POPULATION STRUCTURE AND GENE FLOW IN THE LOUISIANA IRIS SPECIES COMPLEX

by

Alexander Sultan Zalmat, 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 in
Population and Conservation Biology
December 2018

Committee Members:

Noland H. Martin, Chair

Chris Nice

James R. Ott

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

This work is dedicated to my late younger brother, Adam S. Zalmat, and also to my parents, Catherine and Sultan Zalmat, for enduring extreme hardship and still providing me with support and encouragement to continue my work.

ACKNOWLEDGEMENTS

I would like to first thank my thesis committee for all the advice and patience they showed throughout my tenure as a Master's student. I gratefully acknowledge the funding received for this project from the American Iris Society and the Texas State University Biology Department. I thank the members of the Nice, Ott, and Martin labs for their enduring help throughout the process of completing this project. Several others should be mentioned as well including Aiko Amano, Keith Usry, Amanda Driscoe, Kate Bell, Sarah Bialik, Joy Sung, and Chris Kneupper for help with sample collections, lab work, and computing.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

Identifying the distribution of genetic variation and gene flow is important for understanding how evolutionary dynamics have shaped the genetic structuring of populations undergoing divergence. Natural selection and genetic drift play a role in shaping the distribution of population structure and gene flow throughout the ranges of species and can drive the divergence of taxa. With the advent of next generation DNA sequencing techniques, it is now possible to explore population structure and gene flow at a genomic level throughout the range of such ecologically divergent taxa. The Louisiana Irises (*Iris,* series *Hexagonae*) comprise a group of three or more ecologically and reproductively divergent lineages that occasionally produce hybrids in nature, giving an opportunity to explore the process of speciation as it happens. Here we sampled populations of Louisiana Iris spp. in an attempt to characterize population structure and gene flow throughout their respective ranges. We discovered evidence for gene flow in some parts of the range and tested several standing hypotheses of nominal taxonomy accepted by Louisiana Iris enthusiasts. We also quantified introgression in a newly discovered hybrid zone between *Iris hexagona* and *I. brevicaulis* using a Bayesian Genomic Cline analysis. In addition, we tested the hypothesis that a purportedly hybrid species, *I. nelsonii*, indeed shows ancestry from two or more of the hypothesized parental species. We discovered that a relatively small proportion of the loci we sampled in the hybrid zone are experiencing extreme patterns of introgression. We found evidence that population structure appears to be more complex than previous taxonomic designations suggest, with more variation within the *I. brevicaulis* lineage than between other species

in the group. It was also discovered that *I. nelsonii* appeared to share ancestry with only one of the purported parent species, *I. fulva,* at the sampled loci. This study provides a foundation for future exploration of evolutionary dynamics affecting these taxa and sets the stage for understanding the standing distribution of genomic variation in the clade.

I. INTRODUCTION

The process of divergence that leads to speciation is one of the fundamental means by which biodiversity on Earth increases. Understanding the forces at play during the evolution of reproductive isolation, or species barriers, is therefore central to understanding how biodiversity arises and is maintained. Identifying the distribution of genetic variation and gene flow is a fundamental first step in understanding the evolutionary dynamics and history of populations undergoing divergence. Natural selection and drift dynamics shape population structure and influence the genetic structuring of genomes and can drive divergence between closely related taxa. Vicariance, or physical geographic barriers that result in reduced gene flow between closely related populations, has often been recognized as the main facilitator of divergence. However, speciation with gene flow has recently gained attention and provides an alternate perspective on divergence at the genomic level (Rundle and Nosil, 2005; Taylor et al., 2009; Gompert et al., 2012; Nadeau, 2014). Speciation with gene flow is a gradual process that involves the evolution of sequential reproductive barriers to hybridization between divergent lineages (Ramsey et al., 2003; Taylor et al., 2009; Teeter et al., 2010; Payseur, 2010). Under this model, speciation is a continuous process and partial reproductive isolation can evolve despite contemporary gene flow, which acts to decrease differentiation between populations (Nosil, 2008; Feder et al., 2012).

In diploids, when reproductive isolation is incomplete, interspecific mating may result in F1 progeny, each containing a copy of both parental genomes. Upon gametogenesis in F1 individuals, recombination shuffles the two parental genomes, resulting in haploid gametes containing recombinant chromosomes of both parental species. Backcrossing and recombination in later generations further breaks up the

genome and allows for selection to act with high resolution on recombined regions. Some regions of the genome may be resistant to introgression across species boundaries due to their deleterious effects in the foreign genomic or ecological background. These regions may be linked to loci involved in reproductive isolation, especially if they are found to occur less than expected at random in the heterospecific genome (Wu, 2001; Gompert et al., 2011). Conversely, some heterospecific regions of the genome may be *more* susceptible to introgression if selection acts to increase the fitness of those individuals containing such regions (Gompert et al., 2011). Such haplotypes with increased frequencies in heterospecific genomes may represent regions exhibiting adaptive introgression across species boundaries.

Genetic loci in which selection is actively working to restrict or enhance introgression can be identified using modern DNA sequencing techniques (Wu, 2001; Gompert et al., 2011). These techniques provide the opportunity to assess the relative importance of adaptive versus neutral processes during speciation. The identification of loci responsible for adaptive divergence and reproductive isolation, as well as adaptive introgression, is critical to understanding how genomic variation is partitioned during speciation.

There are several methods for detecting loci under selection that have been developed and used in non-model systems. This research proposes to use two in particular: genomic clines analysis (GCA) – which can detect loci that are under selection in hybrid zones and F_{ST} outlier analysis – which can potentially detect loci that are under selection in allopatric populations. GCA uses natural genetic recombination in hybrid zones to identify loci that are introgressing across species boundaries in a non-neutral fashion (Gompert et al., 2011). The results from a GCA identify genomic regions that

may be linked to reproductive isolation or adaptive introgression (Gompert et al., 2012, Gompert et al., 2011). Once identified, these genomic regions can be explored further to assess whether they are associated with important reproductive traits (Sung et al., 2018), or compared with patterns of differentiation in the parental species to assess patterns of selection (Gompert et al., 2012).

The Louisiana Iris species complex offers an opportunity to investigate divergence and speciation as it happens. The clade provides a system in which to study the evolutionary importance of hybridization and introgression as all species are interfertile and yet appear to maintain their evolutionary independence (Carney & Arnold, 1997;Wesselingh & Arnold, 2000; Taylor et al., 2014). With many phenotypes in the wild, the Louisiana Irises have provided biologists with taxonomic uncertainty since an early description by Small and Alexander (1931). Having observed many diverse phenotypes in the field, Small and Alexander originally suggested that the system consisted of over 80 species (Small & Alexander, 1931). Further investigation revealed that these were actually various hybrid forms resulting from introgressive hybridization between three morphologically, reproductively, and ecologically distinct species; *Iris hexagona, I. fulva,* and *I. brevicaulis*, although some have argued that *I. hexagona* may be more appropriately divided into several distinct species: *I. giganticaerulea, I. hexagona, and I. savannarum* (Viosca 1935; Forster, 1937; Riley, 1938; Arnold et al., 1991; Meerow et al., 2011). A fourth species, *Iris nelsonii*, is now recognized and is thought to be the product of hybrid speciation with parental contributions from all three of the major lineages described above (Arnold, 1993). *Iris brevicaulis, I. fulva,* and *I. hexagona* are geographically widespread in the eastern United States and are usually allopatric with the exception of those that occur in Southern Louisiana. *Iris nelsonii* is

limited to a single known locality in southern Louisiana, where all four species occur in sympatry and make distinguishable hybrid zones (Arnold et al., 1991; Arnold, 1993; Taylor et al., 2014). Most of the evolutionary and ecological work in this system has focused primarily on this area where all the species ranges overlap and hybridization is observed, leaving much of the geographic distribution unstudied and many questions to be answered (Arnold et al., 1991).

Here we aimed to characterize the distribution of genomic variation within the Louisiana Irises across their ranges and to identify localities where gene flow and hybridization have been taking place. We resolve some of the taxonomic confusion associated with *I. hexagona*, and we explore some surprising population structure amongst the other species in the clade, providing some new insights into how genetic variation is partitioned within the clade. We also address the purported hybrid origin of *I. nelsonii* by calculating admixture proportions of individuals collected from the only known locality. Finally, we discovered a hybrid zone occurring between *I. hexagona* and *I. brevicaulis* and assessed genome wide variation in introgression within the hybrid population to identify potential loci under selection.

II. METHODS

-Sample Collection

Tissue samples were collected in 2015 from individually sampled Iris individuals spanning locations throughout much of the Louisiana Iris ranges (see Figure 1). Five to ten geographically separated locations per species were sampled, except for *I. nelsonii*, for which there is only one known location. Collection locales for sites containing *I. hexagona* individuals included two sites in the Florida peninsula and several coastal sites in Louisiana and Texas. A coastal marsh in Brazoria County was sampled where there appears to be a contact zone with *I. brevicaulis*. Collection locales for *I. brevicaulis* included two sites in the northern reaches of the range, and several sites in Texas and Louisiana. *I. fulva* collections were made from five localities spanning a large portion of the range. *I. nelsonii* individuals were sampled from their only known location in Abbeville, Louisiana. Leaf tissue from 2-142 individuals per sampling locale was collected, placed in a coin envelope, and then dried in silica gel prior to DNA extraction.

-DNA sequence generation, assembly, and variation

DNA was extracted from leaf tissue samples using a standard CTAB protocol in a 96-well plate format. A reduced representation library was generated for each individual following the methods of Gompert et al. (2012) and Parchman et al. (2013). DNA from each individual was digested with restriction enzymes EcoRI and MseI. An 8 to 10 base pair oligonucleotide barcode adaptor for individual identification was then ligated to the generated fragments. These restriction ligation products were then amplified for two rounds of PCR using standard Illumina primers. After PCR, the products were then pooled and size selected for 300-400 bp length fragments using Blue Pippin technology at the University of Texas Genomic Sequencing and Analysis Facility.

The final DNA library was then sequenced in the same facility over two lanes on an Illumina HiSeq 4000 platform. Single-end, 100bp sequence reads were then assembled to the PhiX genome to remove any sequences that were known to belong to other organisms (i.e. DNA sequences that do not belong to *Iris*). The reads were then processed using custom scripts to remove barcodes and adaptor sequences. Final reads ranged from 84-86 bp in length.

No reference genome is available for Iris species, so a *de novo* assembly of a random subset of 45 million reads using the dDocent assembly (Puritz, Hollenbeck, &Gold, 2014). If reads were shared by fewer than four individuals in the dataset or were represented fewer than four times, they were removed. The resulting set of reads was then assembled using an 80% sequence similarity threshold using CD-hit software. From this, reference scaffolds were generated, and all sequence data was then assembled to the reference using Burrows Wheeler Aligner, BWA version 0.7.5a-r405 (Li et al., 2009).

Single nucleotide polymorphisms (SNPs) were then identified using SAMtools (ver. 0.1.19) and BCFtools (ver. 0.1.19.) In order for SNPs to be identified they had to be present in at least 50% of individuals in the data set, and had to have minimum of one read at that site. Genotype likelihood estimates were generated for the resulting 218,743 loci spanning 645 individuals. Allele frequency estimates were generated from the genotype likelihood estimates and loci with a global minor allele frequency of less than 0.05 were excluded. To reduce the effects of linkage disequilibrium amongst SNPs, one variable site was chosen per reference scaffold. The resulting dataset consisted of 2,693 loci, which were used for all downstream analyses.

-Population structure and gene flow

To quantify the geographic distribution of genomic variation, population genetic

parameters were estimated using ENTROPY (Gompert et al., 2014), a hierarchical Bayesian model similar to the correlated allele frequencies admixture model in STRUCTURE (Pritchard, Stephens, & Donnelly, 2000). ENTROPY differs from STRUCTURE in that it accounts for variation in sequence coverage and alignment errors. ENTROPY makes population genetic parameter estimates with no a-priori knowledge of sample localities, and accounts for variation in sequence coverage and genotyping errors in a Bayesian framework. The user designates the assumed number of clusters (k), and the ENTROPY algorithm estimates parameters based on the designated number of clusters and the posterior probability of the allele frequencies for each k. We compared models assuming 2 and up to 10 clusters (k2-k10). For each model we iterated 100,000 MCMC steps, sampling for each parameter every 10 steps, and dropping the first 5,000 steps. We ran two chains for each model of k. We did not assume that there was a "best" k solution, but rather compared each model to gain insight into different levels of population structuring. Each model can provide biologically pertinent information about genomic structuring at different scales. The Gelman-Rubin diagnostic statistic was used to check chain convergence of each run. Once it was confirmed that runs had chain convergence, admixture proportions and genotypes were averaged across chains. Runs above k10 showed poor mixing and lack of chain convergence. They were therefore not used in further analyses.

To summarize the distribution of genomic variation in our dataset across sites we used a principal components analysis using the prcomp function in R. For this, we input genotype probabilities estimated in ENTROPY to generate a genetic covariance matrix and then used this matrix to generate principal component scores.

Genome-wide variation in introgression

The ENTROPY analysis revealed patterns of admixture consistent with a hybrid zone occurring between clusters that assigned to the species designations *I. hexagona* and *I. brevicaulis* (see Results). To quantify genome-wide variation in introgression within the admixed individuals (N = 106), the Bayesian genomic clines (BGC) model was used (Gompert & Buerkle, 2011; 2012). BGC is a hierarchical model that estimates the probability of ancestry at a locus as a function of the distribution of hybrid index (h) in the dataset. For each locus the parameters α and β were estimated. The α parameter reflects either an increase (+α) or decrease (-α) in locus specific ancestry probability as a function of hybrid index. The β parameter indicates a locus specific increase or decrease in the rate of change of the cline, with positive values indicating limited rates of introgression between parental genomes and negative values indicating increased rates (Gompert and Buerkle, 2011; Gompert et al., 2012; Parchman et al., 2013). To estimate the α and β parameters two chains of the model were run, each with 50,000 MCMC steps, sampling every 5 steps and dropping the first 10,000 steps. The chains were combined after it was confirmed that they reached convergence. Medians and 95% credible intervals of α and β were reported. Loci with values of α or β whose 95% CI did not intersect zero were considered exceptional loci.

III. RESULTS

-*Sequence coverage, assembly, and sampling*

A total of 572,371,746 useable sequence reads were obtained and the de novo assembly produced 49,785 scaffolds onto which the rest of the dataset was assembled. Across these scaffolds, 218,743 variable sites were discovered in the variant calling process. Once loci with a minor allele frequency of less than 0.05 were removed and one variable site per scaffold was chosen, a total of 2,693 SNPs were included in the dataset. The final dataset included 645 individuals from 21 locations across four Louisiana Iris species ranges (Figure 1; Table 1). Sample sizes of each locality ranged from 2 to 142 (Table 1) and the mean individual coverage (the average number of reads per locus per individual) was found to be 15.66 (SD = 2.09; Figure 2).

-*Population structure*

Pairwise G_{ST} ranged from 0.0036 to 0.0786 (Table 2). The largest amount of genetic differentiation was found between a Floridian *I. hexagona* locality and a population of *I. brevicaulis* in Texas. Another notable comparison was of the same Floridian *I. hexagona* population (uuf) and the northernmost population of *I. brevicaulis*, indicating that the Floridian population is highly differentiated from *I. brevicaulis* samples across their range. Interestingly, a second Floridian *I. hexagona* locality (hgf) showed relatively less differentiation from either of the aforementioned *I. brevicaulis* populations at the edges of the range. Within *I. hexagona* comparisons, relatively little differentiation was found between individuals sampled from the Texas site and either of the Florida sites, indicating little differentiation between the opposite edges of the range. Relatively higher levels of differentiation were actually found between the two *I. hexagona* localities in Florida than between each of those sites and others at disparate

locations in the range. The *I. nelsonii* samples showed relatively low levels of differentiation from each of the five *I. fulva* localities. Moderate levels of divergence were observed for the *I. nelsonii* samples in all other population comparisons.

Principal component I explained 15.37% of the variation and principal component II explained 11.36% of the variation in the genotypic data (Figure 3). Individuals appeared to fall into five clusters defined across the first three PC axes.. Individuals designated as *I. hexagona* formed a single cluster far removed from other clusters along the first PC axis. Individuals designated as *I. brevicaulis* formed three clusters separated along the second PC axis. The two northernmost localities formed one cluster, the two Louisiana *I. brevicaulis* localities forming a second, and the Texas *I. brevicaulis* localities forming the third. Interestingly, individuals designated as *I. fulva* and *I. nelsonii* together formed a single cluster that was not resolved across the first three PC axes. Individuals sampled from the *I. hexagona/I. brevicaulis* contact zone mostly fell into one of two clusters with some intermediate individuals spanning the first PC axis in between, indicating gene flow between the two clusters at the Texas locality. The third PC axis mainly pulled out individuals identified as *I. brevicaulis* from Louisiana, however a few Floridian *I. hexagona* individuals were also included in the cluster (Figure 4).

Admixture proportions were calculated in ENTROPY for k2-k10. For k = 2, the model separated individuals designated as *I. hexagona* from all other sample localities (Figure 5). This finding is supported in the PCA where PC I mainly separates *I. hexagona* individuals from the other three clusters. For k = 3, the model adds a cluster that includes individuals designated as *I. brevicaulis* from the Texas localities, however there is some admixture within the northern and Louisiana localities (Figure 6). The northern localities appear to have ancestry from all three clusters in the model, while the Louisiana localities

show mixed ancestry from two clusters. Interestingly, the model shows all *I. nelsonii* and *I. fulva* individuals as belonging generally to the same cluster, indicating that there is more variation within the *I. brevicaulis* lineage than there is between *I. nelsonii* and *I. fulva*. The model at k = 3 also shows some admixture between the Texas *I. brevicaulis* cluster and the *I. hexagona* cluster in the sympatric contact zone indicating some hybridization (Figure 6). At k = 4 the model still does not differentiate between individuals designated as *I. nelsonii* and *I. fulva.* Instead a new cluster is formed containing the individuals designated as *I. brevicaulis* from the two localities in Louisiana (Figure 7). The two northern *I. brevicaulis* sites remain admixed with ancestry from mainly two of the clusters in the model. At k = 5 and above the model begins to break down, adding new clusters that do not show high levels of assignment in any individuals in the dataset while still maintaining the major clusters resolved under the k = 4 model (Figure 8).

-*Genome wide patterns of introgression*

ENTROPY analysis revealed the hybrid zone to be occurring between the Texas *I. brevicaulis* genotype cluster and the *I. hexagona* genotype cluster (Figure 8). Hybrid index in this hybrid zone ranged from 0.037 to 0.955 (Figure 9). Posterior estimates of genomic cline parameter α were variable across loci and ranged from -4.43 to 8.60. The β parameter was somewhat less variable and ranged from -3.15 to 1.23. In total, 238 loci (8.83%) were found to have exceptional α values (Figures 10 and 11), while only 15 loci (0.56%) were found to have exceptional β values (Figures 12 and 13). Of the exceptional alpha loci, 165 showed patterns of introgression from *I. hexagona* to *I. brevicaulis* genomic backgrounds (Figure 10), while the other 73 loci showed exceptional patterns of introgression in the other direction (Figure 11). Of the exceptional β loci, only 9 showed

reduced gene flow between species (i.e. were overrepresented in conspecific genomic backgrounds and underrepresented in heterospecific backgrounds; Figure 12). The other 6 loci showed exceptionally increased rates of introgression into heterospecific genomic backgrounds and are likely experiencing bidirectional selective introgression (Figure 13).

IV. DISCUSSION

Previous studies investigating population structure in the Louisiana Iris species have either focused on a subsection of the range or only included two of the major lineages (Hamlin & Arnold 2014). Here we sampled Louisiana Irises from all the major lineages throughout the known ranges to better understand how genomic variation is structured in the context of the entire species complex. The current results indicate that there are indeed four major lineages, however the distribution of that variation did not corroborate the current standing hypotheses regarding species designations. As might be expected, we observed individuals that showed intermediate admixture proportions in the Louisiana area where all of the major ranges overlap. We might expect this because hybridization is known to occur in this part of the range (Arnold et al., 1991; Arnold, 1993; Taylor et al., 2014; Hamlin & Arnold, 2014; Sung et al., 2018). We did not expect, however, to find admixture in a locality that contained individuals showing typical *I. brevicaulis* phenotypes. These individuals from the Iberia and St. Martin parishes (labeled bsf and bil in Fig. 1) both showed admixture with contributions from each of the three clusters in the k = 3 ENTROPY model, indicating a multiple sources of admixture in these individuals. Furthermore, the k = 4 model showed these two localities to form a single cluster with strong assignment across individuals (Figure 7), consistent with the expectations of a stable hybrid lineage. To confirm this hypothesis, an interclass ancestry analysis can reveal whether the individuals in these localities indeed exhibit ancestry classes that are consistent with a stable hybrid lineage (Fitzpatrick, 2012). Our analysis revealed that there was more variation within the *I. brevicaulis* lineage than within any of the other lineages sampled, with three distinct clusters forming. The three-way admixed cluster described above formed one such cluster. Individuals designated as *I. brevicaulis*

from Texas formed a separate distinct cluster, with some admixture observed in the Texas coastal region where there appears to be hybridization with individuals designated as *I. hexagona*. And the northernmost *I. brevicaulis* sampling localities in Illinois and Ohio (bfi and blo in Figure 1) showed admixture with genotypes containing alleles from the *I. fulva/I. nelsonii* cluster. This admixture suggests that there is hybridization occurring between *I. fulva* and *I. brevicaulis* in the northern parts of their ranges, which overlap with each other. In all, we discovered that *I. brevicaulis* populations appear to be frequently involved in gene flow with other species throughout their range and also appear to contain the most diversity within the dataset.

Another interesting finding was that *I. nelsonii* could not be resolved from *I. fulva* in any of the ENTROPY models that were run. This is contrary to the current standing hypothesis that *I. nelsonii* is a homoploid hybrid lineage with parental contributions from each of the other three *Iris* lineages. Under this hypothesis, we would expect to find individuals from the *I. nelsonii* lineage to show intermediate assignment probabilities to three clusters. We did not observe this in any of the ENTROPY models. Interestingly, *I. nelsonii* individuals assigned strongly to the *I. fulva* cluster, which contained individuals from throughout the range.

I. hexagona individuals from across the range assigned mainly to one cluster across runs of k, however a locality in Florida (designated as hgf in Fig. 1) appeared at first to show some admixture despite its close proximity and phenotypic resemblance to the the other Florida locality (uuf in Figure 1). This observation can be explained by the fact that the hgf locality showed the lowest coverage of any of the localities in the study (Figure 2), and therefore this pattern could be an artifact of low coverage for those individuals. It is also observed that individuals from this locality clustered in the middle

of the PC space in the PCA, which would be expected when the clustering algorithm has a hard time assigning individuals dude to low coverage. It should be noted that individuals from the three-way admixed group in Louisiana clustered in the same PC space, however these individuals showed much higher coverage and assigned strongly to a single cluster in the k=4 ENTROPY model. In none of the models do the hgf individuals assign strongly to a single cluster (as with the potential stable hybrid lineage discovered in Louisiana), however this alone does not necessarily mean that these individuals are not admixed as they may represent a locality experiencing ongoing hybridization. This might be the case if not for the fact that the admixture shown is from the *I. fulva* cluster, for which there are not any known populations in the vicinity, or even in the entire state of Florida. It is possible that these admixed genomic regions have introgressed from across the range into the Florida population of question, however if this were the case we might also expect to find those introgressed regions in individuals from the other Florida sampling locality (which was only several miles down the same road), a pattern we did not observe. It has been proposed that some Floridian populations of what we would designate here as *I. hexagona* represent distinct lineages (dubbed *I. giganticaerulea* and *I. savannarum*) when compared to populations in the rest of the range. We did not observe evidence for this in these data. Individuals designated as *I. hexagona* mainly formed one cluster that held throughout the tested ENTROPY models. In addition, the only somewhat distinct cluster that contained Floridian *I. hexagona* individuals was the admixed locality described above (labeled hgf in Figure 1) and there is not evidence to support that this admixture represents a stable lineage in these data, even if we believe that the admixture is not an artifact of low coverage.

In Texas, we discovered what phenotypically appeared to be a zone of contact

between individuals of *I. hexagona* and *I. brevicaulis*. The ENTROPY results indicated that individuals sampled from this location did indeed show admixture. This finding is supported by the PCA, which showed individuals from that locality occupying PC space spanning along the first PC axis (Figure 3). We took the opportunity to assess patterns of introgression at the site and performed a BGC analysis. Estimates of hybrid index from individuals within the hybrid zone ranged from 0.037 to 0.955 (Figure 9), consistent with the expectations of a hybrid zone. Our sample appeared to contain more admixed individuals with primarily *I. hexagona* genomes, indicating asymmetric introgression, or alternatively a lack of spatial sampling on the *I. brevicaulis* end of the contact zone. The zone of contact did not appear to be mosaic, as other Louisiana Iris hybrid zones have been found to be (Sung et al., 2018), but rather exhibited a gradual change in phenotype from *I. brevicaulis* to *I. hexagona* as we traveled closer to the coastal salt marsh. Introgression was found to be asymmetric in our sample, with a bias in gene flow from *I. hexagona* to *I. brevicaulis* (Figures 10 and 11). A small proportion of the genome was found to show patterns of exceptional introgression in either direction (8.8% of SNPs) as observed with the alpha parameter. Significantly more loci (more than twofold) were observed introgressing into *I. hexagona* genomes rather than in the other direction, suggesting some adaptive variation that originated in *I. brevicaulis* populations. The beta parameter showed exceptional patterns in notably fewer loci, with 9 exceptionally high positive beta loci. Positive beta indicates a lack of introgression that is outside neutral expectations. This suggests that a very small proportion of the genome (> 0.3%) is experiencing resistance to gene flow. Even fewer loci were found to have exceptionally low negative Beta values. These loci presumably are undergoing bidirectional selective introgression.

Conclusions

In all, we found that patterns of genomic structuring were surprisingly contrary to the currently accepted nominal taxonomy. We discovered that *I. nelsonii* does not appear to show patterns of genomic variation consistent with the hypothesis that it is a homoploid hybrid species. Furthermore, we discovered evidence that elsewhere in Louisiana, where species ranges significantly overlap, there appears to potentially be a stable hybrid lineage phenotypically similar to *I. brevicaulis*. We found that there is more genomic variation within what is designated as *I. brevicaulis* than there is in any of the other nominal species designations. In addition, we provided evidence that *I. hexagona* appears to be generally panmictic throughout its range save a few areas where admixture is evident. We discovered a new hybrid zone near the Texas coast and quantified patterns of genomic introgression. We found that genomic introgression occurred in a very small proportion of the sampled genome, and that an even smaller proportion was resistant to introgression. These findings may significantly change our understanding of the distribution of genetic variation within the Louisiana Iris clade and lay the foundation for future studies to investigate how natural selection shapes genomic structuring.

Table 1: Sample localities that were included in the study. Each location has a three letter ID, a species designation based on the phenotype of individuals sampled at the location, and the number of individuals collected from the locality. Also shown are the latitude and longitude coordinates of each location.

Location	ID	Species	N	Latitude	Longitude
Lucas, Ohio	blo	*brevicaulis*	2	41.683	-83.367
Fayette, Illinois	bfi	*brevicaulis*	16	38.927	-89.114
St. Landry, Louisiana	bsf	*brevicaulis*	28	30.547	-91.981
Iberia, Louisiana	bil	brevicaulis	16	29.978	-91.754
Brazos, Texas	bzt	*brevicaulis*	34	30.569	-96.202
Galveston, Texas	ugt	*brevicaulis*	6	29.509	-95.116
Fort Bend, Texas	uft	*brevicaulis*	9	29.379	-95.583
Matagorda, Texas	umt	*brevicaulis*	7	28.915	-95.756
Harris, Texas	uht	*brevicaulis*	7	28.915	-95.756
Brazoria, Texas	ubt	*brevicaulis*	25	29.009	-95.486
Brazoria, Texas	bbt/hbt	*brevicaulis*/*hexagona* contact zone	142	28.882	-95.584
St. Mary, Louisiana	hml	*hexagona*	67	29.776	-91.774
Assumption, Louisiana	ual	*hexagona*	19	29.907	-91.185
Lee, Florida	uuf	*hexagona*	60	26.575	-81.822
Glades, Florida	hgf	*hexagona*	22	26.806	-81.448
Union, Illinois	fui	*fulva*	25	37.443	-89.396
Fulton, Kentucky	ffk	*fulva*	20	36.525	-89.315
Carroll, Mississippi	fcm	*fulva*	10	33.413	-90.155
St. Landry, Louisiana	fll	*fulva*	49	30.546	-91.864
St. Martin, Louisiana	fml	*fulva*	8	30.165	-91.814
Vermillion, Louisiana	nvl	*nelsonii*	73	29.902	-92.096
Totals	21	4	645		

Table 2: Pairwise G_{ST} matrix for each locality comparison. Values ranged from 0.0036 to 0.0786.

	bbt	bfi	bil	blo	bsf	bzt	fcm	ffk	fll	fml	fui	hbt	hgf	hml	nvl	ual	ubt	uft	ugt	uht	umt	uuf
bbt	0.0000																					
bfi	0.0249	0.0000																				
bil	0.0388	0.0401	0.0000																			
blo	0.0341	0.0193	0.0502	0.0000																		
bsf	0.0395	0.0392	0.0053	0.0498	0.0000																	
bzt	0.0036	0.0262	0.0419	0.0360	0.0426	0.0000																
fcm	0.0419	0.0319	0.0316	0.0446	0.0327	0.0440	0.0000															
ffk	0.0381	0.0274	0.0287	0.0379	0.0284	0.0404	0.0078	0.0000														
fll	0.0413	0.0328	0.0306	0.0429	0.0305	0.0439	0.0104	0.0060	0.0000													
fml	0.0413	0.0355	0.0265	0.0461	0.0263	0.0443	0.0171	0.0133	0.0082	0.0000												
fui	0.0382	0.0264	0.0293	0.0380	0.0289	0.0404	0.0076	0.0038	0.0070	0.0139	0.0000											
hbt	0.0356	0.0429	0.0323	0.0526	0.0323	0.0409	0.0447	0.0404	0.0425	0.0407	0.0406	0.0000										
hgf	0.0216	0.0196	0.0120	0.0298	0.0114	0.0249	0.0160	0.0117	0.0133	0.0136	0.0118	0.0136	0.0000									
hml	0.0532	0.0562	0.0382	0.0658	0.0379	0.0592	0.0540	0.0490	0.0504	0.0477	0.0493	0.0046	0.0201	0.0000								
nvl	0.0403	0.0336	0.0280	0.0433	0.0283	0.0430	0.0103	0.0058	0.0041	0.0108	0.0066	0.0397	0.0122	0.0473	0.0000							
ual	0.0514	0.0525	0.0376	0.0625	0.0375	0.0570	0.0486	0.0443	0.0460	0.0447	0.0446	0.0062	0.0187	0.0045	0.0431	0.0000						
ubt	0.0038	0.0257	0.0396	0.0352	0.0404	0.0050	0.0412	0.0378	0.0414	0.0423	0.0376	0.0373	0.0224	0.0548	0.0403	0.0527	0.0000					
uft	0.0086	0.0307	0.0475	0.0409	0.0484	0.0090	0.0487	0.0452	0.0491	0.0498	0.0451	0.0451	0.0295	0.0637	0.0482	0.0613	0.0091	0.0000				
ugt	0.0137	0.0370	0.0517	0.0467	0.0520	0.0143	0.0531	0.0495	0.0535	0.0540	0.0492	0.0498	0.0336	0.0680	0.0523	0.0657	0.0138	0.0179	0.0000			
uht	0.0129	0.0381	0.0536	0.0470	0.0541	0.0135	0.0565	0.0528	0.0556	0.0555	0.0532	0.0518	0.0360	0.0705	0.0545	0.0681	0.0141	0.0184	0.0241	0.0000		
umt	0.0108	0.0355	0.0502	0.0452	0.0513	0.0115	0.0527	0.0488	0.0524	0.0527	0.0490	0.0479	0.0325	0.0663	0.0515	0.0640	0.0109	0.0159	0.0205	0.0209	0.0000	
uuf	0.0612	0.0653	0.0474	0.0742	0.0473	0.0672	0.0643	0.0588	0.0604	0.0580	0.0596	0.0093	0.0276	0.0063	0.0565	0.0096	0.0627	0.0716	0.0755	0.0786	0.0738	0.0000

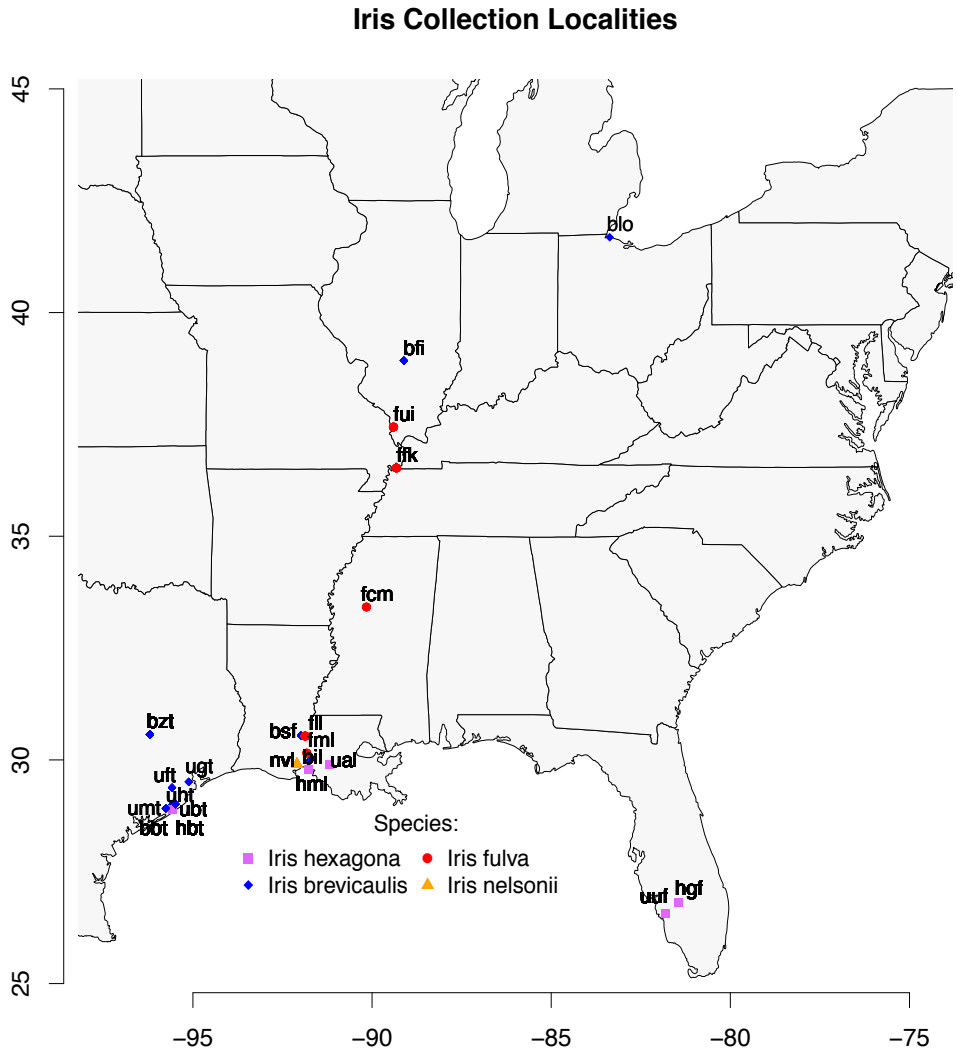

Figure 1: Map of collection localities. Species designations were based on phenotype of the sampled individuals at each locality.

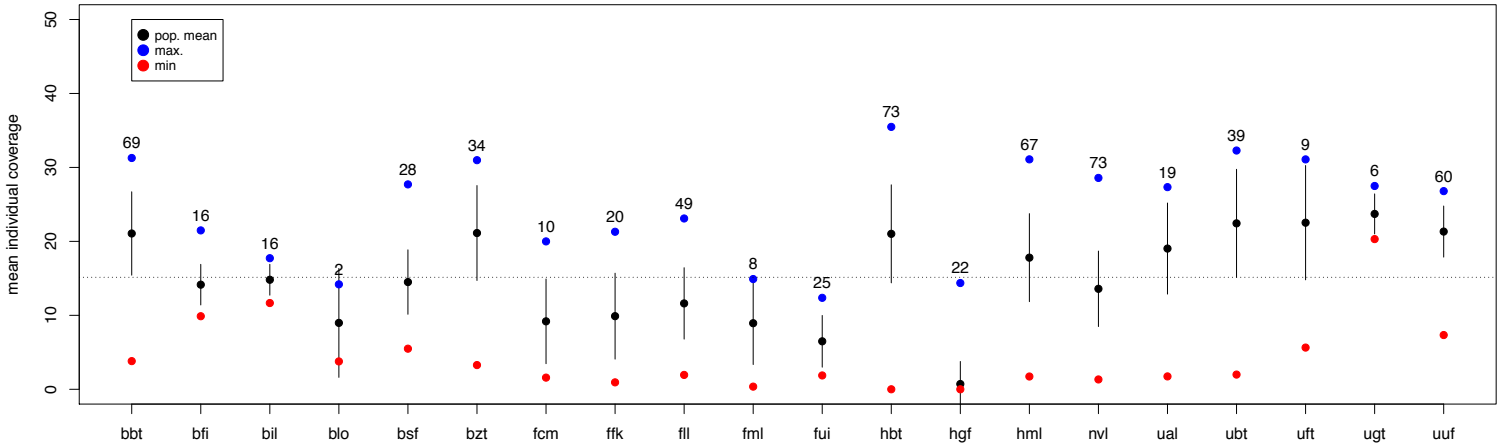

Figure 2: Individual mean coverage (number of reads) per locus shown by locality. Minimum, maximum, and mean values are shown. The dotted line represents the mean across all localities.

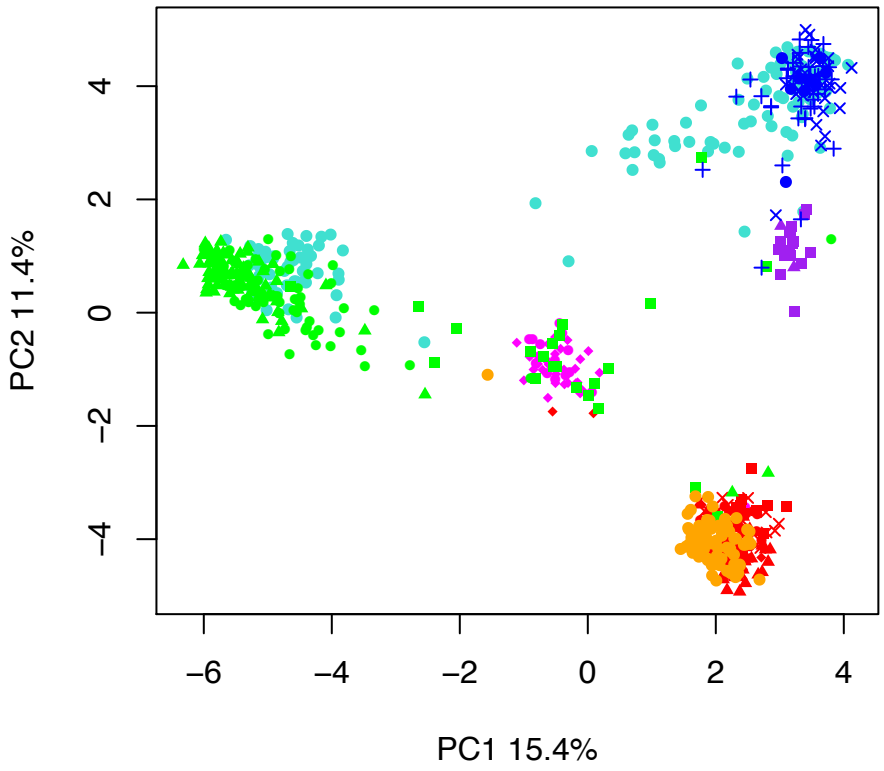

Figure 3: PC 1 versus PC 2. Individuals are marked by color for cluster designation as informed by the K=4 model in ENTROPY, and by shape for sample locality within each species. *I. nelsonii* individuals are colored orange, *I. fulva* are shown in red, *I. brevicaulis* are blue, purple and magenta, and *I. hexagona* individuals are shown in green.

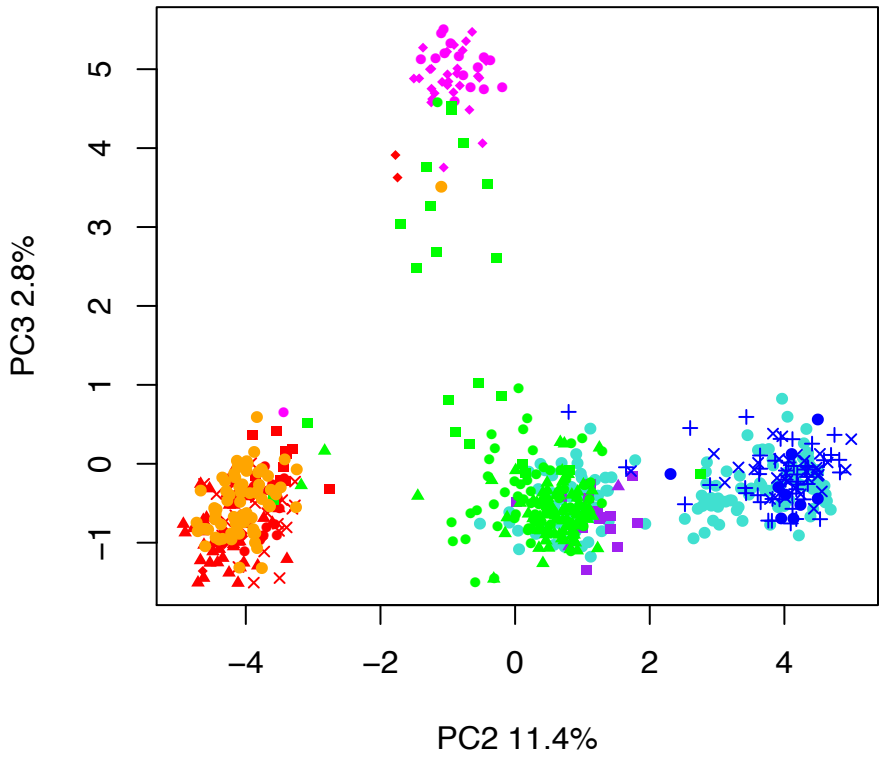

Figure 4: PC 2 versus PC 3. Individuals are marked by color for cluster designation as informed by the K=4 model in ENTROPY, and by shape for sample locality within each species. *I. nelsonii* individuals are colored orange, *I. fulva* are shown in red, *I. brevicaulis* are blue, purple and magenta, and *I. hexagona* individuals are shown in green.

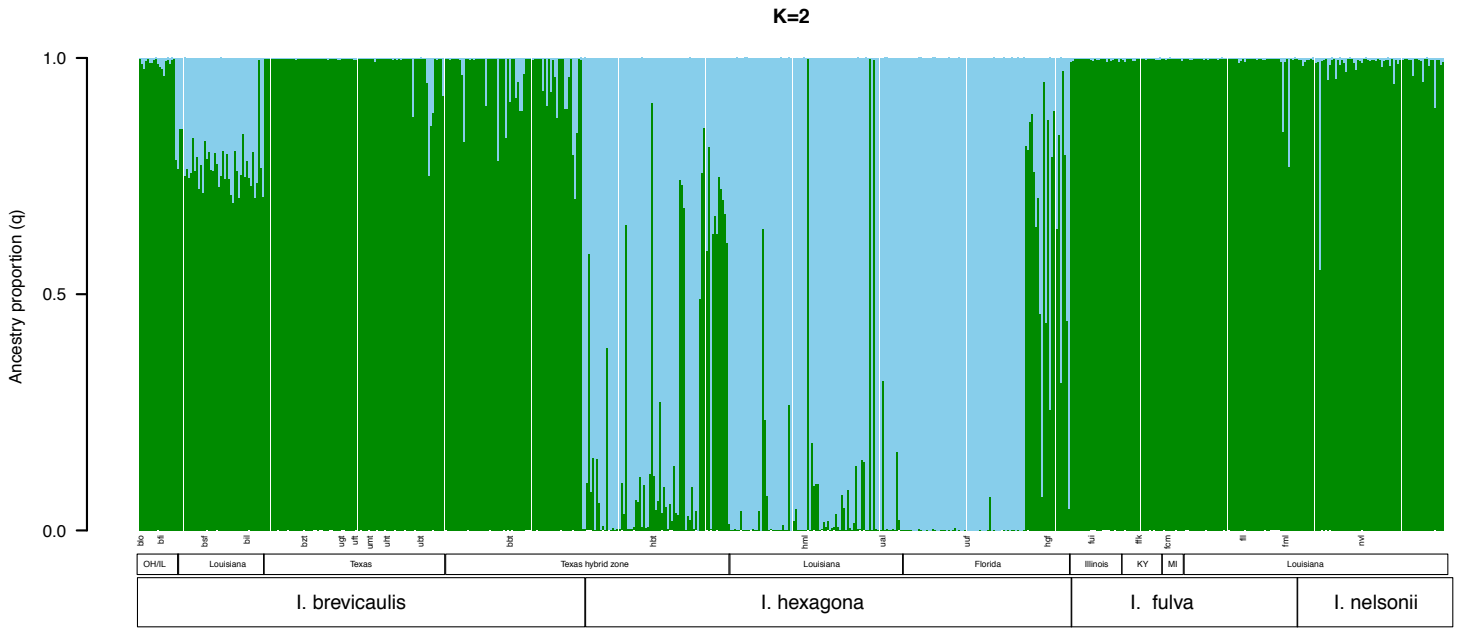

Figure 5: ENTROPY model output for K=2, with n=645, and 2693 loci included in the model. The x-axis is labeled first by locality, then by state, and then by species designation.

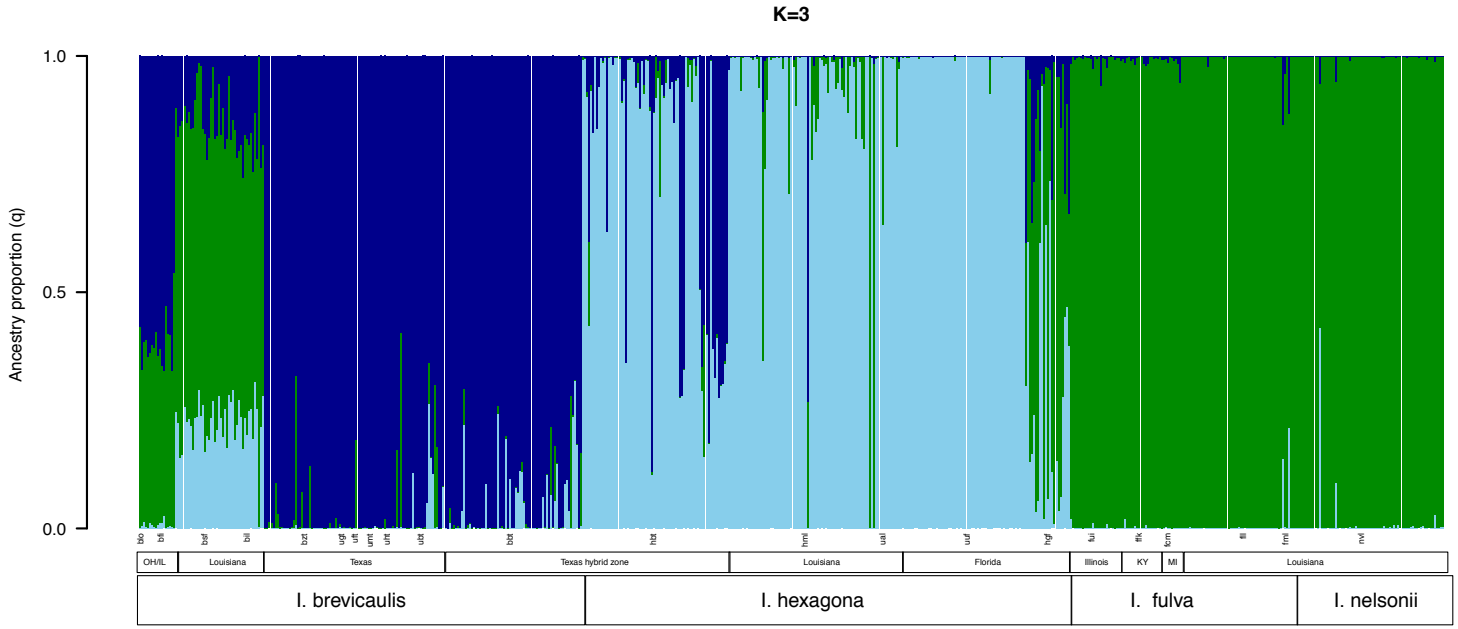

Figure 6: ENTROPY model output for K=3, with n=645, and 2693 loci included in the model. The x-axis is labeled first by locality, then by state, and then by species designation.

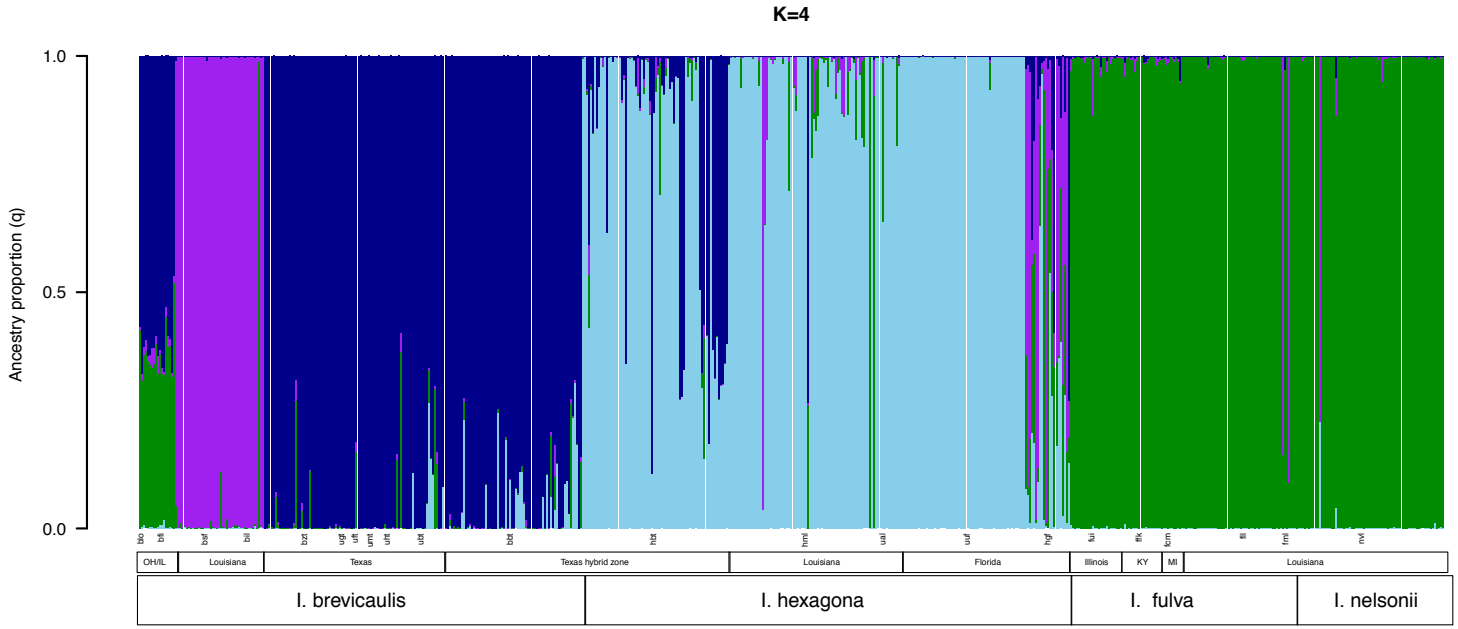

Figure 7: ENTROPY model output for K=4, with n=645, and 2693 loci included in the model. The x-axis is labeled first by locality, then by state, and then by species designation.

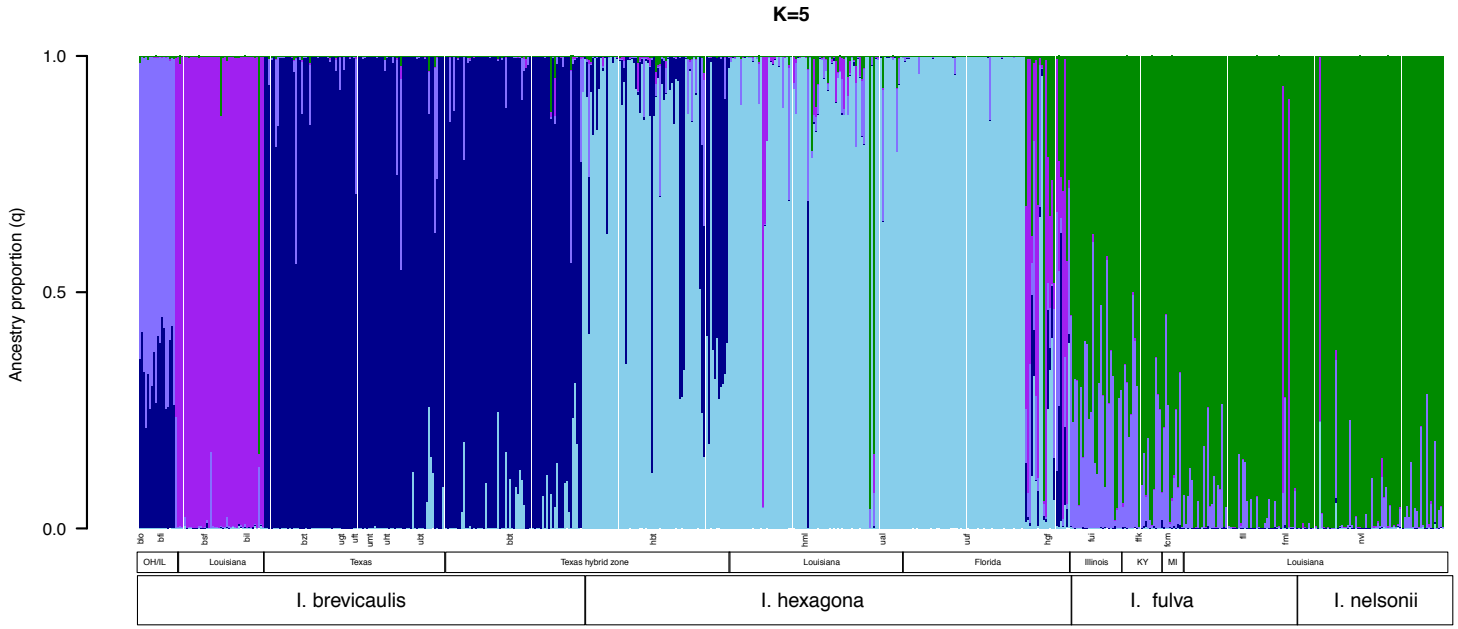

Figure 8: ENTROPY model output for K=5, with n=645, and 2693 loci included in the model. The x-axis is labeled first by locality, then by state, and then by species designation. In this run we began to see model breakdown with the 5th cluster showing only intermediate assignments.

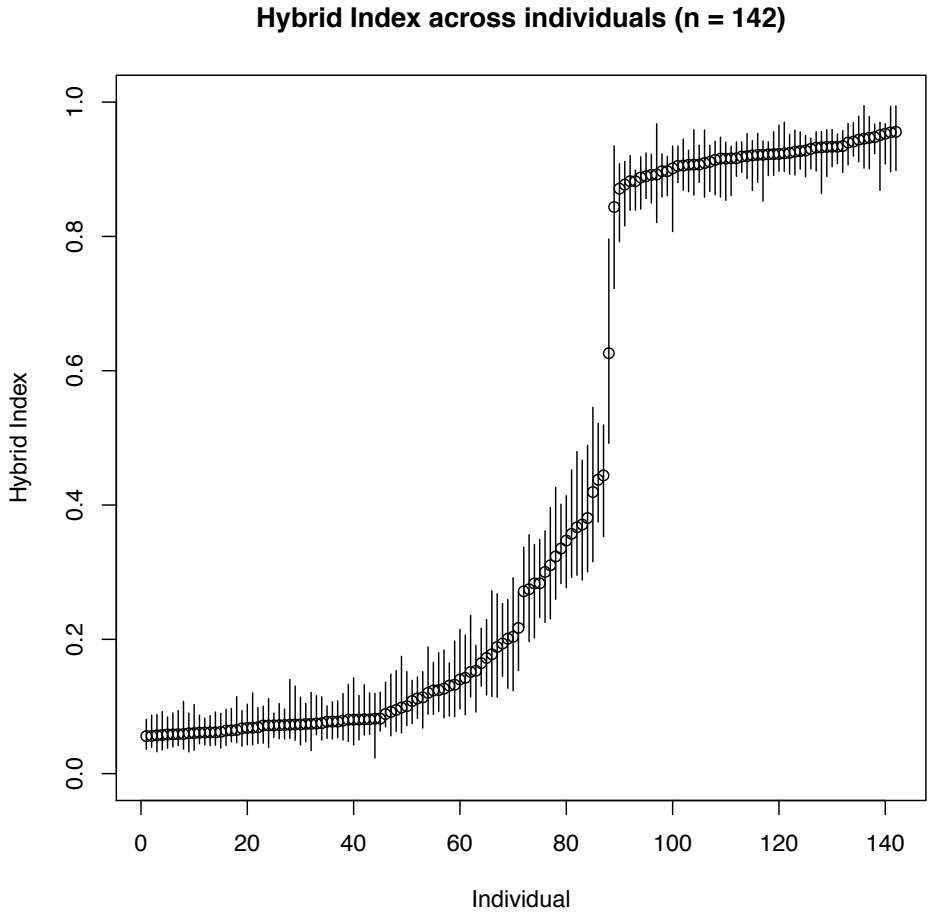

Figure 9: Hybrid index (HI) estimated from individuals in the Texas hybrid zone. HI of 0 denotes pure *I. hexagona*, while HI of 1 indicates pure *I. brevicaulis.*

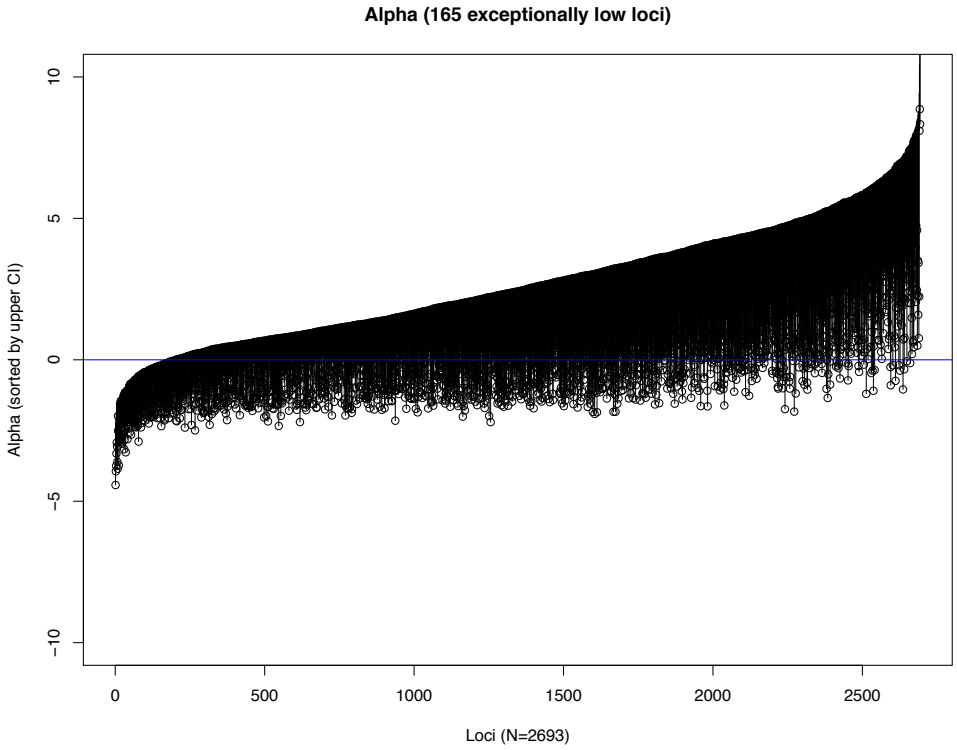

Figure 10: Distribution of alpha across loci. The data are sorted by upper CI to show the proportion of the sampled loci that were exceptionally low (upper CI did not intersect zero).

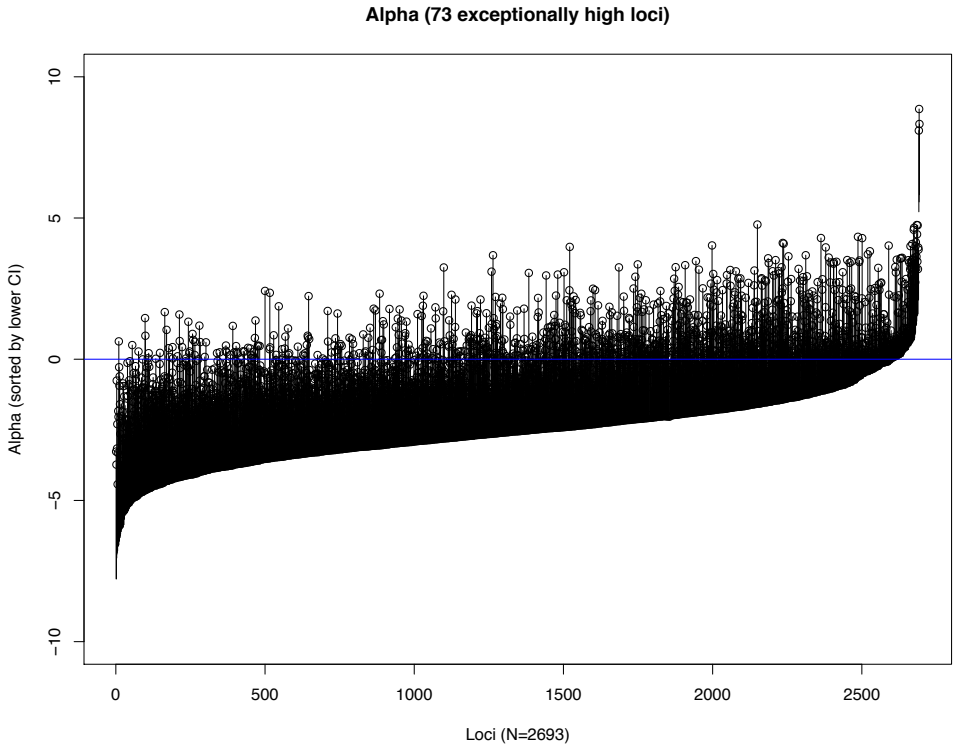

Figure 11: Distribution of alpha across loci. The data are sorted by lower CI to show the proportion of the sampled loci that were exceptionally high (lower CI did not intersect zero).

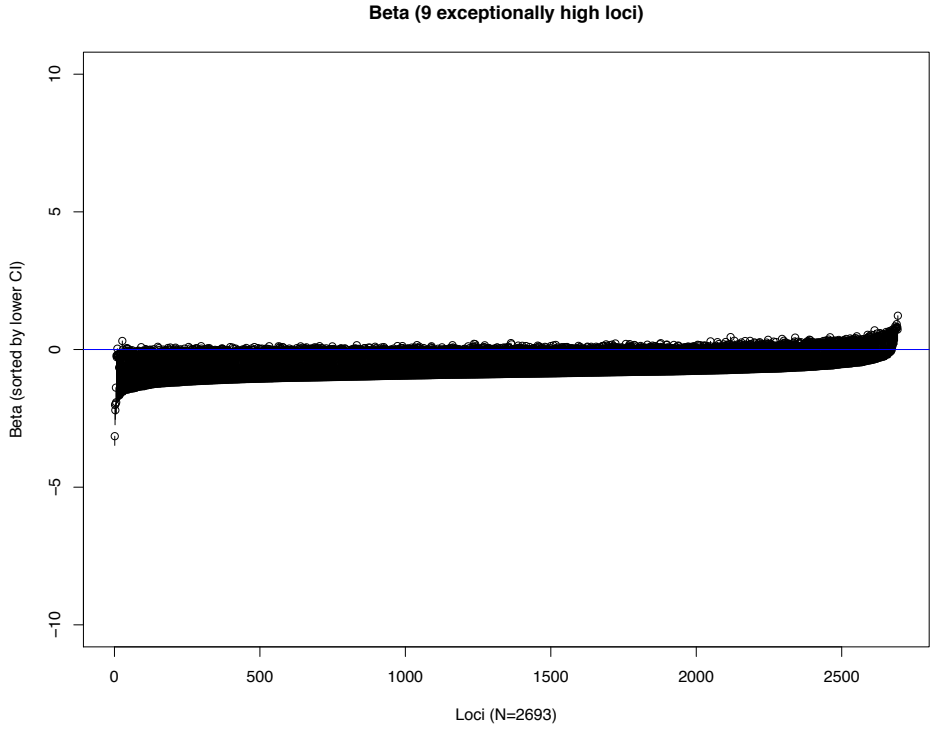

Figure 12: Distribution of beta across loci. The data are sorted by upper CI to show the proportion of the sampled loci that were exceptionally low (upper CI did not intersect zero).

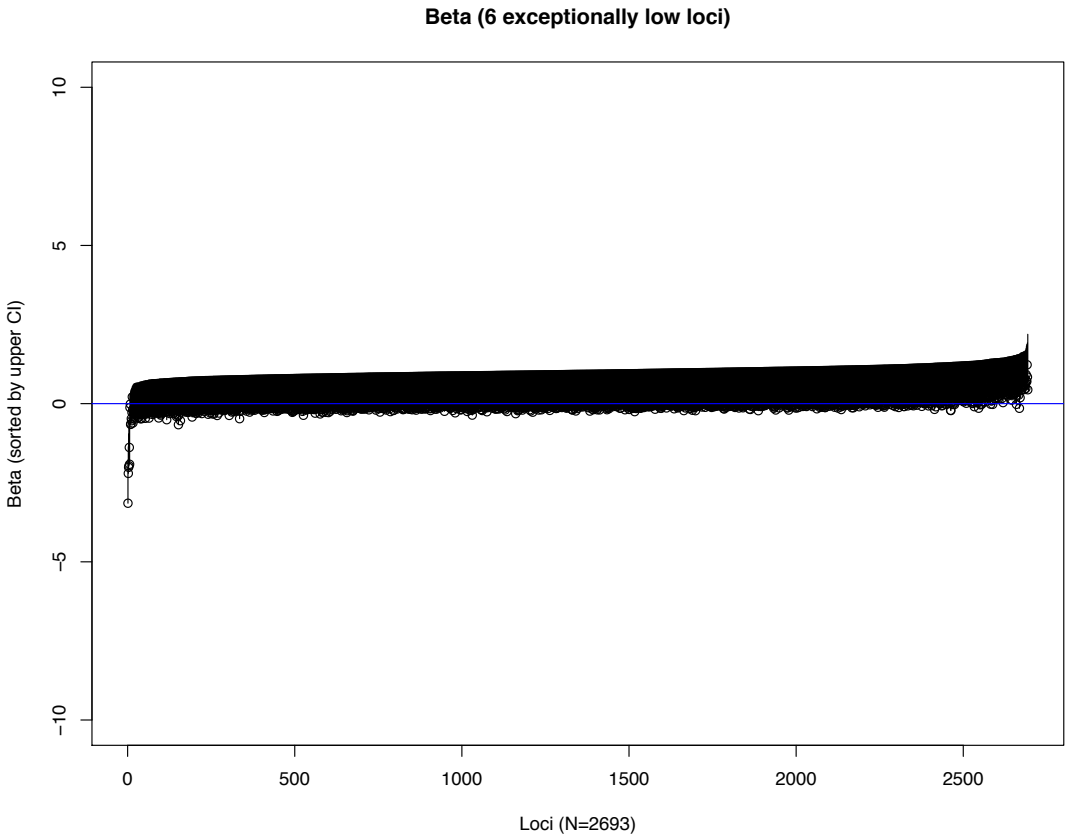

Figure 13: Distribution of beta across loci. The data are sorted by upper CI to show the proportion of the sampled loci that were exceptionally low (upper CI did not intersect zero).

LITERATURE CITED

1. Arnold ML, Buckner CM, Robinson JJ. *Pollen-mediated introgression and hybrid speciation in Louisiana irises*. Proceedings of the National Academy of Science U S A. 1991 Feb 15; 88(4):1398–402.

2. Arnold ML. *Iris nelsonii (Iridaceae): origin and genetic composition of a homoploid hybrid species*. American Journal of Botany. 1993; 80(5): 577.

3. Carney SE, Arnold ML. *Differences in pollen-tube growth rate and reproductive isolation between Louisiana irises*. Journal of Heredity. 1997; 88(6): 545–9.

4. Cheng-Jung S, Bell KL, Nice CC, Martin NH. *Integrating Bayesian genomic cline analysis and association mapping of morphological and ecological traits to dissect reproductive isolation introgression in Louisiana Iris hybrid zone.* Molecular Ecology. 2018; 27: 959-78.

5. Feder JL, Egan SP, Nosil P. *The genomics of speciation-with-gene-flow*. Trends in Genetics. Elsevier Ltd; 2012; 28(7): 342–50.

6. Fitzpatrick BM, *Estimating ancestry and heterozygosity of hybrids using molecular markers.* Evolutionary Biology. 2012; 12: 131.

7. Foster, R.C. *A cyto-taxonomic survey of the North American species of Iris. Contributions from the Gray Herbarium,* 1937; no. CXIX, pp. 3-80.

8. Gompert Z, Buerkle CA. *A hierarchical bayesian model for next-generation population genomics*. Genetics. 2011; 187(3): 903–17.

9. Gompert Z, Buerkle AC. *Bayesian estimation of genomic clines*. Molecular Ecology. 2011; 20: 2111–27.

10. Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. *Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species*. Evolution (N Y). 2012; 66(7): 2167–81.

11. Gompert Z, Lucas LK, Nice CC, Buerkle CA. *Genome divergence and the genetic architecture of barriers to gene flow between Lycaeides idas and L. Melissa.* Evolution. 2013 Sep; 67(9): 2498–514.

12. Hamlin JP, Arnold ML. *Determining population structure and hybridization for two iris species*. Ecology and Evolution. 2014 Mar; 4(6): 743–55.

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

14. Martin NH, Willis JH. *Ecological divergence associated with mating system causes nearly complete reproductive isolation between sympatric mimulus species*. Evolution (NY). 2007; 61(1): 68–82.

15. Meerow AW, Gideon M, Kuhn DN, Mopper S, Nakamura K. *The Genetic Mosaic of Iris Series Hexagonae in Florida: Inferences on the Holocene History of the Louisiana Irises and Anthropogenic Effects on Their Distribution*. International Journal of Plant Science. 2011; 172(8):1026–52.

16. Nadeau N. *Butterfly genomics sheds light on the process of hybrid speciation.* Molecular Ecology. 2014; 4441–3.

17. Narum SR, Hess JE. *Comparison of F_{ST} outlier tests for SNP loci under selection.* Molecular Ecology Resources. 2011; 11:184–94.

18. Nosil P. *Speciation with gene flow could be common*. Molecular Ecology. 2008; 17(9): 2006–8.

19. Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, & Buerkle, CA. (2012). *Genome-wide association genetics of an adaptive trait in lodgepole pine.* Molecular Ecology. 2012; 21: 2991-3005.

20. Payseur B A. *Using differential introgression in hybrid zones to identify genomic regions involved in speciation*. Molecular Ecology Resources. 2010; 10: 806–20.

21. Pritchard JK, Stephens M, Donnelly P. *Inference of Population Structure Using Multilocus Genotype Data.* Genetics. 2000; 155(2): 945-959.

22. Puritz JB, Hollenbeck CM, Gold JR. *dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms.* PeerJ. 2014; 2:e431.

23. Ramsey J, Bradshaw H, Schemske D. *Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M. cardinalis (Phrymaceae)*. Evolution (NY). 2003; 57(7): 1520–34.

24. Riley, H.P. *A character analysis of colonies of Iris fulva, Iris hexagona var. giganticaerulea and natural hybrids*. American Journal of Botany 1938; 25, 727-738.

25. Rundle HD, Nosil P. *Ecological speciation.* Ecology Letters. 2005; 8(3): 336–52.

26. Small, J.K. and Alexander, E.J. *Botanical interpretation of the Iridaceous plants of the Gulf States*. New York Botanical Garden Contribution. 1931; 327, 325-357.

27. Taylor SJ, Arnold M, Martin NH. *The genetic architecture of reproductive isolation in Louisiana irises: Hybrid fitness in nature*. Evolution (NY). 2009; 63(10): 2581–94.

28. Taylor SA, Curry RL, White TA, Ferretti V, Lovette I. *Spatiotemporally consistent genomic signatures of reproductive isolation in a moving hybrid zone*. Evolution (NY). 2014; 3066–81.

29. Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK. *The variable genomic architecture of isolation between hybridizing species of house mice.* Evolution (NY). 2010; 64: 472–85.

30. Viosca P., Jr. *The irises of southeastern Louisiana-a taxonomic and ecological interpretation.* Bulletin of the American Iris Society 1935; 57, 3-56

31. Wesselingh RA, Arnold ML. *Pollinator behaviour and the evolution of Louisiana iris hybrid zones.* Journal of Evolutionary Biology. 2000 Mar; 13(2):171–80.

32. Wu CI. *The genic view of the process of speciation.* Journal of Evolutionary Biology. 2001; 14(6): 851–65.