common mistakes in population genetics and how to avoid them. *Molecular ecology*, 24(13), pp.3223-3231.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010.6: pdb-prot5448.

Neag, M. A., Mocan, A., Echeverría, J., Pop, R. M., Bocsan, C. I., Crișan, G., & Buzoianu, A. D. (2018). Berberine: Botanical occurrence, traditional uses, extraction methods, and relevance in cardiovascular, metabolic, hepatic, and renal disorders. *Frontiers in pharmacology*, 9, 557.

Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., & Buerkle, C.A. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, 21, 2991-3005.

Parchman, T.L., Gompert, Z., Braun, M.J., Brumfield, R.T., McDonald, D.B., Uy, J.A.C., Zhang, G., Jarvis, E.D., Schlinger, B.A., & Buerkle, C.A. (2013). The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, 22, 3304–3317.

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular ecology*, 25(11), 2337-2360.

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959.

Puritz, J. B., Hollenbeck, C.M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431; DOI 10.7717/peerj.431.

Qiagen. “QIAquick Gel Extraction Kit and QIAquick PCR & Gel Cleanup Kit - (EN).” QIAGEN, www.qiagen.com/us/resources/resourcedetail?id=a72e2c07-7816-436f-b920-98a0ede5159a&lang=en.

Ramsey, J., H. D. Bradshaw, and D. W. Schemske. (2003). Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M. cardinalis (Phrymaceae). *Evolution* 57:1520–1534.

Rieseberg, L. H., Van Fossen, C., & Desrochers, A. M. (1995). Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature*, 375(6529), 313-316.

Schluter, D. (2001). Ecology and the origin of species. *Trends Ecol. Evol.* 16:372–380.

Shastry, V., Adams, P. E., Lindtke, D., Mandeville, E. G., Parchman, T. L., Gompert, Z., & Buerkle, C. A. (2020). Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. *bioRxiv*.

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual review of plant biology*, 60, 561-588.

Sotola, V., Ruppel, D. S., Bonner, T. H., Nice, C. C., & Martin, N. H. (2019). Asymmetric introgression between fishes in the Red River basin of Texas is associated with variation in water quality. *Ecology and Evolution*, 9(4), 2083-2095.

Sung, C. J., Bell, K. L., Nice, C. C., & Martin, N. H. (2018). Integrating Bayesian genomic cline analyses and association mapping of morphological and ecological traits to dissect reproductive isolation and introgression in a Louisiana Iris hybrid zone. *Molecular ecology*, *27*(4), 959-978.

Whitham, T. G., Martinsen, G. D., Keim, P., Floate, K. D., Dungey, H. S., & Potts, B. M. (1999). Plant hybrid zones affect biodiversity: tools for a genetic-based understanding of community structure. *Ecology*, 80(2), 416-428.

Wu, Y., Bogdanowicz, S. M., Andres, J. A., Vieira, K. A., Wang, B., Cossé, A., & Pfister, S. E. (2020). Tracking invasions of a destructive defoliator, the gypsy moth (Erebidae: Lymantria dispar): Population structure, origin of intercepted specimens, and Asian introgression into North America. *Evolutionary applications*, 13(8), 2056.

Wu, C. I. (2001). Genes and speciation. *Journal of evolutionary biology*, 14(6), 889-891