THE GENOMICS OF SPECIATION

by

Katherine L. Bell, B.S., M.S.

A dissertation submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Aquatic Resources
and Integrative Biology
December 2018

Committee Members:

Chris C. Nice, Chair

Noland Martin

James Ott

James Fordyce

C. Darrin Hulsey

COPYRIGHT

by

Katherine L. Bell

2018

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

Duplication Permission

As the copyright holder of this work I, Katherine L. Bell, refuse permission to copy in excess of the "Fair Use" exemption without my written permission.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Chris Nice, and all of my committee members for their advice and support.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES vi

LIST OF FIGURES vii

ABSTRACT xii

CHAPTER

1. PATTERNS OF INTROGRESSION IN A NORTH AMERICAN PITCHER PLANT HYBRID ZONE 1

2. POPULATION GENOMIC EVIDENCE REVEALS SUBTLE PATTERNS OF DIFFERENTIATION IN THE TROPHICALLY POLYMORPHIC CUATRO CIÉNEGAS CICHLID, *HERICHTHYS MINCKLEYI* 29

3. THE GENOMIC ARCHITECTURE OF JAW POLYMORPHISM IN THE CUATRO CIÉNEGAS CICHLID (*HERICHTHYS MINCKLEYI*) 50

4. HOW MUCH OF GENOMIC DIFFERENTIATION IS REPEATABLE?: A CONTINENT- AND GENOME-WIDE COMPARISON OF PATTERNS 70

REFERENCES 95

LIST OF TABLES

Table **Page**

1.1 Hyper parameter estimates for morphological traits from GEMMA . . 27

1.2 Hyper parameter estimates for morphological traits from GEMMA . . 28

3.1 Number of *H. minckleyi* Individuals Sampled from Cuatro Ciénegas Valley 68

3.2 Hyper Parameter Estimates from GEMMA 69

4.1 Sample information for 67 *Lycaeides* collection localities. Locality numbers, names, nominal species designations, Subgroup assignment (see text for details) latitude, longitude, and number of individuals are provided. The last column indicates previously published sequence data (G&C) (Gompert et al., 2014; Chaturvedi et al., 2018) or sequence data presented here for the first time (Present). 90

LIST OF FIGURES

Figure **Page**

1.1 Example of *S. rubra* and *S. minor* phenotypes and list of 6 traits measured in the hybrid zone. 17

1.2 Plot of genome average admixture proportions (q), each bar represents one individual. 17

1.3 Plot of genome average admixture proportions (q) by inter-source ancestry estimates (Q).
Dark green triangles = allopatric *S. rubra*, light green triangles = allopatric *S. minor*, small green circles = hybrids 18

1.4 Plot of genomic cline center (α) vs. genomic cline width (β). Black contour lines show kernal density estimate. 19

1.5 Plot of genomic cline center (α) and F_{ST} for SNPs with significantly negative α's, and significantly positive α's.
Black contour lines show kernal density estimate. 19

1.6 Plot of genomic cline width (β) and F_{ST} for SNPs with significantly negative α's, and significantly positive α's.
Black contour lines show kernal density estimate. 20

1.7 Plot of morphology trait scores against admixture proportions estimated from ENTROPY. 20

1.8 Plot of z-transformed invertebrate group abundance against admixture proportions estimated from ENTROPY. 21

1.9 Posterior inclusion probabilities from the BSLMM models, dark purple=leaf scape score, medium purple=fenestration, light purple = hood apex, light green = hood angle, medium green=wing shape, dark green=tube shape. 21

1.10 Posterior inclusion probabilities multiplied by β from the BSLMM models, dark purple=LA, medium purple=F, light purple=HAp, light green=HA, medium green=WS, dark green=TS. 22

1.11 Posterior inclusion probabilities from the BSLMM models, darkest purple = Coleoptera, dark purple = Collembola, purple = Diptera, light purple = Hemiptera, lightest green = Hymenoptera, light green = Formicidae, green= Orthoptera, dark green = Arachnids, darkest green = grubs. 22

1.12 Posterior inclusion probabilities multiplied by β from the BSLMM models, darkest purple = Coleoptera, dark purple = Collembola, purple = Diptera, light purple = Hemiptera, lightest green = Hymenoptera, light green = Formicidae, green= Orthoptera, dark green = Arachnids, darkest green = grubs. 23

1.13 For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β.) TS = tube shape, WS = wing shape, HA = hood angle, HAp = hood apex, F = fenestration, LSS = leaf scape score. Black contour lines show kernal density estimate. 24

1.14 For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β). Black contour lines show kernal density estimate. 25

1.15 For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β). Black contour lines show kernal density estimate. 26

2.1 Map of sampling locations in the Cuatro Ciénegas valley, Coahuila Mexico. ©Google 2018. Purple = individuals of both morphotypes from Escobedo, green = individuals of both morphotypes from Juan Santos, orange = *H. cyanoguttatus* from Rio Salado. 47

2.2 Principal component analysis (PCA) of genotype probabilities. Each dot represents an individual, whose position in ordination space is determined by it's multi-locus genotype. Orange = *H. cyanoguttatus*, squares = papilliform, circles = molariform, triangles = intermediate, purple = *H. minckleyi* from Escobedo, green = *H. minckleyi* from Juan Santos A: *H. cyanoguttatus* and *H. minckleyi* common (6,220) SNP's, B: *H. minckleyi* common (6,587) SNP's. 48

2.3 Barplot of admixture proportions for k=2 model in ENTROPY for *H. cyanoguttatus* and *H. minckleyi*. The top barplot shows admixture proportions for common genetic variants, the bottom graph shows admixture proportions for rare genetic variants. Each bar represents one individual's assignment proportions to each of the 2 source populations.
Number of individuals differs between common and rare data sets due to differences in coverage (see text). Numbers underneath bars represent sampling location and morphotype: 1 = *H. cyanoguttatus*, 2 = *H. minckleyi* intermediate from Escobedo, 3 = *H. minckleyi* molariform from Escobedo, 4 = *H. minckleyi* molariform from Juan Santos, 5 = *H. minckleyi* papilliform from Escobedo, 6 = *H. minckleyi* papilliform from Juan Santos 48

2.4 Barplot of admixture proportions for k=2 through k=5 in ENTROPY based on common genetic variants for the minckleyi data set. Each bar represents one individual's assignment proportions to each of k source populations.
Numbers underneath bars represent sampling location and morphotype: 1 = Intermediate from Escobedo, 2 = Molariform from Escobedo, 3 = Molariform from Juan Santos, 4 = Papilliform from Escobedo, 5 = Papilliform from Juan Santos 49

2.5 RDA plot demonstrating the relationship between *H. minckleyi* common genetic variants, geographic location, and standardized tooth size.
Squares = papilliform, circles = molariform, triangles = intermediate, purple/dark grey = *H. minckleyi* from Escobedo, green/light grey = *H. minckleyi* from Juan Santos. The arrow represents the vector for standardized tooth size. 49

3.1 Figure provided by C. Darrin Hulsey. Papilliform pharyngeal jaw (A) and molariform pharyngeal jaw (B) of *Herichthys minckleyi*. The pharyngeal tooth size of *H. minckleyi* (C) has a bimodal distribution and individuals with intermediate pharyngeal morphology are rare. 63

3.2 Principal component analysis of genotype probabilities estimated from ENTROPY. Dark colors = molariform individuals, light colors= papilliform individuals. Tierra Blanca = blue, Tio Candido = green, Escobedo = red, Juan Santos = orange, Mojarral Este = purple. . . . 64

3.3 Barplot of admixture proportions estimated from ENTROPY for k = 2 through k = 5. 65

3.4 Violin plots showing estimates of proportion of variance explained, proportion of genetic variance explained by SNPs with sparse effect, number of SNPs with sparse effect.
TS = tooth size, SL = standard length, Prot = jaw protrusion, Gut = gut length, Gape = gape, AscPro = ascending promaxilla length. . 66

3.5 Posterior Inclusion Probabilities (PIP's) from BSLMMs conducted using GEMMA for six traits of interest.
Ascending premaxilla (AscPro) = dark brown, gape = medium brown, gut length = light brown, protrusion = light purple, standard length = medium purple, standardized tooth size = dark purple. 67

3.6 Posterior Inclusion Probabilities (PIP's) multiplied by β from BSLMMs conducted using GEMMA for six traits of interest.
Ascending premaxilla (AscPro) = dark brown, gape = medium brown, gut length = light brown, protrusion = light purple, standard length = medium purple, standardized tooth size = dark purple. 67

4.1 Map of sampling locations across North America. Full details for localities can be found in Table 4.1. 84

4.2 Principal Component Analysis (PCA) on genotype probabilities across 21,166 SNPs. Each data point represents one individuals genotype probabilities across all 21,166 SNPs.
Data points are color coded by their *Lycaeides* lineage. Light purple = Karners center, dark purple = Karners edge, darkest blue = Whites/Sierra hybrids, blue = Jackson hybrids, light blue = Warners hybrids, lightest blue = Alpines, orange = *L. anna*, pink = *L. idas*, dark green = *L. melissa* Rockies, green = *L. melissa* west, light green = *L. melissa* east, red = *L. ricei*. 85

4.3 Plot of admixture proportions for k = 2 through k=5. 85

4.4 Plot of admixture proportions for k = 6 through k=9. 86

4.5 Estimates of Genome-Wide F_{ST} for pairwise comparisons between all evolutionary lineages, across all SNPs with minor allele frequency $\geq$ 0.5% (6,245). 86

4.6 Correlation between pairwise F_{ST} and population estimates of θ. . . . 87

4.7 Correlation between pairwise F_{ST} and population estimates of π for each linkage group. 88

4.8 Correlation between population estimates of θ and π for each linkage group. 89

ABSTRACT

Speciation, the process by which reproductive isolation evolves between diverging lineages, is pivotal to our understanding of evolution. Across multiple wild populations I explored the genetic architecture of reproductive isolation and adaptive traits, the interaction between gene flow and genetic architecture of traits and their impact on the process of speciation, and finally I assessed the repeatability of genetic differentiation and absolute diversity across the genome, across multiple species pair comparisons. My dissertation includes investigations of hybridization between pitcher plants *(Sarracenia sp.)*, a repeated trophic polymorphism within the Cuatro Ciénagas cichlid fish *(Herichthys minckleyi)*, and a species complex of blue butterflies *(Lycaeides sp.)* that have a complicated evolutionary history that includes repeated, independent evolution of hybrid species. I generated genome-wide population genetic data to quantify patterns of genomic differentiation in all of these case studies. I used a combination of analyses to dissect the relationships between trait architecture, adaptation, and reproductive isolation. Bayesian clustering was used to describe patterns of variation and identify areas of admixture. Bayesian Sparse Linear Mixed Models (BSLMM) were used to map the genetic architecture of a variety of traits and I compared estimates of introgression for genomic regions that contribute to trait variation to understand if these traits are associated with fitness in admixed individuals. Bayesian Genomic Clines models were used to identify patterns of introgression and excess ancestry in admixed individuals. Patterns of

differentiation measured along chromosomes was used to assess the repeatability of differentiation and potential adaptation. I found remarkable variation in trait architecture, ranging from very simple to highly complex. Many genomic regions were associated both with trait variation and patterns of strong selection, though this was not universal. Repeatable patterns were detected in some regions of the genome which suggests that evolution can be predictable, yet there are also instances of unrepeated differentiation suggesting a role for historical contingency. Overall, my results contribute to our understanding of the process of speciation and highlight the power of genome-wide data to resolve important questions in evolution.

1. PATTERNS OF INTROGRESSION IN A NORTH AMERICAN PITCHER PLANT HYBRID ZONE

Introduction

Speciation, the process by which reproductive isolation evolves between diverging lineages, is pivotal to our understanding of evolution. Speciation occurs via an accumulation of both pre- and post-zygotic isolating barriers which impede gene flow between lineages (Coyne and Orr, 2004). From the time when Darwin first formally introduced the idea of speciation, our understanding of this process has advanced dramatically (Darwin, 1859; Mayr, 1963; Bush, 1994). Recent theoretical and empirical research into speciation argues that evolutionary outcomes can be complex, and species boundaries may be porous or semi-permeable (Gompert et al., 2012a; Wu, 2001; Harrison and Larson, 2014; Schluter, 2009; Nosil, 2012). This raises important questions about the genomic architecture of reproductive isolation and the role of stochastic (such as genetic drift) vs deterministic (such as natural selection) evolutionary processes.

In hybrid zones, when reproductively isolated species come into contact, patterns of introgression, natural admixture, and recombination across the genome provide information about the maintenance of species boundaries and can be used to map the genomic architecture of reproductive isolation (Barton and Hewitt, 1985; Baack and Rieseberg, 2007; Abbott et al., 2013; Hvala et al., 2018; Gompert and Buerkle, 2016; Payseur and Rieseberg, 2016). Hybrid zones can have complex and important evolutionary outcomes. Hybridization may lead to

the break down of barriers to gene flow, resulting in a loss of differentiation (Rhymer and Simberloff, 1996), contrastingly there may be an increase in the strength of reproductive barriers via processes such as reinforcement (Servedio and Noor, 2003; Wu, 2001). Introgression may introduce novel phenotypes that contribute to adaptive divergence between populations (Borge et al., 2005; Whitney et al., 2010), or may result in the formation of new, admixed populations which are reproductively isolated from both parental species (Mallet, 2007). With continuing advances in sequencing technology and statistical models it is now possible to not only approach the study of hybridization at the genomic level, but also to study naturally occurring hybrid zones in non-model systems. In these systems we can identify regions that show either under or over representation in an alternate genomic background, referred to as excess ancestry or introgression.

Exploring introgression in hybrid zones can provide insight into regions of the genome that are responsible for maintaining reproductive isolation (i.e. barrier loci), or regions associated with fitness. We can use patterns of differential introgression to explore how many regions of the genome are associated with reproductive isolation or fitness in hybrid zones and ask what these patterns tell us about the process of speciation and the nature of species boundaries (Abbott et al., 2013; Gompert et al., 2017). Patterns of differential introgression are relatively common in hybrid zones; introgression may be affected by stochastic processes such as drift, therefore increased or decreased introgression should be interpreted carefully (Gompert et al., 2012b). Regions that are associated with reproductive isolation should have reduced introgression within an hybrid zone

compared to genome wide patterns. Additionally introgression is often restricted in regions of the genome that are rearranged or have reduced rates of recombination, this could be explained by increased effects of selection on linked loci when recombination is restricted (Baack and Rieseberg, 2007; Gompert et al., 2017). Introgression may be asymmetric toward one parental species if alleles from one species confer higher fitness (Gompert et al., 2017). If regions of the genome that are highly differentiated in parental species also show restricted introgression in hybrids this could indicate that these regions contain barrier loci, those that contribute directly to reproductive isolation. If however these highly differentiated regions show introgression in hybrids then either they are not adaptive or selection may be context dependent. For example, in a hybrid zone between two species of butterfly from the genus *Lycaeides* highly differentiated regions were shown to have more excess ancestry for one parental species (*L. idas*), compared to the other. This indicates that highly differentiated loci affect fitness of the hybrid and parental species in ways which depend on habitat or genomic background, hybrids in this study were found in habitat more similar to *L. idas* and had a higher proportion of *L. idas* ancestry (Gompert et al., 2012a). Differences in introgression for regions that are associated with the genetic architecture of adaptive traits can provide insight into the influence of these traits on fitness in hybrids or their role in maintaining species boundaries. Pitcher plants of the genus *Sarracenia* commonly form hybrid zones and present the opportunity to explore the genomic architecture of reproductive isolation and adaptation in a natural setting. Pitcher plants are unique in the plant kingdom in their production of modified leaves that form hollow, water containing vessels

which are used to trap invertebrate prey(McPherson, 2007). These traps are thought to attract prey via nectar, scent and coloration. Due to their ability to obtain nutrients from their prey pitcher plants can survive in hostile environments. *Sarracenia* is endemic to wet pine savannah, seepage slopes, and fens in North America and is distributed primarily throughout the southeastern United States (with the exception of one species) (Stephens et al., 2015). Diversification of *Sarracenia* is thought to have occurred less than 3 million years ago during the Pleistocene (Ellison et al., 2012). Hybridization between species is relatively common where species boundaries overlap. All *Sarracenia* are inter-fertile and self-fertile and species are pollinated by the same insects, primarily bees (McPherson, 2007). The hybrid zone we explore is between *Sarracenia rubra*, commonly referred to as the sweet pitcher plant, and *S. minor*, the hooded pitcher plant. Both species are found in North and South Carolina, Florida, and Georgia while *S. rubra* is also found from Alabama to the southeastern edge of Mississippi but with a fragmented range (McPherson, 2007). We explore patterns of introgression and admixture in this *S. minor* and *S.rubra* hybrid zone and ask four specific questions: 1) What is the distribution of hybrids within the hybrid zone? 2) Are there patterns of excess ancestry? 3) What proportion of loci that show excess ancestry also show high levels of differentiation between parental species? 4) What proportion of loci associated with phenotypic traits also show excess ancestry?

Methods

Collection and Sampling

We sampled 60 hybrids (*S. rubra* X *S. minor*) from a site in Francis Marion National Forest in South Carolina (latitude: 33.08, longitude: -79.7). Tissue was taken for genetic analyses, prey contents of traps was recorded, and morphological measurements were made. For each morphological trait, an individual was scored on a scale form zero to ten, individuals with a pure *S. rubra* trait were scored zero, while individuals with a pure *S. minor* like trait were scored a ten. In total 6 traits were measured; tube shape, wing shape, hood angle, hood apex, fenestration, and leaf scape ratio. *S. minor* individuals have a funnel shaped tube, the wing is wider in the middle of the leaf, the angle of the hood is arching and the apex is acute, fenestration (translucent areas on the leaf) are present, and the scape height is lower than the leaf height. In contrast, for *S. rubra* tube shape is tubular, the wing is wide below the middle of the leaf, the hood angle is suberect, hood apex is acuminate, fenestration is not present, and scape height is taller than leaf height (Figure 1.1). For the measurements of prey abundance the contents of each pitcher plant trap was removed and individuals were classified as belonging to one of 11 invertebrate groups (Table 3.2, *Details need to be clarified with A. Strand*). The location of the hybrid zone is fairly disturbed, and consists of a boggy area with areas of higher dry ground and a gas and power road has been added. The site appears to have been continually disturbed since at least 1989 based on historical satellite imagery. Tissue samples for 30 individuals from the two parental species were collected at

locations within 4km of the hybrid site.

Molecular Methods

We prepared reduced representation genotype-by-sequencing libraries for each individual following the protocol of Parchman et al. (2012) and Gompert et al. (2012a). In brief, for each individual, DNA was fragmented using EcoR1 and Mse1, we then ligated sequencing adaptors and unique 8-10 base pair multiplex identifier sequences (barcodes sequences), conducted two rounds of PCR and then used BLUE PIPIN (Sage Science) to size select fragments (between 350-450bp). Individuals were sequenced across 2 lanes of Illumina HiSeq 4000 technology at the University of Texas Genome Sequencing and Analysis facility (GSAF). This resulted in just over 339 million reads. In order to remove contaminants, raw reads were assembled to the PhiX genome using BOWTIE (Langmead, 2010). Sequences that did not align to the PhiX genome were used in all further analyses. We used custom perl scripts to remove barcode sequences and carried out a *de novo* assembly following the dDocent protocol with minor modifications (Puritz et al., 2014a,b). The dDocent *de novo* assembly resulted in 125,079 scaffolds. These scaffolds were used in the reference based assembly. We used BWA SAMSE and ALN to align to the *de novo* assembly. We used a combination of custom perl scripts, SAMTOOLS, and BCFTOOLS to call variants and required 80% of individuals to have data at a site in order for a variant to be called. We removed variants that were only present in a single individual as these may be due to sequencing error. This resulted in 38,882 single nucleotide polymorphisms (SNPs).

Population Genomics, Introgression, and Admixture Mapping

Estimates of genotype likelihoods from BCFTOOLS were used in the program ENTROPY to obtain estimates of genome average admixture (q), genotype probabilities, and admixture class (Q) frequency which estimates how much of the genome is heterozygous for ancestry (inter-source ancestry) vs. homozygous for ancestry (intra-source ancestry) (Gompert et al., 2014). Estimates of inter-source ancestry provide information about the type of hybrids present, recent hybrids with non-admixed parents should have high inter-source ancestry, while late generation hybrids that represent a stable hybrid lineage should have low inter-source ancestry (Gompert et al., 2014; Buerkle and Lexer, 2008; Gravel, 2012). The program ENTROPY was developed by Gompert et al. (2014) and implements a Bayesian hierarchical model similar to that used in STRUCTURE but takes into account sequence and alignment error (Pritchard et al., 2000; Falush et al., 2003). Given our focus on the hybrid zone and two parental species we ran the model for k=2. Posterior probability estimates for each parameter were obtained using Markov Chain Monte Carlo (MCMC). We ran two chains for 170,000 steps, with a burn in of 25,000, saving every 20th step. We checked that the model had reached convergence and stabilization by estimating effective sample size and Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992) in R using the package CODA (Plummer et al., 2006). To visualize the relationship between sampling groups we conducted a principle component analysis (PCA) on genotype probability estimates using the PRCOMP function in R (R Core Team, 2016). To explore the type of hybrids present in the hybrid zone we plotted genome average admixture estimates (q)

against estimates of admixture class (Q).

We estimated locus-specific genomic introgression using Bayesian Genomic Cline model (BGC) developed by Gompert and Buerkle (2011). BGC quantifies locus-specific patterns of introgression using two parameters; α the genomic cline center and β the genomic cline width. For locus i, α describes the increase or decrease in probability of ancestry relative to hybrid index. For locus i, β describes the increase or decrease in rate of transition from low to high probability of ancestry relative to hybrid index (Gompert and Buerkle, 2011). Loci that show extreme patterns of introgression may be associated with adaptive divergence or reproductive isolation. BGC requires the specification of pure parental species, and individuals that are sampled from the hybrid zone. Parameter estimates are obtained using MCMC. We ran the model for four chains, 100,000 steps, a burn in of 25,000 and saved every 10th step. As with ENTROPY we estimated effective sample size and calculated Gelman and Rubin's diagnostic in order to ensure the model had reached a stable sampling distribution. We identified outlier loci as those that had credible intervals that did not overlap zero, and whose median was either below the 2.5% quantile or above the 97.5% quantile in the distribution of α or β across all loci.. Significant loci were classified as those whose credible intervals did not overlap zero. To explore if patterns of selection experienced in the parental species contribute to fitness in the hybrid zone we first calculated locus-specific Wright's F_{ST} in R between the two allopatric parental species samples (Wright, 1943). Then we compared estimates of locus-specific F_{ST} for SNPs identified from BGC as having exceptional patterns of ancestry in the hybrid zone with estimates of

either α or β that were classified as outliers.

We fit Bayesian Sparse Linear Mixed Models (BSLMM) using the program GEMMA to map the genetic architecture of the six morphological traits and the abundance of eleven groups of invertebrate prey measured in the hybrids (Zhou et al., 2013). Linear mixed models and sparse linear models involve different assumptions about the genetic architecture of a trait; it is often unclear which of these assumptions are most appropriate for a data set *a priori*. The BSLMM approach implements a hybrid of these two models which adapts to the genetic architecture detected in the data as the model is run. This approach is able to accurately model data where both a small number of large effect markers (sparse effect) and a large number of small effect markers (polygenic effect) are present. The sparse effect of this model is estimated by the parameter β (hereafter we refer to this parameter as β_{GEMMA} to avoid confusion with β estimated from BGC). β_{GEMMA} estimates the effect of an individual SNP on trait values, a SNP can be pulled into and out of the model - whereby a β_{GEMMA} value of zero effectively removes that SNP from the model. The probability that a SNP is kept in the model is referred to as the posterior inclusion probability (PIP). BSLMM also simultaneously models a polygenic effect μ which can be interpreted as representing the combined effect of a large number of small effects across all measured markers (Zhou et al., 2013). Before running the model we z-transformed the trait measurements (Figure 1.1), and estimated a kinship matrix to remove spurious associations that may result from genetic relatedness. We ran the model separately for each trait and used MCMC to obtain parameter estimates. For each trait the model was run for 1,000,000 steps with a burn in of

100,000 and saved every tenth value. We obtained several parameter estimates but in particular we are interested in exploring the posterior inclusion probability (PIP) and β_{GEMMA} estimates of each SNP, and comparing those SNPs associated with a trait to the SNPs which show excess ancestry as estimated from BGC.

Results

We used estimates of genome average admixture proportions (q) and inter-source ancestry (Q) to explore the ancestry of hybrids within the hybrid zone. We found that for genome average estimates of admixture there were hybrid individuals that clustered with the allopatrically sampled parental species and in general hybrid individuals evenly spanned ancestry between *S. minor* and *S. rubra* (Figure 1.2). For the plot of q against Q the maximum possible inter-source ancestry given the global genetic ancestry is shown with the solid grey line (Figure 1.3) and we found that the majority of hybrid individuals fell along this line, indicating that they had at least one non-admixed parent. This provides evidence of back crossing between hybrid individuals and one or other of the parental species. Slightly more hybrid individuals showed genetic similarity to *S. minor*.

Our analysis using the Bayesian genomic clines (BGC) model identified outlier loci that had either α or β estimates where the median was either below the 2.5% quantile or above the 97.5% quantile in the distribution and the credible intervals did not overlap zero. We identified 1,282 loci with positive α's, which indicates introgression of *S. rubra* alleles. We identified 786 loci with negative α's, which indicates introgression of *S. minor* alleles. We identified 20 loci with

positive β's which demonstrates that the rate at which probability of ancestry switches from one parental species to the other is high, creating a narrow cline. We found 195 loci with negative β's which shows a slower rate of transition of probability of ancestry and therefore a wider genomic cline. We found 1,458 SNPs with α values whose credible intervals did not overlap zero, and 786 SNPs with significantly negative α values (the same number as the number of outlier SNPs). For β we found 1,973 with significantly positive values of β and 417 SNPs with significantly negative values. When α was plotted against β we found a significant negative correlation (r=-0.36, n=38,882, P<0.01) (Figure 1.4). This demonstrates that SNPs that show *S. rubra* introgression (positive α) tended to have a wider genomic cline (negative β) and those that show *S. minor* introgression (negative α) tended to have a narrow genomic cline (positive β). A narrow genetic cline indicates that selection may be acting against introgression in these regions, while a wider cline indicates either neutral or positive selection. We compared F_{ST} estimates to α and β estimates for those SNPs with outlier genomic cline values. We might expect that those loci that are differentiated between the two parental species (i.e. high values of F_{ST}) would not introgress in the hybrid zone as they might play a role in maintaining reproductive isolation. Therefore, for those SNPs that have positive α estimates we would expect a negative correlation between α and F_{ST}. However we found that Pearson's correlation calculated between SNPs that had positive outlier α's and F_{ST} and between SNPs with negative outlier estimates of α and F_{ST} were not significant. We did not find any significant relationship between SNPs with outlier β values and F_{ST}.

To visualize the relationship between trait measurements and hybrid individuals we plotted trait values against genome average admixture proportions (q) (Figure 1.7 and 1.8). We used BSLMMs to estimate the genetic architecture of six traits that distinguish the two parental species *S. rubra* and *S. minor*, and for prey abundance in pitcher traps for 11 groups of invertebrates. We found that all morphological traits appear to have complex architectures, which involve a large number of SNPs, each of relatively small effect. Overall PIP's were relatively low (Figure 3.5) and proportion of variance explained (PVE) ranged from 0.09 to 0.35 (Table 1.1). The models that we fit for the abundance of 11 invertebrate prey species found a very different pattern, we found that for each trait a small number of SNPs showed very high values of PIP, and many also had large effect sizes. This indicates a relatively simple genetic architecture underlies the unmeasured trait or traits associated with prey capture. For each trait we identified the SNPs which had PIP's in the top 2.5% of estimates. We explored the relationship between SNPs with top PIPs and F_{ST}, genomic cline center (α), and the genomic cline width (β) using density plots and by estimating Pearson's correlation coefficient. Overall we did not find a clear relationship between SNPs associated with the genetic architecture of traits and differentiation in parental species or with introgression in the hybrid zone.

Discussion

We used a natural hybrid zone to explore patterns of genomic introgression and differentiation in order to understand the genetic architecture and maintenance of reproductive isolation in two species of pitcher plants. Recent theoretical and

empirical work has demonstrated that species boundaries may be more complex and porous than once thought, natural hybrid zones can provide insight into the process of speciation and recent technical advances provide the opportunity to study reproductive isolation at a genomic scale (Gompert et al., 2012a, 2013; Wu, 2001; Harrison and Larson, 2014; Schluter, 2009; Nosil, 2012; Abbott, 2017; Mandeville et al., 2015).

The hybrid individuals we sampled spanned a nearly continuous range of genome average ancestry (q), from putatively pure *S. minor* individuals to putatively pure *S. rubra* individuals (Figure 1.2). When q was plotted against inter-source ancestry (Q) we found that the majority of hybrid individuals likely had at least one non-admixed parent (Figure 1.3). This suggests that hybrid individuals rarely reproduce with other hybrids, and that the majority of hybrids in this zone are likely back-crossed. This could indicate relatively recent establishment of this hybrid zone. The presence of what appear to be F1 hybrids (individuals with inter-source ancestry of 1 and genome average ancestry of 0.5) clearly show that hybridization is ongoing between *S. minor* and *S. rubra* at this site (Figure 1.3).

We found that a relatively small proportion of SNPs showed excess introgression. For the genomic cline center estimate (α), 3.297% of SNPs had estimates of α that were positive indicating *S. rubra* introgression while 2.82% of SNPs had negative values of α, indicating *S. minor* introgression. The habitat of the hybrid zone is relatively disturbed and contains both boggy areas and drier areas where there is higher ground. Typical *S. rubra* habitat contains standing water, while that of *S. minor* prefers slightly drier boggy habitat. The higher levels of introgression in the direction of *S. rubra* may mean that selection is favoring *S.*

rubra genotypes in the hybrid zone. For the genomic cline width estimate (β), we found 0.05% of SNPs had estimates of β that were positive, this represents a narrow genomic cline where there is an increase in the rate of transition from a low to high probability of ancestry relative to hybrid index. We found a higher proportion of SNPs with negative β's (0.5%), indicating that SNPs were more likely to have a wider genomic cline than a narrow one. When we compared estimates of α and β across all SNPs we found we found a significant negative correlation whereby SNPs that show *S. minor* introgression had narrower clines than those that showed *S. rubra* introgression. This could indicate that there is less selection acting upon SNPs with *S. rubra* introgression, or that selection favors *S. rubra* alleles, as there is a decrease in the rate of transition at these sites from low probability of ancestry to high probability of ancestry (Figure 1.4). In order to explore if those regions contributing to fitness in the hybrids also experience selection allopatrically, we compared values of locus-specific F_{ST} calculated between the allopatric populations of the two parental species to values of α and β (Figure 1.5 and 1.6). If those SNPs which are showing high values of F_{ST} in the parental species are associated with the maintenance of reproductive isolation one might expect that they would not show patterns of excess introgression. If however those SNPs are associated with adaptation to a particular habitat, then depending on the selection pressures in the hybrid zone, we might expect to see increased introgression for those regions whereby selection is favoring the phenotype of one of the parental species. We did not find an overall correlation between outlier values for α's or β's and F_{ST}. While we did not find an association in this hybrid zone, this pattern may be variable

depending on the genomic make up of the hybrid individuals and the selective pressures they experience.

In addition to comparing regions with introgression to regions that were highly differentiated in the parental species, we were also interested in understanding their relationship with SNPs associated with trait variation. We fit BSLMMs for each of the six morphological traits, and abundance of prey from 11 invertebrate groups. Overall we found that genetic architecture for morphological traits was highly polygenic, involving a high number of SNPs each with small effect sizes (Figure 3.5). While we obtained accurate estimates of parameters from our model we may have relatively little power to detect the true architecture of these traits for two reasons: first the GWAS was confined to the 60 hybrid individuals due to a lack of morphological measurements in the parents, secondly measurements were estimated on a scale (rather than measured quantitatively) and therefore may lack resolution to detect fine scale differences. For the models that we fit for the 11 invertebrate prey groups we found very simple genetic architectures. The mechanisms used to attract prey to pitcher traps are not fully understood, but are thought to involve scent, nectar composition, coloration of the pitcher, and UV fluorescence. It is possible that one of these mechanisms is involved with the differences in prey that we have identified, but further research would be required to disentangle this relationship and identify the specific mechanism responsible for variation in prey abundances. For both the morphological traits and the prey abundance we did not find any significant correlations between F_{ST}, α, or β and top PIPs (Figure 1.13, 1.14, and 1.15). While there were no overall patterns, we did find a small number of SNPs

associated with a trait which also showed excess ancestry. This indicates that selection may be acting upon this trait in the hybrid zone.

In conclusion, we found the hybrid zone contained a nearly continuous range of admixed individuals. We found that the majority of individuals were back-crossed and evidence that hybridization is still ongoing. Our analysis of introgression identified outlier loci for both the genomic cline center (α) and genomic cline width (β). We found higher levels of introgression in the direction of *S. rubra*, and found clines tended to be wider rather than narrow between loci with exceptional patterns of excess ancestry as estimated by BGC. When we examined the relationship between outlier BGC SNPs and F_{ST} we did not find any evidence that SNPs which were highly differentiated in parental species also showed excess ancestry. This could indicate that highly differentiated SNPs are involved in the maintenance of reproductive isolation, and therefore do not introgress in the hybrid zone. We fit BSLMMs to map the genetic architecture of various traits and found that the morphological traits we measured had complex architectures while the underlying traits associated with the prey abundance of 11 different invertebrate groups had simple architectures. We found a small proportion of SNPs associated with trait architecture showed patterns of excess introgression, indicating that selection may be acting upon these traits in the hybrid zone. Overall these results indicate that introgression is relatively limited and favors *S. rubra* alleles and selection may be acting upon a limited number of regions involved in trait adaptation.

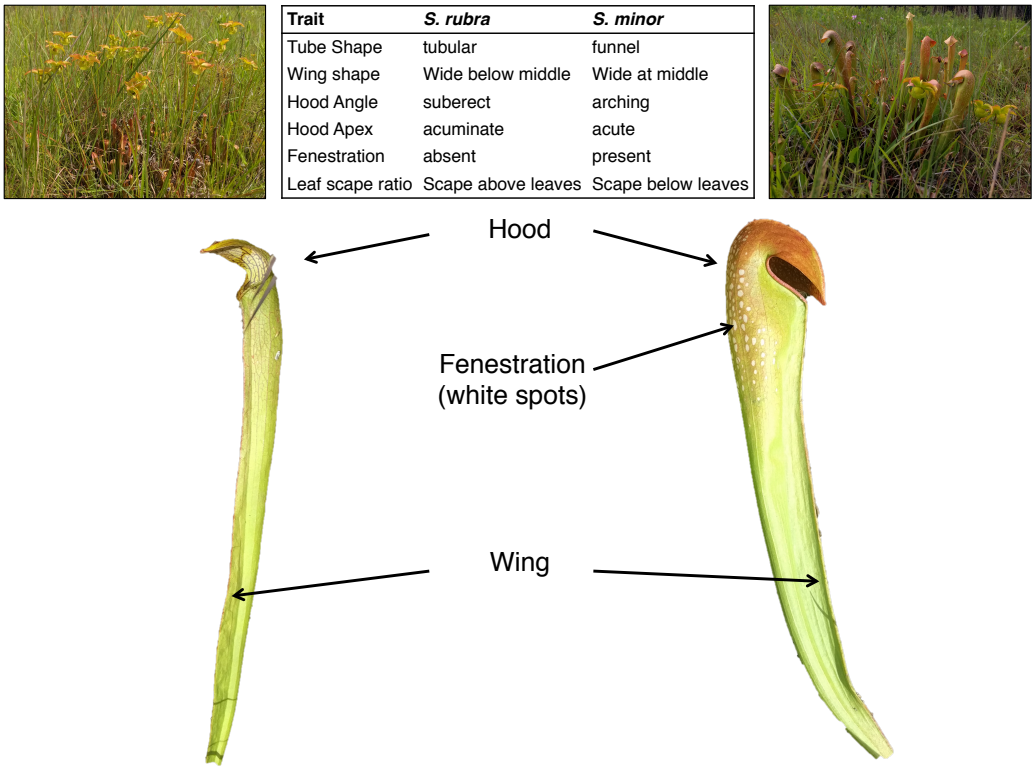

Figure 1.1: Example of *S. rubra* and *S. minor* phenotypes and list of 6 traits measured in the hybrid zone.

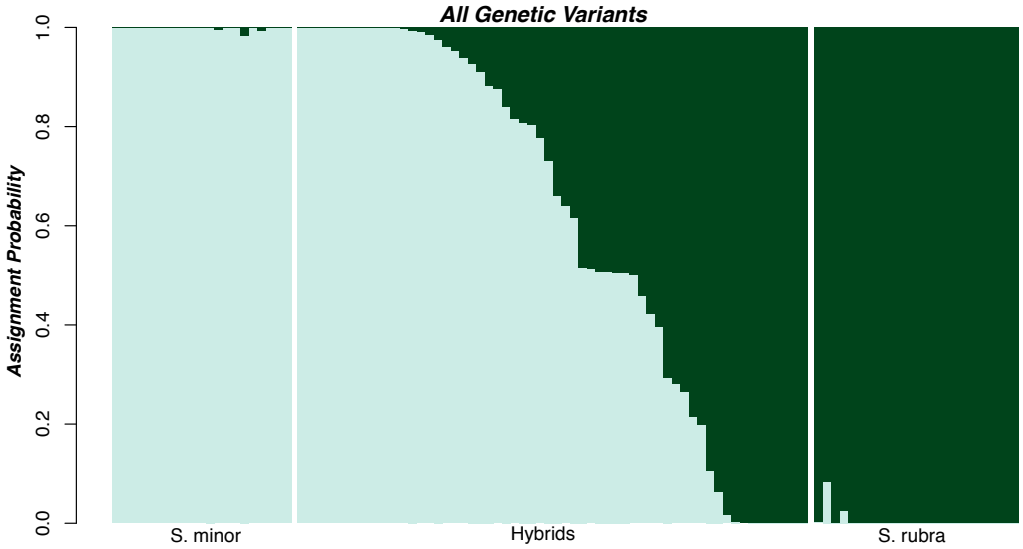

Figure 1.2: Plot of genome average admixture proportions (q), each bar represents one individual.

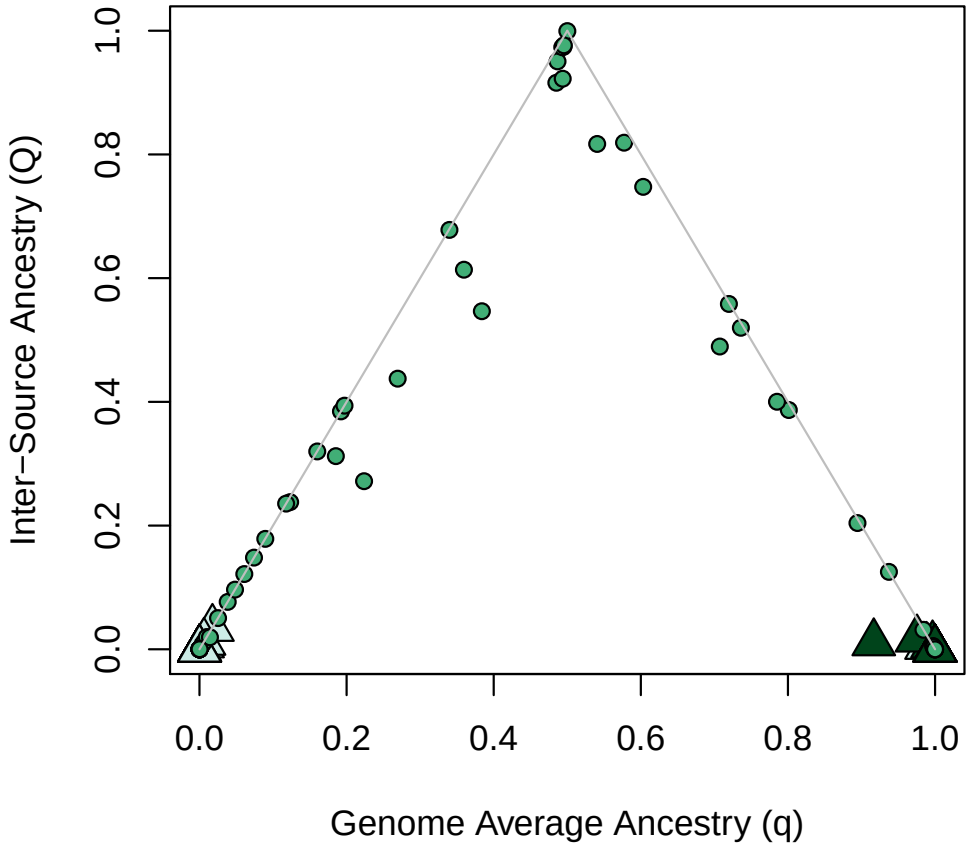

Figure 1.3: Plot of genome average admixture proportions (q) by inter-source ancestry estimates (Q). Dark green triangles = allopatric *S. rubra*, light green triangles = allopatric *S. minor*, small green circles = hybrids

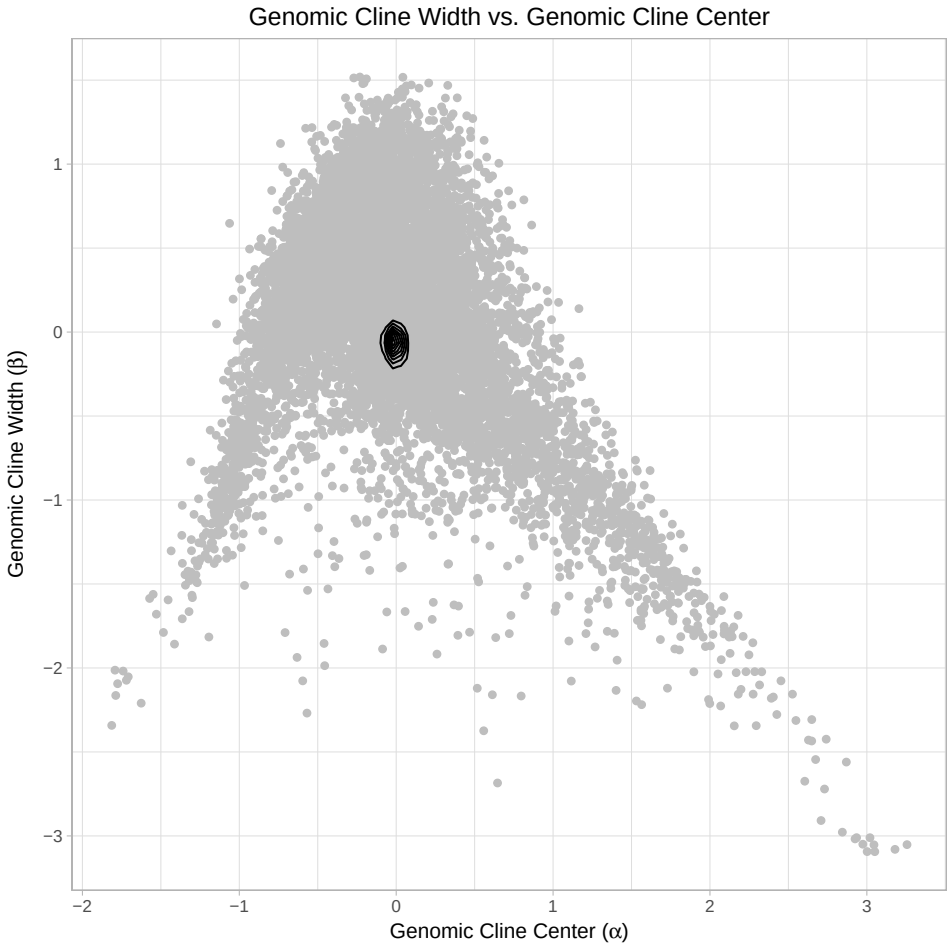

Figure 1.4: Plot of genomic cline center (α) vs. genomic cline width (β). Black contour lines show kernal density estimate.

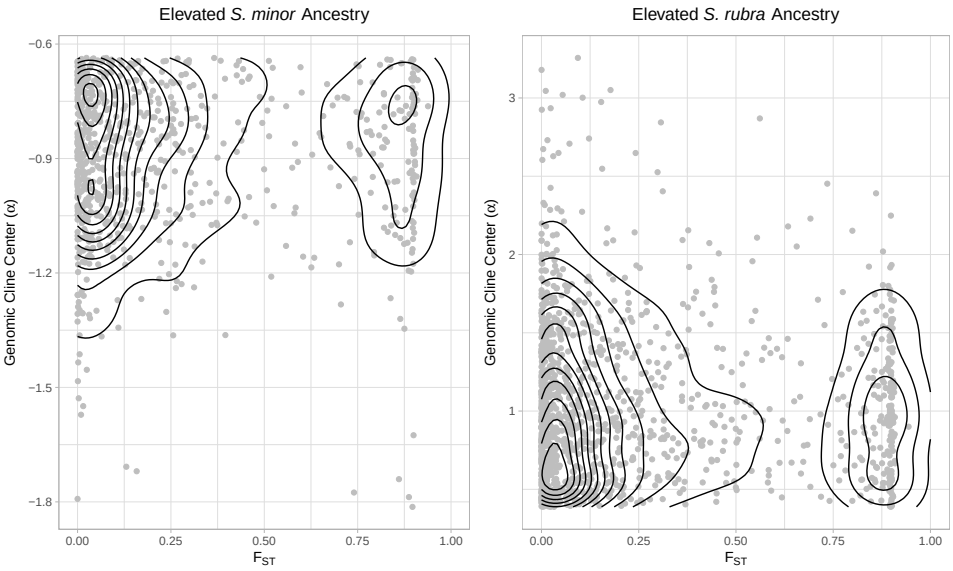

Figure 1.5: Plot of genomic cline center (α) and F_{ST} for SNPs with significantly negative α's, and significantly positive α's. Black contour lines show kernal density estimate.

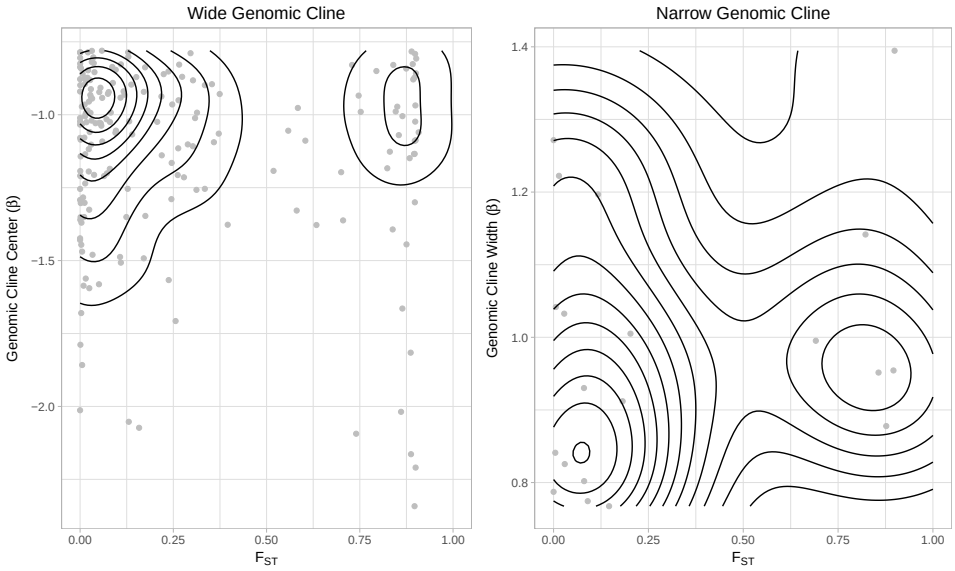

Figure 1.6: Plot of genomic cline width (β) and F_{ST} for SNPs with significantly negative α's, and significantly positive α's. Black contour lines show kernal density estimate.

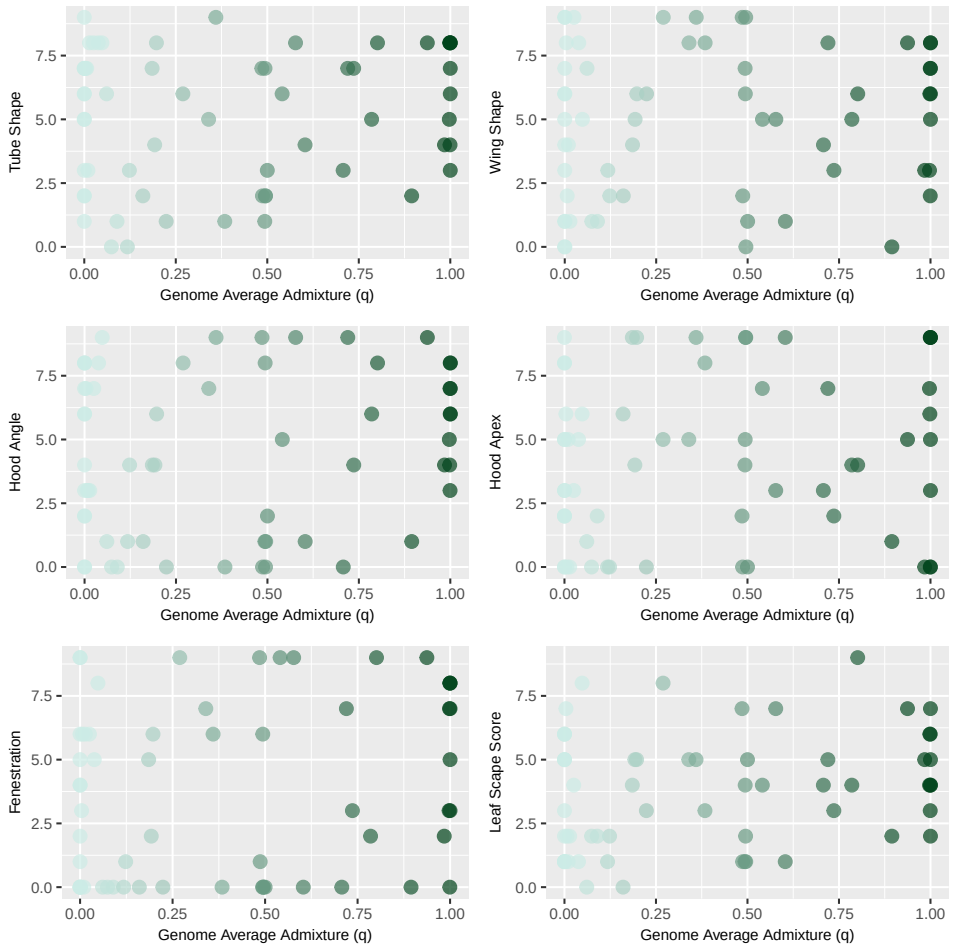

Figure 1.7: Plot of morphology trait scores against admixture proportions estimated from ENTROPY.

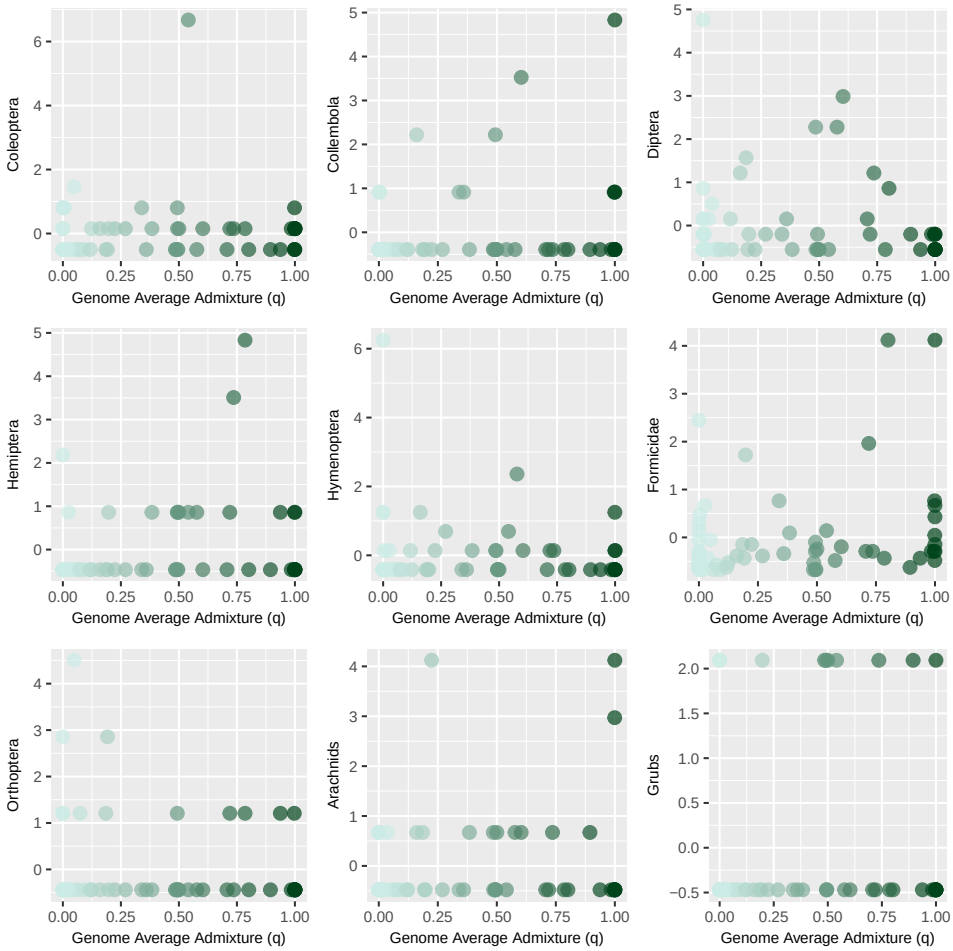

Figure 1.8: Plot of z-transformed invertebrate group abundance against admixture proportions estimated from ENTROPY.

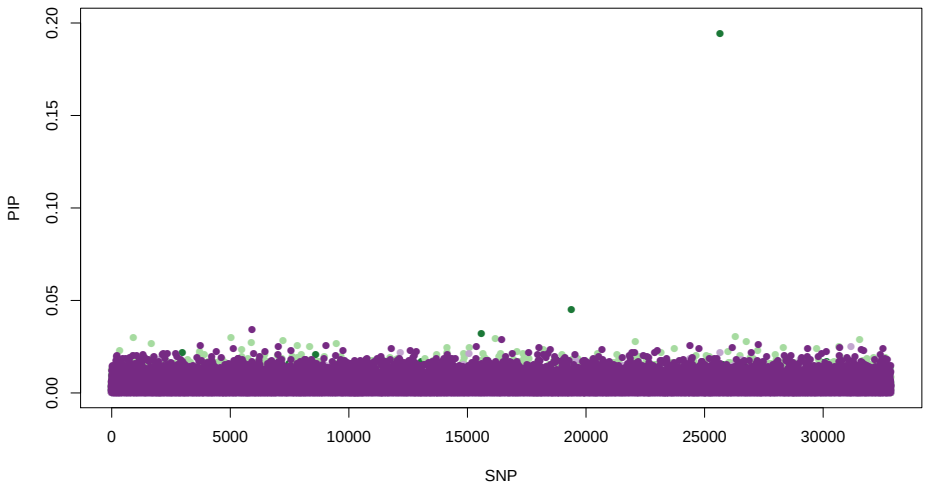

Figure 1.9: Posterior inclusion probabilities from the BSLMM models, dark purple = leaf scape score, medium purple = fenestration, light purple = hood apex, light green = hood angle, medium green = wing shape, dark green = tube shape.

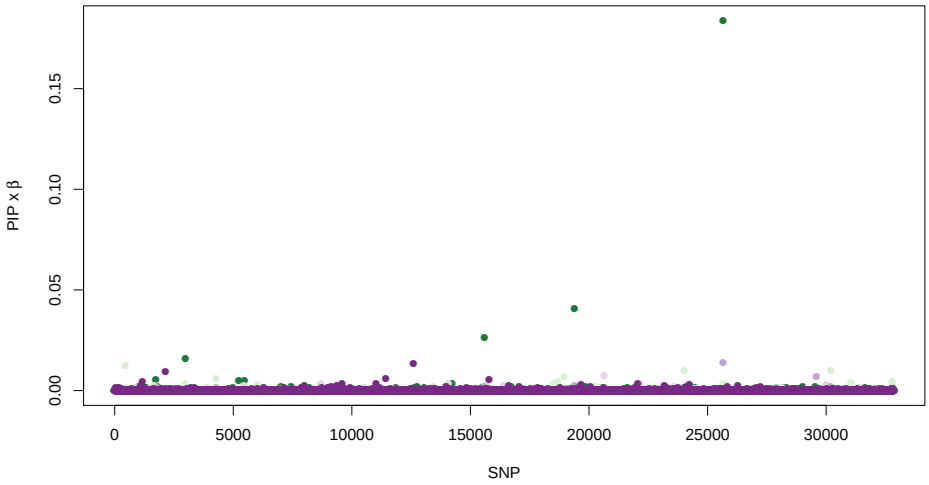

Figure 1.10: Posterior inclusion probabilities multiplied by β from the BSLMM models, dark purple = leaf scape score, medium purple = fenestration, light purple = hood apex, light green = hood angle, medium green = wing shape, dark green = tube shape.

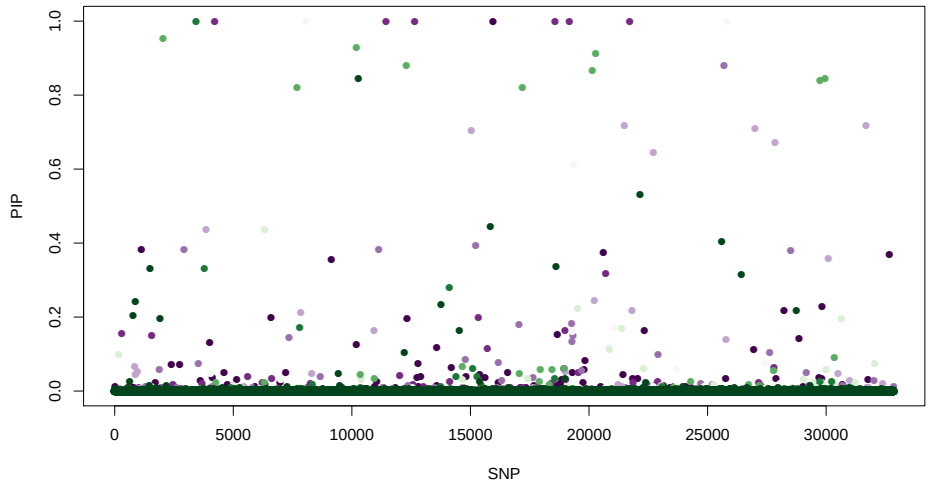

Figure 1.11: Posterior inclusion probabilities from the BSLMM models, darkest purple = Coleoptera, dark purple = Collembola, purple = Diptera, light purple = Hemiptera, lightest green = Hymenoptera, light green = Formicidae, green= Orthoptera, dark green = Arachnids, darkest green = grubs.

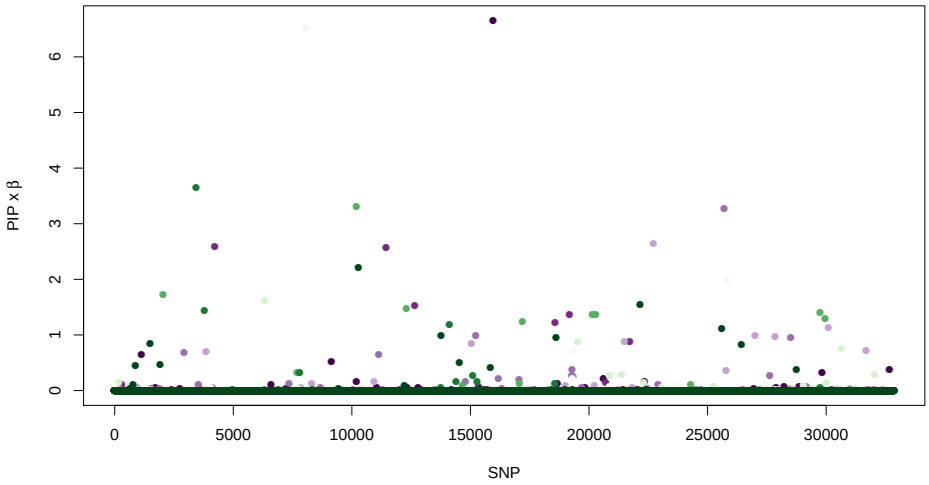

Figure 1.12: Posterior inclusion probabilities multiplied by β from the BSLMM models, darkest purple = Coleoptera, dark purple = Collembola, purple = Diptera, light purple = Hemiptera, lightest green = Hymenoptera, light green = Formicidae, green= Orthoptera, dark green = Arachnids darkest green = grubs.

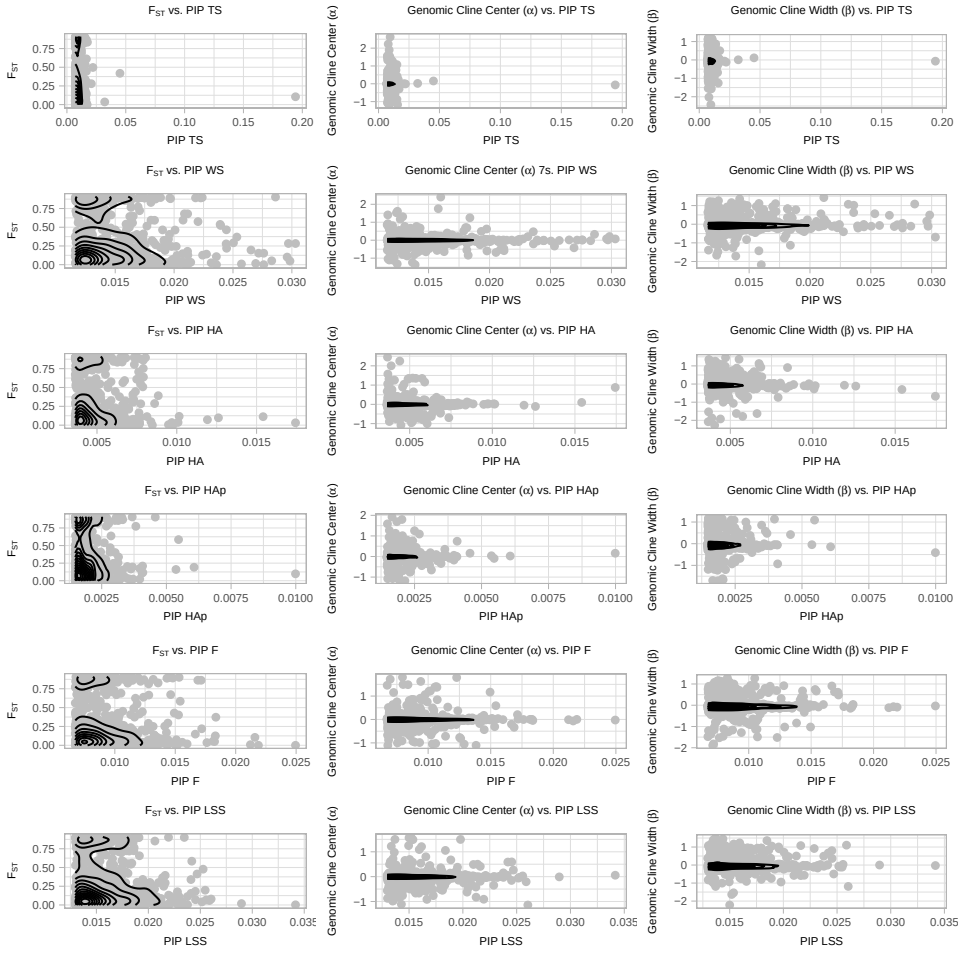

Figure 1.13: For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β.) TS = tube shape, WS = wing shape, HA = hood angle, HAp = hood apex, F = fenestration, LSS = leaf scape score. Black contour lines show kernal density estimate.

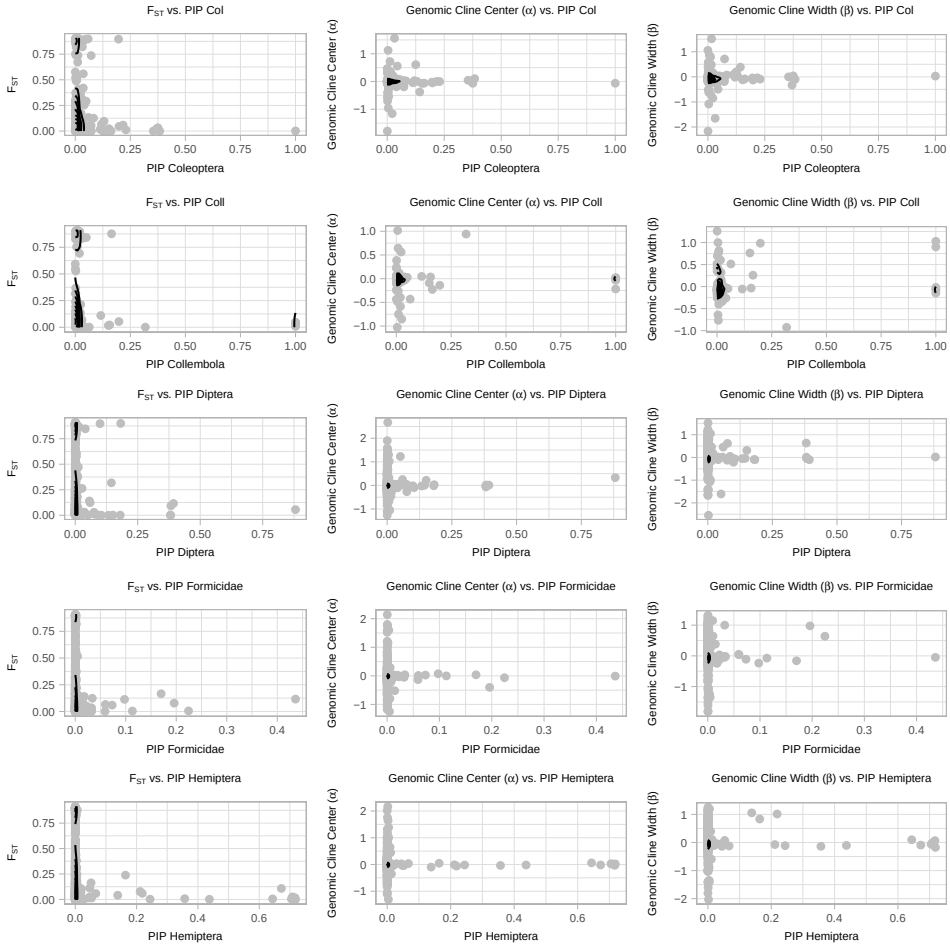

Figure 1.14: For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β). Black contour lines show kernal density estimate.

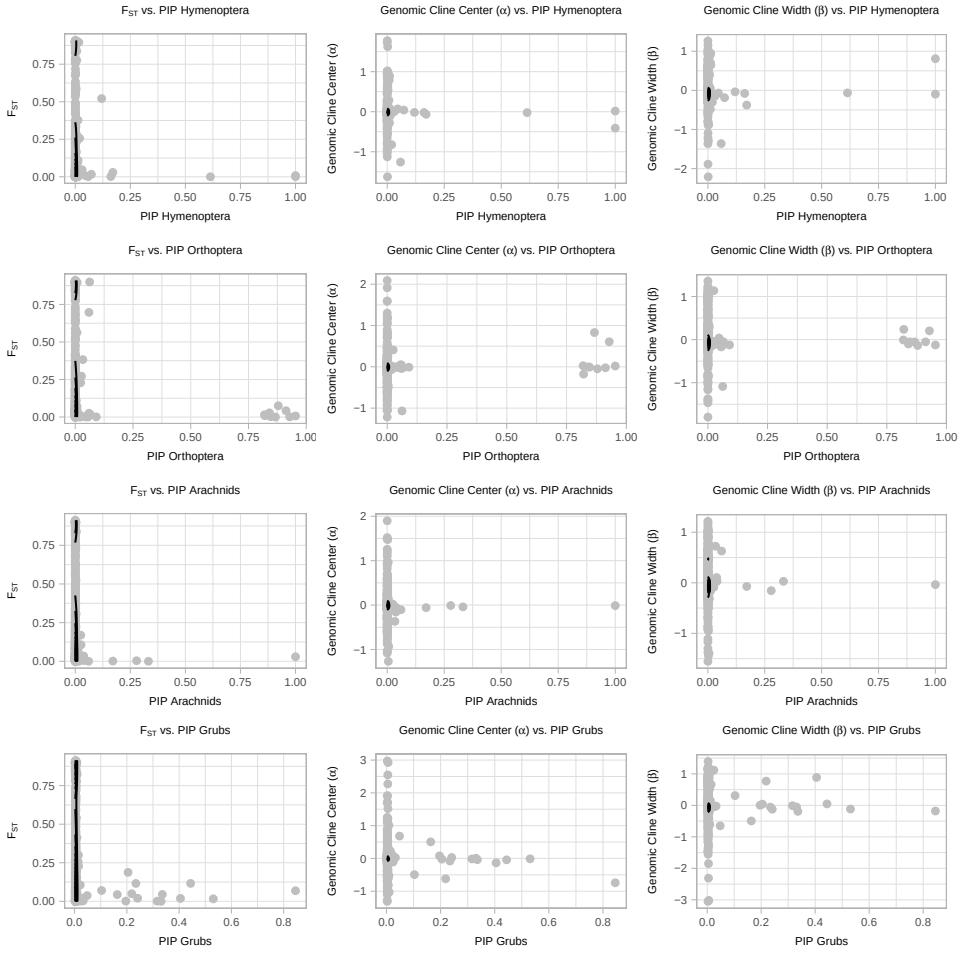

Figure 1.15: For each of the 6 traits, SNPs with the top 2.5% of PIP values plotted against F_{ST}, genomic cline center (α), and genomic cline width (β). Black contour lines show kernal density estimate.

Table 1.1: Hyper parameter estimates for morphological traits from GEMMA]

Hyper Parameter	**Trait**	**Mean**	**Median**	**2.5%**	**97.5%**
Number of SNPs in Model	Tube Shape	60.42878	12	0	284
	Wing Shape	116.612	90	0	284
	Hood Angle	35.32268	19	0	134
	Hood Apex	14.09573	8	0	55
	Fenestration	56.5876	19	0	269
	Leaf Scape Score	149.4537	157	1	296
Proportion of Variance Explained (Sparse Effect)	Tube Shape	0.3202863	0.2647428	0.007280894	0.8850891
	Wing Shape	0.1464671	0.09558705	0.003302213	0.5759683
	Hood Angle	0.2440812	0.1733368	0.006534758	0.8176664
	Hood Apex	0.2023665	0.1470488	0.004369809	0.6872134
	Fenestration	0.2057504	0.1536184	0.006795504	0.6749063
	Leaf Scape Score	0.4171999	0.3523368	0.02489606	0.9910857
Proportion of Variance Explained (Sparse and Polygenic effects)	Tube Shape	0.5308482	0.5685336	0	0.9779126
	Wing Shape	0.4648246	0.4569723	0	0.9690313
	Hood Angle	0.4473741	0.4388361	0	0.9701809
	Hood Apex	0.4164261	0.4028587	0	0.9587071
	Fenestration	0.4379039	0.4190639	0	0.9690269
	Leaf Scape Score	0.4797011	0.4745758	0.006483416	0.9718332

Table 1.2: Hyper parameter estimates for morphological traits from GEMMA

Hyper Parameter	Trait	Mean	Median	2.5%	97.5%
Number of SNPs in Model	Coleoptera	6.33495	6	3	9
	Collembola	7.90532	8	6	10
	Diptera	7.90532	8	6	10
	Hemiptera	6.76874	7	1	10
	Hymenoptera	3.7843	4	3	6
	Formicidae	2.55126	2	1	6
	Orthoptera	8.68566	9	2	11
	Arachnids	2.69048	2	2	5
	Grubs	16.55697	6	4	158
Proportion of Variance Explained (Sparse Effect)	Coleoptera	0.9219189	0.941879	0.7764938	0.9945228
	Colleoptera	0.9744788	0.9717914	0.9440674	0.9999608
	Diptera	0.9029111	0.9369309	0.5360128	0.999989
	Hemiptera	0.9311893	0.9621845	0.6308181	0.9997274
	Hymenoptera	0.94638	0.965508	0.8280959	0.9998575
	Fromicidae	0.6871526	0.6938801	0.4188062	0.9493719
	Orthoptera	0.9499712	0.9832148	0.5820314	0.9994954
	Arachnids	0.7817404	0.781956	0.6195643	0.9378958
	Grubs	0.8247603	0.9263541	0.03293014	0.9973576
Proportion of Variance Explained (Sparse and Polygenic effects)	Coleoptera	0.8635357	0.868832	0.7221262	0.9880326
	Collembola	0.9736944	0.9783448	0.9241641	0.9990435
	Diptera	0.7778372	0.8516214	0.3807374	0.9976342
	Hemiptera	0.9643339	0.9862438	0.8341046	0.9981226
	Hymenoptera	0.9157426	0.9132832	0.7765247	0.9924562
	Formicidae	0.8898544	0.9130659	0.6641823	0.996672
	Orthoptera	0.9824981	0.9981865	0.8169468	0.9996691
	Arachnids	0.9335268	0.9490036	0.7854863	0.9991667

2. POPULATION GENOMIC EVIDENCE REVEALS SUBTLE PATTERNS OF DIFFERENTIATION IN THE TROPHICALLY POLYMORPHIC CUATRO CIÉNEGAS CICHLID, *HERICHTHYS MINCKLEYI*

Introduction

Understanding patterns of genomic variation, and the evolutionary processes that lead to that variation, facilitates our understanding of the speciation continuum. In recent decades a mechanistic approach to the study of evolution and an increased understanding of molecular ecology has led to a reinterpretation of the role of gene flow during the evolution of reproductive isolation and novel biological diversity (Via, 2001; Bolnick and Fitzpatrick, 2007; Nosil, 2008; Feder et al., 2012). Historically, gene flow between divergent populations was thought to erode any accumulated genetic differences, thereby blurring species boundaries and disrupting the process of speciation (Coyne and Orr, 2004). While this certainly can be the case in many circumstances, it is now understood that the outcome of on-going gene flow can be more complex (e.g. Gompert et al. (2014); Crawford et al. (2015); Dupuis and Sperling (2015); Wen et al. (2016)). For example on-going gene flow could result in a selection-migration equilibrium, or could lead to speciation with gene flow. Between two different species gene flow may introduce novel biological diversity via hybridization followed by introgression. Accompanying this renewed interest in the role of gene flow during the evolution of reproductive isolation and

biological diversity are predictions about the distribution of genetic variation across the genome. In cases of speciation with gene flow evolutionary theory predicts that a small number of regions, that are potentially under strong selection, will be highly differentiated while the rest of the genome will show weak differentiation because of the homogenizing impact of on-going gene flow (Feder et al., 2012). However testing this hypothesis can be challenging as similar genomic patterns may result from other evolutionary processes such as incomplete lineage sorting, or reduced diversity in regions of the genome that show elevated differentiation (e.g. Cruickshank and Hahn (2014)).

An excellent opportunity to explore the mechanisms underlying the evolution of biological diversity and reproductive isolation exists in species that show resource polymorphisms, where discrete intraspecific morphs show differential resource use. In vertebrates, resource polymorphisms are wide spread and occur in many different taxa, from birds through to fish (Skulason and Smith, 1995; Smith and Skúlason, 1996). These polymorphisms may represent the early stages of evolutionary divergence, the collapse of a hybrid lineage, or may be maintained in a stable equilibrium through mechanisms such as density-dependent selection (Wimberger, 1994; Smith and Skúlason, 1996; Kopp and Hermisson, 2006; Rueffler et al., 2006). Investigating genomic patterns of differentiation between morphotypes in wild populations can contribute substantially to our understanding of speciation. The opportunity to explore patterns of genetic and genomic variation, and the role of gene flow, exists in a trophically polymorphic species of fish, the Cuatro Ciénegas cichlid (*Herichthys minckleyi*). *H. minckleyi* is endemic to the Cuatro Ciénegas valley in northern Mexico and is trophically

polymorphic (Kornfield and Koehn, 1975; Kornfield et al., 1982; Liem, 1984; Sage and Selander, 1975). In *H. minckleyi* morphotypes can be distinguished based on the pharyngeal jaw morphology. Individuals either have a papilliform pharyngeal jaw, with small needle-like teeth, or a molariform pharyngeal jaw with large molar-like teeth. These two morphotypes also have associated feeding differences: molariform individuals have been found to have a much higher proportion of snails in their diet relative to papilliform individuals (Hulsey et al., 2006).

In other species of cichlid, divergent pharyngeal jaw morphotypes are phenotypically plastic and tooth size differences develop in response to diet (Muschick et al., 2011). Research examining dentition in African cichlids in Lake Malawi found evolution of novel diversity in dentition likely results from changes in a small number of conserved genetic regions (Albertson et al., 2003; Streelman and Albertson, 2006; Loh et al., 2008; Fraser et al., 2009). While there is evidence that the alternative jaw morphotypes in *H. minckleyi* has a plastic component, previous work has also demonstrated that pharyngeal jaw type likely has an underlying genetic basis (Stephens and Hendrickson, 2001; Trapani, 2003). Although there is potentially a genetic basis for pharyngeal jaw morphology and on-going gene flow, within the pools where *H. minckleyi* is found both morphotypes are always present and few intermediate individuals exist (Hulsey and García-de León, 2013). Recent molecular work, such as that by Hulsey and García-de León (2013) and Magalhaes et al. (2015) found no evidence of genetic differentiation between morphotypes and found geographic structure among different pools within the valley. Additionally, both studies found evidence of mitochondrial introgression between *H. minckleyi* and *H.*

cyanoguttatus, a closely related species whose range is known to overlap that of the Cuatro Ciénegas cichlid. This suggests that there could be either historic or substantial on-going hybridization between *H. minckleyi* and *H. cyanoguttatus* (Hulsey et al., 2016). Here we employ markers that provide a substantially broader view of overall genomic divergence compared to previous studies.

In order to understand the role that gene flow may play in this system we explore patterns of genetic and genomic variation at three hierarchical levels; 1) between the two species, *H. minckleyi* and *H. cyanoguttatus*, 2) between *H. minckleyi* individuals from two different geographic locations (pools) within the Cuatro Ciénegas valley, 3) between *H. minckleyi* individuals with alternate morphotypes within pools. For each level of this hierarchy we use both common (minor allele frequency $\geq 5\%$) and rare (minor allele frequency $< 5\%$) genetic variants. A comparison of common versus rare genetic variants may reveal differing evolutionary histories (Gompert et al., 2014). Common genetic variants may represent historic patterns of gene flow while rare genetic variants may represent newer mutations that are spatially restricted and therefore represent more recent gene flow (Gravel et al., 2011; Gompert et al., 2014). Thus, stratifying loci as common and rare variants could be a powerful approach for distinguishing between historical and more recent, or contemporary gene exchange. To analyze patterns of genetic variation at a genome-wide scale, for both classes of genetic variant, we use a clustering algorithm and principal component analysis (PCA), at a locus-specific scale we estimate F_{ST} and in order to further assess the amount of genotypic variance attributable to tooth size and location we use a redundancy analysis (RDA).

Methods

Sampling and Collecting

All fish were collected from the wild in 2008. *H. minckleyi* were collected from two pools, Escobedo and Juan Santos, in the Cuatro Ciénegas valley, Coahuila Mexcio (n=69) (Figure 4.1, Supplemental Table 2). At each location *H. minckleyi* individuals were morphotyped in the field using an otoscope placed into the throat of the fish, following methods of Kornfield and Taylor (1983); Hulsey et al. (2005). The presence of large molar-like teeth was used to diagnose fish as molariform. The absence of molar-like teeth was used to diagnose fish as papilliform. Individuals from both morphotypes were collected from Escobedo and Juan Santos. A small number of individuals from Escobedo (n=5) were identified as having intermediate tooth size. *H. cyanoguttatus* individuals were collected from Rio Salado, just outside of the Cuatro Ciénegas valley (n=10). Fish were fin clipped then stored in formalin for further studies. To determine tooth size on the lower pharyngeal jaws of the *H. minckleyi* individuals, we dissected the fifth ceratobranchial, or lower pharyngeal jaw, from the fish. These bony elements were cleaned of all muscle and fascia and allowed to dry. Then, we took a digital image of the dorsal surface of the jaws and imported it into ImageJ (Schneider et al., 2012). Using a size-standard placed in each image, we digitally drew a circle around the top-most right and left tooth on the lower pharyngeal jaw to determine tooth area. Tooth size was standardized by averaging the two teeth measurements, taking the square root, and then converting this to a proportion of standard length.

Molecular Methods

DNA was extracted from fin clips using QIAgen's DNeasy Blood and Tissue kit (QIAgen Inc.) following the manufacture's protocol. DNA sequence data were obtained following the methods of Gompert et al. (2012a) and Parchman et al. (2012). Briefly, each individual's genome was fragmented using two restriction enzymes (EcoR1 and Mse1). Next, customized Illumina adapter sequences and a unique eight to ten base pair barcode sequence were ligated to each DNA fragment. The unique barcode allows us to identify which fragments come from which individual and allows for the pooling of samples during sequencing. Following ligation of the adapter sequences and barcodes, fragments are amplified during two rounds of Polymerase Chain Reaction (PCR). Pooled PCR product was run on a two percent agrose gel and fragments between 200 and 500 bp were size selected by excising them from the gel using QIAgen Gel Purification kit (QIAgen Inc.), following the manufactures protocol. Samples were sequenced using Illumina GAII technology at the National Center for Genomic Research (Santa Fe, NM). Sequences have been previously used in the study by Hulsey et al. (2016).

Sequence reads were processed using a combination of custom Perl scripts, BCFtools and SAMtools (Li et al., 2009), following the methods of Gompert et al. (2012a), Parchman et al. (2012) and Gompert et al. (2014) (All custom scripts are available by contacting the authors) As we do not have a genome for this species we carried out a *de novo* assembly of the 15,800,517 sequence reads using SEQMAN NGEN ver. 11.0.0.172 (DNASTAR). We used a minimum match percentage of 62%, mismatch penalty of 15, and gap penalty of 30. This resulted

in 142,618 consensus sequences. These were treated as an artificial chromosome during our reference-based alignment which was conducted using the aln and samse algorithms in BWA ver 0.7.5 (Li and Durbin, 2009). We used a maximum difference of two base pairs between the reference and sequence being aligned, a maximum gap of one and only assembled reads that had a unique best match. At this point we created two data sets, one that contained individuals from both focal species and a second data set that only had individuals from *H. minckleyi*. This allowed us to look for fine-scale differences within *H. minckleyi* using SNPs that may not be present in the full data set due to differences in either alleles, or restriction sites, between the two species. For both data sets variable sites were called using SAMtools and BCFtools (Li et al., 2009). We required 75% of individuals to have at least one read at a site in order for it to be called as variable. We kept one variable site per contig and sorted variable sites into common (minor allele frequency $\geq 5\%$) and rare variants (minor allele frequency $< 5\%$) using allele frequency point estimates obtained from BCFtools. For the full data set this resulted in 6,220 common single nucleotide polymorphisms (SNPs) and 3,009 rare SNPs. For the *H. minckleyi* only data set this resulted in 6,587 common SNPs and 3,256 rare SNPs. Genotype likelihoods for SNPs were used in all downstream analyses.

Distribution of Genetic Variants

Genotype probabilities and admixture proportions were estimated using the program ENTROPY, developed by Gompert et al. (2014). ENTROPY implements a Bayesian hierarchical model similar to that used in STRUCTURE (Pritchard

et al., 2000; Falush et al., 2003). An important difference between ENTROPY and STRUCTURE is that ENTROPY incorporates sequence coverage, sequence error and alignment error into the model (Gompert et al., 2014). As with STRUCTURE, ENTROPY requires specification of the number of ancestral clusters (k) and does not incorporate any prior information about which cluster an individual is assigned to. For the full data set our primary focus was to understand patterns of gene flow between the two species *H. cyanoguttatus* and *H. minckleyi* therefore ENTROPY was run for k=2 for both common and rare genetic variants. For the *H. minckleyi* data set ENTROPY was run for k=2 through k=9 for common and rare genetic variants. For each value of k, parameter estimates were obtained using Markov Chain Monte Carlo (MCMC). For each model we ran four chains, 270,000 steps, a burn-in of 50,000 and retained every 20th value. This resulted in 11,000 samples from the posterior distribution for each chain (total of 44,000 samples). MCMC chains were checked for convergence and stabilization by estimating effective sample size and Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992). In order to visualize the relationships between sample groups a principal components analysis (PCA) was conducted using estimates of genotype probabilities from ENTROPY for both the full and *H. minckleyi* data sets, for both the common and rare SNPs. PCA's were conducted using the statistical program R (using the prcomp function) (R Core Team, 2016). To explore variation at a locus-specific scale genotype probability estimates from ENTROPY for common SNPs were used to calculate locus-specific Wright's F_{ST} for both the full data set and for the *H. minckleyi* data set. For the *H. minckleyi* data set, in order to identify those SNPs that

have F_{ST}'s significantly different from zero we created a null distribution for each SNP from 1,000 permutations of individuals across morphotypes and across pools. A redundancy analysis (RDA) was used to estimate the relationship between common genetic variants, locality, and tooth size in *H. minckleyi* (Van Den Wollenberg, 1977). Redundancy analysis is a canonical ordination method used to estimate the influence of geographic location and tooth size as predictors of genotype probabilities for the common SNP data set for *H. minckleyi* using the vegan package in R (R Core Team, 2016; Oksanen et al., 2013) (Peres-Neto et al., 2006). We used a permutational ANOVA with 999 replicates in the vegan package to test for significance of each term in the model. For the scaffolds containing the 10 highest, and 10 lowest SNP RDA scores for each axis we used the BLAST command line tool to assess if they were associated with any specific genes or genomic features (Altschul et al., 1990). We also compare F_{ST}'s for all pairwise comparisons for those SNP's that were found to have high RDA scores. The reduced representation Illumina reads are available on the NCBI SRA database (SAMN04523166).

Results

We sequenced over 15 million 125bp DNA fragments from 10 *H. cyanoguttatus* individuals (one location) and 69 *H. minckleyi* individuals (two locations), including both papilliform and molariform morphotypes (Figure 4.1). After assembly and variant calling this resulted in 6,220 common (minor allele frequency $\geq 5\%$) single nucleotide polymorphisms (SNPs) and 3,009 rare SNPs (minor allele frequency $< 5\%$) for the data set with all individuals and 6,587

common SNPs and 3,256 rare SNPs for the *H. minckleyi* data set.

To explore genome-wide genetic differentiation between *H. cyanoguttatus* and *H. minckleyi* we used a PCA, and estimated admixture proportions for k=2 (Figure 2.2A and 2.3). The PCA conducted using common genetic variants showed a clear distinction between *H. cyanoguttatus* and *H. minckleyi* individuals along the PC1 axis, which explained 59.94% of the variation in the data set. The PC2 axis explained 1.66% of the variation and divides individuals of *H. minckleyi* based on sampling location (Figure 2.2A). The PCA conducted using the rare genetic variants showed similar patterns (Supplemental Figure 1A). Overall the PCA analyses showed that the distribution of both the common and rare genetic variants reflects the species boundaries (Figure 2.2A). The barplots of admixture proportions for common and rare genetic variants for the k=2 model in ENTROPY reflect similar patterns, at a genome-wide scale we found little evidence of introgression or gene flow between the two species (Figure 2.3). For the common data set a small number of *H. minckleyi* individuals showed low levels of shared ancestry with *H. cyanoguttatus* ($< 4\%$). This was not present in admixture proportions estimated from the rare data set. Estimates of locus-specific F_{ST} of common SNP's for pairwise comparisons between *H. cyanoguttatus* and *H. minckleyi* showed a U-shaped distribution of F_{ST} (Supplemental Figure 3), where there were a high number of loci with either F_{ST} of zero, or F_{ST} of one. There was a smaller number of loci with intermediate values of F_{ST}.

For our exploration of the *H. minckleyi* data set the first PC axis (6.82%) of the PCA of common genetic variants divides individuals based on sampling locality (Figure 2.2B). Individuals of both morphotypes from Juan Santos cluster at one

end of the axis, while all individuals from Escobedo cluster at the opposite end. The PC2 axis does not show any distinctions between morphotypes within locations. Additionally bar plots of admixture proportions using common genetic variants do not show evidence of genetic differentiation between different morphotypes within the same geographic location (Figure 2.4). Together, for common genetic variants, these results demonstrate geographic genetic structure. However, we found little evidence of genetic differences between individuals with different morphotypes. For the *H. minckleyi* data set rare genetic variants showed no clear patterns of differentiation, this could reflect recent gene flow or could be because these markers have a lack of resolution (Supplemental Figure 1B and Figure 2).

At the locus-specific level comparisons between *H. minckleyi* groups had values of F_{ST} that were lower than comparisons between the two species and there were no fixed differences between alternate alleles (Supplemental Figure 4). In contrast to the between species comparisons, the distribution of F_{ST} between *H. minckleyi* samples was L-shaped rather than U-shaped, meaning that the majority of locus specific F_{ST}'s were close to zero and a small number of loci had higher F_{ST}'s but there were no fixed differences. Our significances test of F_{ST} showed that a higher number of SNPs had F_{ST}'s significantly different from zero for comparisons between pools (for α <0.05 number of SNPs ranged from 1,194 to 1,329 and between 425 and 604 for α <0.01), than comparisons between morphotypes within pools (for α <0.05 MJS vs PJS has 291 SNPs and for MES vs PES there were 359 SNPs, for α <0.01 there were 35 SNPs for MJS vs PJS and 42 SNPs for MES vs PES). To further explore differences between pools, and

between morphotypes within pools we conducted an RDA with standardized tooth size and geographic sampling location as a predictor for common genetic variants, respectively. We found the individuals divided along the RDA1 axis (6.35%) based on sampling geography (Figure 2.5). Along the RDA2 axis (1.43%) individuals were separated based on tooth size, therefore showing a relationship between genetic variants and standardized tooth size. We conducted a permutational ANOVA on the RDA using the R package Vegan. We found that geographic location was significant ($P<0.0009$) but tooth size was not ($P = 0.7992$). Despite a lack of significance we still believe that a small number of SNP's that have high RDA scores may have biological significance (Legendre et al., 2011). We examined locus-specific F_{ST}'s for those SNP's identified from the RDA and found that several were significantly different from zero (Supplemental Table 1). This includes SNPs identified from the RDA2 axis which were significantly different from zero in pairwise comparisons between different morphotypes within a pool. Thus, both the RDA and individual locus F_{ST} values suggest a non-random association of variation at a small number of genomic regions and pharyngeal tooth size.

Discussion

In this study we explored the evolutionary history of gene flow in a trophically polymorphic species of cichlid. We explored genetic variation at three hierarchical levels; between *H. cyanoguttatus* and *H. minckleyi*, between *H. minckleyi* individuals from two geographic locations, and finally between *H. minckleyi* individuals with alternate morphotypes. We employed a multifacited approach to

explore the evolutionary patterns at both a genome- and locus-specific scale. Previous research into this system using mitochondrial DNA suggested gene flow between *H. minckleyi* and a closely related species *H. cyanoguttatus* (Hulsey and García-de León, 2013; Magalhaes et al., 2015). However, previous analyses using nuclear markers found limited support for on-going introgression between the two species, despite the patterns found in the mitochondrial genome (Hulsey and García-de León, 2013; Magalhaes et al., 2015). In the last 100 years canals have been built in the Cuatro Ciénegas valley which have brought *H. minckleyi* and *H. cyanoguttatus* into close contact (Chaves-Campos et al., 2011a,b). It is has been suggested that hybridization between *H. minckleyi* and *H. cyanoguttatus* could have led to the jaw polymorphism observed among *H. minckleyi* individuals (Hulsey and García-de León, 2013). For common genetic variants we found a small number of *H. minckleyi* individuals had a low level of shared ancestry with *H. cyanoguttatus* (Figure 2.3A). However for rare genetic variants we found no evidence of on-going gene flow between any of the *H. minckleyi* populations and *H. cyanoguttatus* (Figure 2.3 and Supplemental Figure 1A). At a locus-specific scale for common genetic variants, estimates of F_{ST} showed both shared alleles and fixed differences between the two species, and overall F_{ST}'s between the two species were much higher than those calculated for comparisons within *H. mickleyi* (Supplemental Figures 3 and 4). We found limited evidence of gene flow or shared ancestry between the two species for common genetic variants, but not for rare variants. This pattern could result from incomplete lineage sorting. Alternatively, this could suggest historical gene flow but it is possible that recent gene flow has occurred between the two species in genomic regions not tagged by

our SNP markers. Therefore we cannot rule out that limited introgression of the nuclear genome between *H. cyanoguttatus* and *H. minckleyi* has taken place. While initially considered as different species, early genetic work comparing *H. minckleyi* individuals with alternate morphotypes using allozymes, and more recent work using a larger number of nuclear markers, found scant evidence of genetic differentiation between the morphotypes (Hulsey and García-de León, 2013; Kornfield and Koehn, 1975; Kornfield et al., 1982; Magalhaes et al., 2015; Sage and Selander, 1975). We build upon this previous genetic work and take advantage of next generation sequencing technology to generate a much higher number of genetic markers than were previously available and additionally we explore patterns at both a genome- and locus-specific scale. For common genetic variants we found that all our analyses identified geographic structure as the primary isolating factor. A PCA of genotype probabilities of common genetic variants demonstrated that individuals cluster together based on geographic sampling location, either Escobedo or Juan Santos (Figure 2.2B). The PCA does not show any differences between individuals of alternate morphotypes from within the same pool. Consistent with the PCA, admixture proportions for common genetic variants demonstrated geographic differences but did not detect any differentiation between different jaw types within the same pool. Previous research on other species within the valley, and on *H. minckleyi*, have demonstrated that geographic structure between pools is not unusual, particularly between Escobedo and Juan Santos due to the topology of the region (Coghill et al., 2013). For rare genetic variants we found little evidence of genetic structure between geographic locations at a genome-wide scale

(Supplemental Figure 1B and 2). This could indicate recent gene flow, perhaps from a recent flooding event, but this pattern could also be the result of poor resolution from the rare genetic variants as they may only be present in one or two individuals in the data set.

At a locus-specific scale, pairwise calculations of F_{ST} for common genetic variants within *H. minckleyi* found no fixed differences (Supplemental Figure 4) and overall values were much lower on average than comparisons between *H. cyanoguttatus* and *H. minckleyi*. The majority of loci showed low values of F_{ST}, while a small proportion of loci showed higher values of F_{ST}. This L-shaped distribution of F_{ST} is consistent with what would be predicted under a speciation with gene flow model, where a small proportion of the genome, under strong selection, shows pattens of differentiation while the rest of the genome remains undifferentiated due to on-going gene flow. We identified several SNP's that has F_{ST}'s significantly different from zero for all pairwise comparisons. As expected we found more between individuals from different pools, but still identified significant structure between morphotypes within a pool. While this is consistent with speciation with gene flow, much more exploration would be needed in order to conclude that this is occurring in *H. minckleyi*. In *H. minckleyi* it is unclear how the bi-modal distribution of jaw type is maintained while there is limited genetic structure between individuals with different jaw types. It could be that, as described above, only a small proportion of the genome is under selection. This could explain the bi-modal distribution of jaw type in *H. minckleyi* that is maintained with no detectable genetic differentiation between individuals of alternate morphotypes at a genome-wide scale.

Therefore we aimed to explore the relationship between genetic variation and tooth size (large teeth represent molariform individuals, small teeth represent papilliform individuals) using an RDA, a canonical ordination method. In addition to exploring the relationship between genetic variants and tooth size we also included geography. As expected, for common genetic variants RDA1 separated individuals based on geographic location, with individuals from Escobedo clustering on one side and Juan Santos individuals on the other (Figure 2.5). The second RDA axis separated individuals based on standardized tooth size, suggesting that there is a small amount of genetic variation that varies with tooth size.Given our expectation about genomic architecture of jaw morphology, that there will be a small number of regions underlying the trait, it is not surprising that the term for tooth size in the model is not statistically significant. Despite this we follow the advice of Legendre et al. (2011) and suggest that those SNPs which show high scores in the RDA, while that term is not statistically significant, may have biological significance and those regions of the genome where they are located may warrant further investigations. We compared pairwise F_{ST}'s for those SNPs identified from the RDA model and found that several also had F_{ST}'s significantly different from zero, providing further evidence that these regions may play a role in differentiation between pools and or between morphotypes within pools. This is the first time a relationship between genetic variation and standardized tooth size has been documented in *H. minckleyi*. Further exploration of those genetic variants that have the highest RDA loadings provides an important direction for future work into the genetic basis of tooth size in *H. minckleyi* (Supplemental Table 1).

The possibility of a simple architecture for the jaw polymorphism in *H. minckleyi* has been suggested before (Sage and Selander, 1975; Wimberger, 1994). This comports with expectations that simple architectures underlie resource polymorphisms (Smith and Skúlason, 1996) and general models of divergence with gene flow (Gavrilets et al., 2007). Strong selection would presumably keep a small number of quantitative trait loci (QTL) in mutation - selection balance. Alternatively, the polymorphism could be protected by a small inversion (e.g. Barth et al. (2017)). While there is some evidence for simple architectures underlying even complex traits (e.g. Boyko et al. (2010)), several cases of local adaptation have demonstrated complex architectures even despite theoretical expectations (e.g. Oppenheim et al. (2018); Marques et al. (2016)). A full exploration of the genomic architecture is not possible with the data presented here. Even though 6,000 SNP markers is an improvement over previous studies of the polymorphism in *H. minckleyi*, these SNPs represent very sparse coverage of the genome, making detection of QTL of small effect quite difficult. Greater coverage of the genome will be required to answer questions about the architecture of this resource polymorphism.

Conclusions

This research used thousands of nuclear genetic markers sampled from throughout the genome to explore the origin and maintenance of alternate jaw morphotypes within the Cuatro Ciénegas cichlid, *Herichthys minckleyi*. At a genome-wide scale we found limited evidence of introgression between *H. minckleyi* and *H. cyanoguttatus* in common variants and no evidence of gene

flow for rare genetic variants. At a locus-specific scale we found relatively high F_{ST}'s but also found shared alleles between the two species. Therefore we cannot rule out introgression between the two species, despite the fact we found little evidence for this. For our genome-wide analyses of *H. minckleyi* genotypes we identified genetic differentiation between geographic locations for common genetic variants, but not for rare. For the first time we were able to identify low levels of genetic differentiation for both common and rare genetic variants between *H. minckleyi* individuals with alternate morphotypes within a pool. However, we were only able to detect genetic differentiation in our calculation of F_{ST} and with an RDA of standardized tooth size, but found no evidence of differences at a genome-wide scale in our admixture model or a PCA of multi-locus genotype. This suggests that only a small region of the genome has accumulated differences between individuals with alternate jaw types. This may provide the explanation for why no genetic differentiation between morphotypes has previously been documented. Together these patterns suggest a complex evolutionary history of intermittent and brief connections between pools, and potentially selection acting upon only a small region or regions of the genome to maintain alternate morphotypes. Future work could use loci identified from our RDA as candidates in identifying the genetic architecture of local adaptation that might lead to reproductive isolation between pools and between morphotypes.

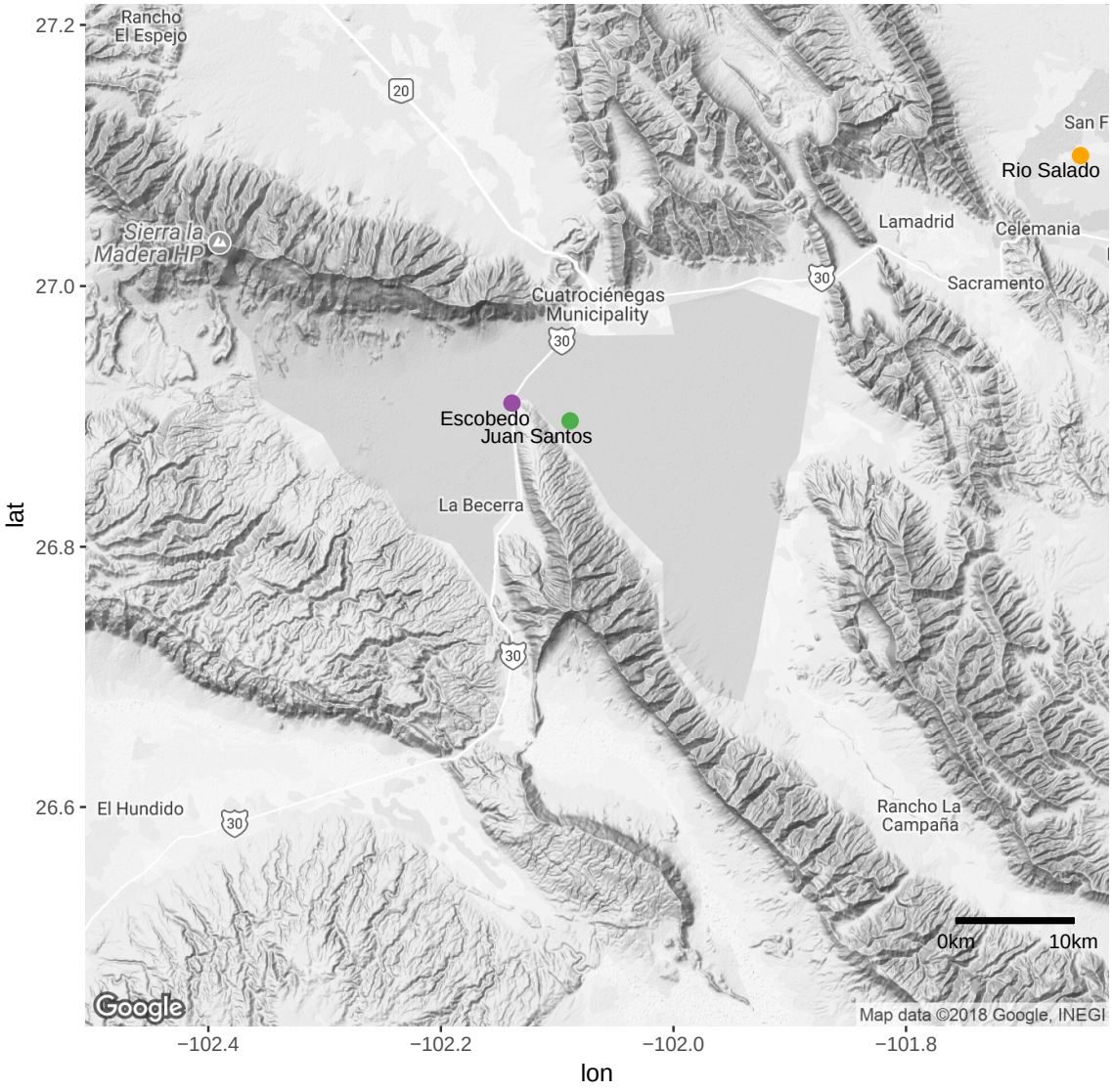

Figure 2.1: Map of sampling locations in the Cuatro Ciénegas valley, Coahuila Mexico. ©Google 2018

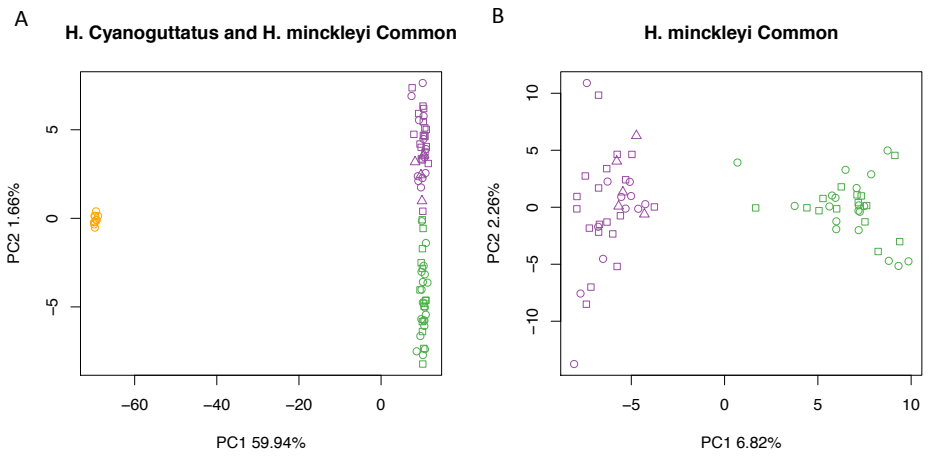

Figure 2.2: Principal component analysis (PCA) of genotype probabilities. Each dot represents an individual, whose position in ordination space is determined by it's multi-locus genotype. Orange = *H. cyanoguttatus*, squares = papilliform, circles = molariform, triangles = intermediate, purple = *H. minckleyi* from Escobedo, green = *H. minckleyi* from Juan Santos A: *H. cyanoguttatus* and *H. minckleyi* common (6,220) SNP's, B: *H. minckleyi* common (6,587) SNP's.

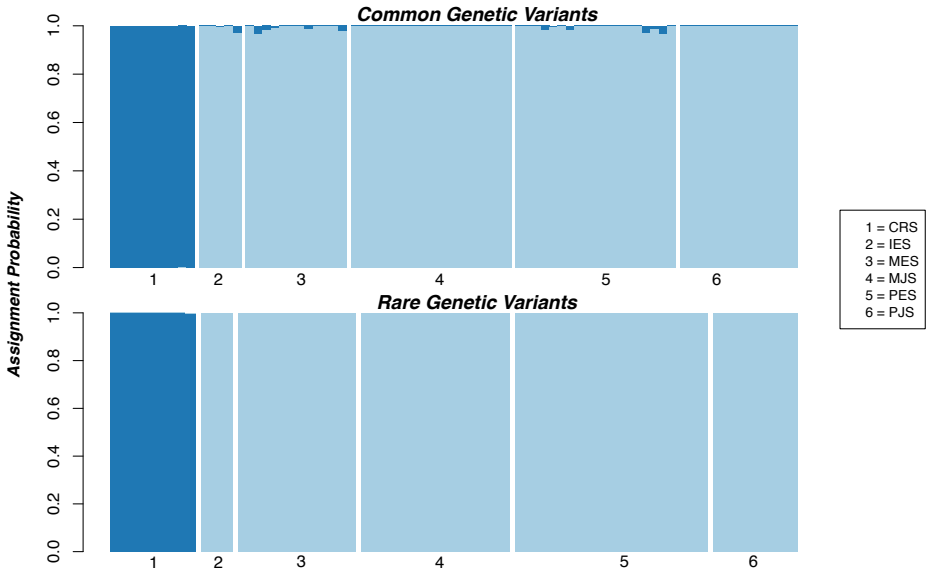

Figure 2.3: Barplot of admixture proportions for k=2 model in ENTROPY for *H. cyanoguttatus* and *H. minckleyi*. The top barplot shows admixture proportions for common genetic variants, the bottom graph shows admixture proportions for rare genetic variants. Each bar represents one individual's assignment proportions to each of the 2 source populations. Number of individuals differs between common and rare data sets due to differences in coverage (see text). Numbers underneath bars represent sampling location and morphotype: 1 = *H. cyanoguttatus*, 2 = *H. minckleyi* intermediate from Escobedo, 3 = *H. minckleyi* molariform from Escobedo, 4 = *H. minckleyi* molariform from Juan Santos, 5 = *H. minckleyi* papilliform from Escobedo, 6 = *H. minckleyi* papilliform from Juan Santos

Figure 2.4: Barplot of admixture proportions for k=2 through k=5 in ENTROPY based on common genetic variants for the minckleyi data set. Each bar represents one individual's assignment proportions to each of k source populations. Numbers underneath bars represent sampling location and morphotype: 1 = Intermediate from Escobedo, 2 = Molariform from Escobedo, 3 = Molariform from Juan Santos, 4 = Papilliform from Escobedo, 5 = Papilliform from Juan Santos

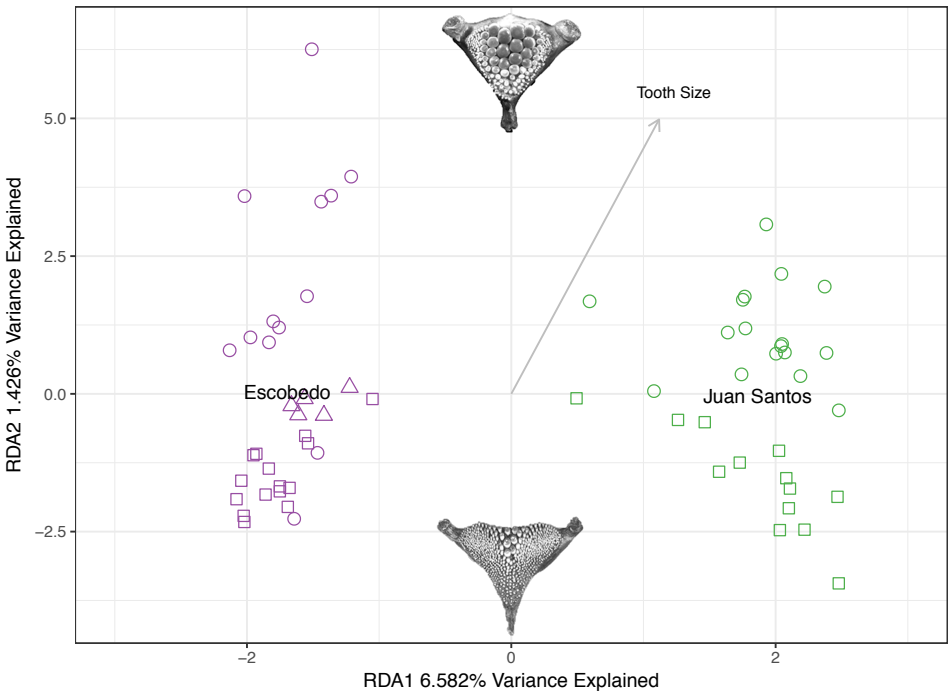

Figure 2.5: RDA plot demonstrating the relationship between *H. minckleyi* common genetic variants, geographic location, and standardized tooth size. Squares = papilliform, circles = molariform, triangles = intermediate, purple/dark grey = *H. minckleyi* from Escobedo, green/light grey = *H. minckleyi* from Juan Santos. The arrow represents the vector for standardized tooth size.

3. THE GENOMIC ARCHITECTURE OF JAW POLYMORPHISM IN THE CUATRO CIÉNEGAS CICHLID (*HERICHTHYS MINCKLEYI*)

Introduction

Understanding the evolution of novel phenotypes remains a central focus in evolutionary ecology. Species that have resource polymorphisms provide the opportunity to determine not only the origin and maintenance of novel phenotypes, but also their adaptive significance. Trophic or resource polymorphisms occur when discrete intraspecific morphs show differential resource use. In vertebrates, resource polymorphisms are wide spread and occur in many different taxa, from birds through to fish (Skulason and Smith, 1995; Smith and Skúlason, 1996). The exploration of intraspecific variation, such as resource polymorphisms, can provide important insight into the process of speciation. It has been proposed that speciation can be thought of as a continuum, which ranges from intraspecific variation in a panmictic population through to the evolution of complete reproductive isolation between divergent populations (Hendry et al., 2009; Nosil, 2012). Within this paradigm there are important questions about the stability of evolutionary outcomes, and the factors that may influence these outcomes.

One aspect which will influence evolutionary outcomes is the genetic architecture of traits which are under selection. Genetic architecture refers to the number of regions of the genome associated with a trait, and the effect size of those regions.

If only a limited number of regions affect a trait which is under selection then local adaptation will necessarily be constrained, but genetic architecture may also by shaped by evolutionary processes such as natural selection and gene-flow (Yeaman and Whitlock, 2011). If genetic architecture is complex, involving numerous regions of the genome each with small effect size, then continuing recombination as a consequence of gene flow between alternate morphotypes would result in a"breaking up" of the genetic combinations that lead to optimal phenotypes. In this case one might expect the population to progress along the speciation continuum, as the evolution of reproductive isolation between morphotypes would be required in order to maintain an optimal phenotype. An alternative possibility is that natural selection may act upon this complex genetic architecture via chromosomal evolution, leading to gradual replacement of multiple regions of small effect, with fewer regions that have larger effect (Yeaman and Whitlock, 2011). Therefore although an adaptive trait may initially have a complex architecture, continued exposure to selection and gene-flow may lead to the evolution of a simpler architecture (Yeaman and Whitlock, 2011). If the genetic architecture of an adaptive trait is simple (i.e. a small number of regions of the genome which have large effect sizes) the evolution of assortative mating may not be required, and alternate morphotypes could be maintained through frequency-dependent disruptive selection (Via, 2012). This would represent a case where the population does not advance along the speciation continuum to the evolution of two different species. Understanding the interaction between gene flow and genetic architecture of an adaptive trait provides information about where populations are on the speciation continuum,

and if those populations may evolve into distinct, reproductively isolated, species. The opportunity to explore the relationship between gene flow and the genetic architecture of adaptive traits exists in the Cuatro Ciénegas cichlid (*Herichthys minckleyi*). Individuals are trophically polymorphic and have either a papilliform pharyngeal jaw with small, needle-like teeth and feed on detritus, or a molariform jaw with large molar-like teeth and feed on snails (Kornfield and Koehn, 1975; Kornfield et al., 1982; Liem, 1984; Sage and Selander, 1975). Previous studies have demonstrated that while tooth size shows some plasticity, there is also an underlying genetic basis (Stephens and Hendrickson, 2001; Trapani, 2003). Numerous questions remain about the architecture of pharyngeal jaw type, the distribution of the discrete jaw types within populations (Figure 3.1), and the lack of genome-wide genetic differentiation between morphs (Bell *et.al. In Review*) leads to an important question about how this polymorphism is being maintained in *H. minckleyi*. As predicted from quantitative genetics theory, if genetic architecture were complex then ongoing recombination would result in continuous variation of jaw type, not the bimodal distribution observed in *H. minckleyi* populations. If jaw morphology were to have a simple architecture, particularly if this architecture involved some level of dominance, then ongoing gene flow between the two morphs would not disrupt the distribution of phenotypes in the same way. There are several possibilities about how these two discrete morphotypes are maintained; 1) If genetic architecture for morphotypes is relatively simple, and there is no detected reduction in gene flow, then it is likely morphotypes are maintained by frequency-dependent disruptive selection. 2) If genetic architecture is simple, but there is detectable genetic

differentiation between morphotypes, this could be indicative of speciation-with-gene-flow. How far the process of assortative mating continues along the speciation continuum may depend on several factors such as the strength of selection. 3) If genetic architecture is complex, we would expect to detect some level of genetic differentiation between morphotypes which would be indicative of ongoing evolution of reproductive isolation. In *H. minckleyi* there is the opportunity to investigate these three possibilities by exploring multiple, isolated populations. It is likely that different populations have different demographic histories and experience different levels of selection, both of which could affect where these populations lie on the speciation continuum.

We sampled *H. minckleyi* individuals of both morphotypes from five pools within the Cuatro Ciénegas valley and used a genotype-by-sequencing approach to answer the following two questions; 1) What are the patterns of genetic differentiation between pools, and between individuals within pools? 2) What is the genetic architecture of pharyngeal jaw tooth size? We also explore the genetic architecture of other relevant traits, such as jaw protrusion, in order to compare their architecture to that of tooth size.

Methods

Sampling and Collection

We collected 169 individuals from the wild from five pools within the Cuatro Ciénegas valley, Coahuila Mexico in 2008 (Table 4.1). From each pool individuals of both morphotypes (papilliform and molariform) were sampled. Morphotypes were identified in the field using an otoscope placed into the throat of the fish,

following methods of Kornfield and Taylor (1983) and Hulsey et al. (2005). An individual was identified as molariform if there was at least one large tooth present. Fin clips were taken, then fish were placed in formalin for further morphological measurements. To determine tooth size on the lower pharyngeal jaws, we dissected the fifth ceratobranchial, or lower pharyngeal jaw, from the fish, bony elements were cleaned of all muscle and fascia and allowed to dry. A digital image of the dorsal surface of the jaws was taken and imported into ImageJ (Schneider et al., 2012). Using a size-standard placed in each image, we digitally drew a circle around the top-most right and left tooth on the lower pharyngeal jaw to determine tooth area. Tooth size was standardized by averaging the two teeth measurements, taking the square root, and then converting this to a proportion of standard length.

Molecular Methods

We used a genotype-by-sequencing approach to produce a reduced representation genomic library for each individual, following the methods of Parchman et al. (2012) and Gompert et al. (2012a). Briefly, we used two restriction enzymes (EcoR1 and MSE1) to digest the genome, ligated a unique 8-10 base pair barcode to each fragment, and then amplified fragments using two rounds of PCR. Genomic libraries were sent to the University of Texas, Austin Genome Sequencing and Analysis Facility (GSAF) where fragments were then size selected (200 to 350 bp) using Blue Pipin (Sage Science). Once size selected, fragments were sequenced using one lane of Illumina HiSeq 2500 technology. This resulted in over 275 million reads. A custom perl script was used to remove

unique barcodes, then we conducted a *de novo* assembly following the dDoccent protocol, with a few minor modifications (Puritz et al., 2014a,b). This resulted in 132,222 scaffolds which we used as an artificial reference for alignment of all sequences using BWA ALN and SAMSE algorithms. We called variants using SAMTOOLS and BCFTOOLS and required 75% of individuals to have a read in order for a sequence to be called as variable. We removed variable sites only present in one individual. This resulted in 54,458 single nucleotide polymorphisms (SNPs).

Analysis of Genetic Differentiation

Genotype likelihoods from BCFTOOLS were used in the program ENTROPY to estimate genotype probabilities and admixture proportions (Gompert et al., 2014). ENTROPY implements a Bayesian hierarchical model, similar to STRUCTURE, but incorporates sequence coverage, sequence error, and alignment error into the model (Pritchard et al., 2000; Falush et al., 2003; Gompert et al., 2014). ENTROPY requires the specification of the number of ancestral clusters K, but does not incorporate any prior information about which cluster an individual is assigned to. We ran the model for $k = 2$ through to $k= 9$. For each value of k we used Markov Chain Monte Carlo (MCMC) to obtain parameter estimates. Each model was run for 2 chains, 170,000 steps with a burn-in of 25,000, and saving every tenth step. This resulted in 30,000 samples from the posterior distribution for each parameter. To check that the model had reached a stable sampling distribution we estimated effective sample size and calculated Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992). To visualize the

relationships between individuals we conducted a principal component analysis (PCA) on genotype probabilities using the prcomp function in R (R Core Team, 2016). Admixture proportions estimated from ENTROPY were plotted for $k = 2$ through $k = 9$.

Genetic Architecture of Pharyngeal Jaw Tooth Size

To map the genetic architecture of tooth size on the pharyngeal jaw we fit Bayesian Sparse Linear Mixed Models (BSLMMs) using the program GEMMA (Zhou et al., 2013). The BSLMM is a hybrid between sparse linear models and linear mixed models. These two types of models make very different assumptions about the genetic architecture of a trait and the validity of these assumptions is often not known *a priori*. BSLMM is able to fit a model for genetic architecture that incorporates both a small number of regions that have a large effect (sparse linear model), and a high number of regions that have a small effect (linear mixed model). The sparse effect of the model is estimated by the parameter β which represents the effect of an individual SNP on the trait, the SNP can be pulled into or out of the model (it is removed from the model by having a β value of zero) and the posterior inclusion probability (PIP) represents how frequently that SNP is included. The parameter μ represents the polygenic effect in the model which can be interpreted as the effect of a large number of markers all of small effect. MCMC is used to estimate several parameters from the model: posterior inclusion probability (PIP), phenotypic variance explained (PVE) this includes the polygenic term of the model and variance explained by SNPs with measurable effects or associations (β). PGE is the proportion of variance

explained by SNPs with measurable effects or associations only. The model also allows for the calculation of a kinship matrix, which accounts for phenotypic covariance among individuals based on their relatedness or genetic similarity, thereby removing the influence of population structure when calculating the association between a SNP and the trait of interest (Zhou et al., 2013). Each trait was first z-transformed, then the model was run for 1 chain for 1,000,000 steps with a burn-in of 100,000. *Top SNPs are currently being blasted, these results will be included once finished.*

Results

Our sequencing approach resulted in 54,458 SNPs. Genotype likelihood estimates from SAMTOOLS and BCFTOOLS were used in the program ENTROPY to obtain genotype probability estimates and admixture proportions for $k = 2$ through $k = 9$.

As is expected for the ENTROPY model, genotype probabilities were strongly correlated across values of k, therefore we arbitrarily chose to use genotype probability estimates from the $k = 2$ model (see Supplemental Figure 1). The principal component analysis (PCA) on genotype probabilities demonstrates genetic differentiation between several pools within the Cuatro Ciénegas valley (Figure 4.2). The PC1 axis explains 9.99% of the variation in genotype and shows the highest differentiation between individuals from Juan Santos and those from Tio Candido. Along PC1 Tierra Blanca and Mojarral Este are genetically similar to Juan Santos individuals and Escobedo individuals are intermediate between these populations and Tio Candido. PC2 explains 5.029%

of variation in the data and separates Tierra Blanca and Mojarral Este from Juan Santos. Individuals from Tierra Blanca and Mojarral Este cluster relatively closely together along both PC axes shown. The PCA showed no differences between individuals with alternate morphotypes within a pool. We plotted admixture proportions estimated form ENTROPY for $k = 2$ through $k = 9$ (Figure 3.3 and Supplemental Figure 2). For the $k = 2$ model individuals from Tio Candido form their own cluster, separate from all other individuals in the sample. For $k = 3$ the next population to form its own cluster is Juan Santos, for $k = 4$ Escobedo individuals form a cluster, but continue to show some shared ancestry with Tierra Blanca and Mojarral Este. For $k = 5$ through $k = 9$ no other samples form clearly delineated clusters. Overall this pattern reflects what we found in the PCA, whereby individuals from Tierra Blanca and Mojarral Este are genetically similar, and no genetic differences were identified at a genome-wide scale between individuals with alternate morphotypes.

We used a Bayesian Sparse Linear Mixed Model (BSLMM) implemented in GEMMA to map the genetic architecture of several traits associated with feeding differences in *H. minckleyi*. Our primary focus was mapping the architecture of tooth size on the pharyngeal jaw, and we use other traits as a basis for comparison. We found substantial genetic variation (PVE) for all traits, which ranged from PVE = 0.33 for gut length to PVE = 0.73 for jaw protrusion (Table 3.2). The PVE for tooth size falls in the middle of this range at 0.5 (Figure 3.4). PVE represents the total variance explained by both the polygenic term of the model, and the sparse effect of the model, (β), which represents variation due to SNPs with measurable effects or associations (Zhou et al., 2013). For PGE,

which is the proportion of variance explained by the sparse effect of the model only, we found a high level of variation in estimates for several traits (AscPro, gape, gut, SL) and therefore can not be very confident in the validity of these estimates (Figure 3.4 and Table 3.2). For protrusion there appears to be lower variation explained by PGE (0.31 compared to 0.73 for PVE) which indicates that the polygenic term of the model is very influential in explaining variation in jaw protrusion. Tooth size had the highest PGE of any of the traits (0.77) which indicates that the sparse effect of the model explains a large proportion of variation in tooth size. Estimates of the number of SNPs with sparse effect were relatively low for AscPro, gape, gut length, and standard length with the mean estimate ranging from 17 to 59. The number of SNPs with sparse effect for jaw protrusion was more variable, and had a mean of 133 SNPs (Table 3.2). Tooth size had the lowest number of SNPs with sparse effect, with a mean of 7. The distribution of posterior inclusion probabilities (PIPs) varied between traits (Figure 3.5). Tooth size had two SNPs that had the highest PIPs, with estimates of 0.98 and 0.97. This is much higher than the maximum PIP for any of the other traits, maximum PIP are as follows: AscPro = 0.52, gape = 0.13, gut = 0.25, prot=0.06, SL = 0.32. In order to estimate the over all effect size we multiple the sparse effect (β) by the PIP for that SNP. We found that tooth size had the two SNPs with the largest effect size. Three other traits (gut length, jaw protrusion, and ArcPro) had SNPs with relatively large effect, but these SNPs had over all much lower PIPs than was found for tooth size (Figure 3.6).

Discussion

We used a genome-wide sequencing approach to explore the evolutionary relationships between individuals with alternate morphotypes in a trophically polymorphic species of cichlid (*H. minckleyi*), and to map the genetic architecture of pharyngeal jaw tooth size. We estimated patterns of genetic differentiation between pools, and between individuals within pools and estimated the genetic architecture of tooth size. This research provides insight into the evolutionary processes that lead to the maintenance of jaw polymorphism in this species and, more broadly, into the process of speciation. In order to quantify genetic differentiation we used a PCA to visualize relationships between genotype probabilities (Figure 4.2) and estimated admixture proportions for $k = 2$ through $k = 9$ (Figure 3.3 and Supplemental Figure 2). Both the PCA on genotype probabilities, and barplots of admixture proportions demonstrated genetic differentiation between several of the pools within the Cuatro Ciénegas valley. These genetic relationships may reflect the different drainages within the valley (Johnson et al., 2007). In both the PCA and the estimates of admixture proportions individuals from Tierra Blanca and Mojarral Este cluster together, and are found in the same drainage within the valley. Escobedo, Juan Santos, and Tio Candido are in separate drainages and show higher levels of genetic differentiation. In the PCA and for the $k = 4$ model from ENTROPY they form their own clusters. While we identified geographic genetic structure, we did not identify any genetic structure between individuals with alternate morphotypes within a pool. This result is inkeeping with previous research that was only able to detect genetic differences using a locus-specific

approach, at a small number of markers in the genome (Bell *et. al. In Review.* This research represents a more comprehensive approach than previous work, both in terms of number of markers, and number of pools surveyed. Despite the bi-modal distribution of morphotypes within pools, there is still no evidence of significant reproductive isolation between individuals with different jaw types. As discussed in the introduction, this leads to important questions about the genetic architecture of tooth size in *H. minckleyi* and how this contributes to the maintenance of this polymorphism.

We mapped the genetic architecture of tooth size using a Bayesian Sparse Linear Mixed Model (BSLMM) in the program GEMMA. In addition to mapping tooth size we also included several other traits which may play a role in the maintenance of alternate diets in this species. This allowed us to compare and contrast genetic architecture within the same system. We found significant differences in the architecture of several traits compared to tooth size. Mapping for tooth size identified two SNPs within our data set that have extremely high posterior inclusion probabilities (0.98, and 0.97) (Figure 3.5). This indicates that these markers are nearly always included in the model to explain variation in tooth size. These PIPs are much higher than what was found for the other traits that we mapped, this indicates that these regions were consistently identified as playing an important role in the variation in tooth size. Additionally, we found that the effect size of these SNPs was much larger than what was found for other traits (Figure 3.6). In addition to identifying SNPs with high PIPs, we also found that the BSLMM for tooth size included overall less SNPs. We found the mean number of SNPs included in the model was 5, while for other traits this

ranged from 17 through to 133 (Table 3.2). These results provide support for the hypothesis that tooth size in *H. minckleyi* has a simple genetic architecture, involving a small number of regions of the genome, which have large effect size. Given the lack of genetic structure between individuals with alternate genotypes at a genome-wide scale, it is unsurprising that the genetic architecture of tooth size is simple. Ongoing recombination as a result of gene flow between morphotypes would cause a “breaking up" of gene combinations in traits that are determined by more complex genetic architecture. Additionally, alleles for a trait with simple genetic architecture may have stronger selection coefficients resulting in a shift of the migration-selection balance (Yeaman and Whitlock, 2011). It has been suggested that under prolonged periods of adaptation under gene flow complex combinations of small effect genetic regions could be replaced by fewer regions that have larger effect sizes (Yeaman and Whitlock, 2011). Alternatively, it could be the case that tooth size was originally a plastic trait but has undergone genetic assimilation. It has been suggested that lineages that experience adaptive radiations may show high levels of plasticity which may facilitate their adaptation to new environments (West-Eberhard, 2005). Recent research into a similar polymorphism in jaw type in the East African cichlids has demonstrated evidence of genetic assimilation in two candidate genes associated with jaw type (Gunter et al., 2017).

Overall these results provide support for the hypothesis that pharyngeal jaw tooth size has a simple genetic architecture. Despite using a large number of markers (54,458 SNPs), distributed throughout the genome, we were still unable to detect genetic differentiation at a genome-wide scale between alternate

morphotypes in any of the pools we sampled. This suggests that either the evolution of reproductive isolation is relatively recent, and we are therefore unable to detect it, or that alternate morphotypes are being maintained by frequency-dependent selection.

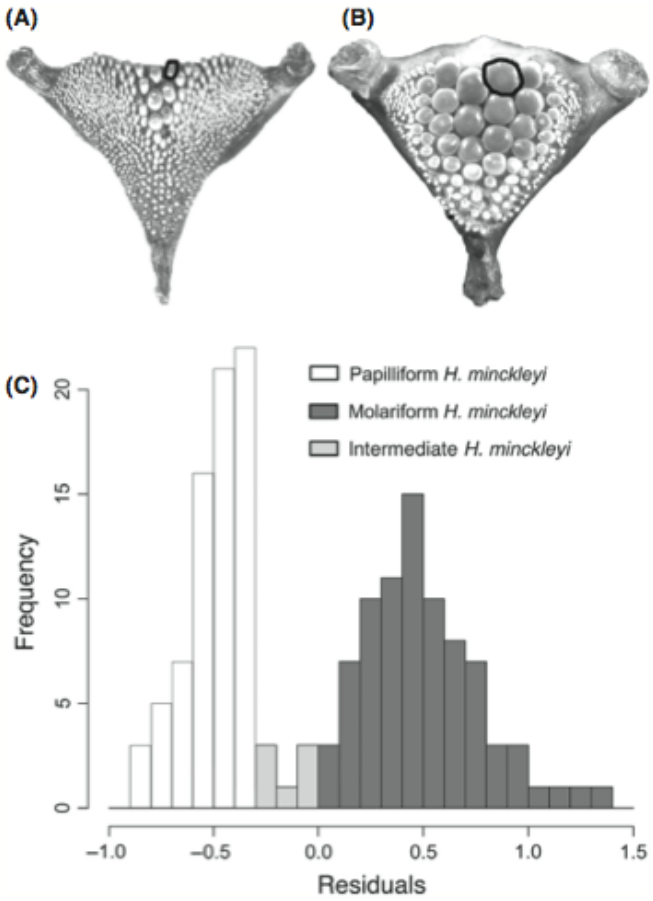

Figure 3.1: Figure provided by C. Darrin Hulsey. Papilliform pharyngeal jaw (A) and molariform pharyngeal jaw (B) of *Herichthys minckleyi*. The pharyngeal tooth size of *H. minckleyi* (C) has a bimodal distribution and individuals with intermediate pharyngeal morphology are rare.

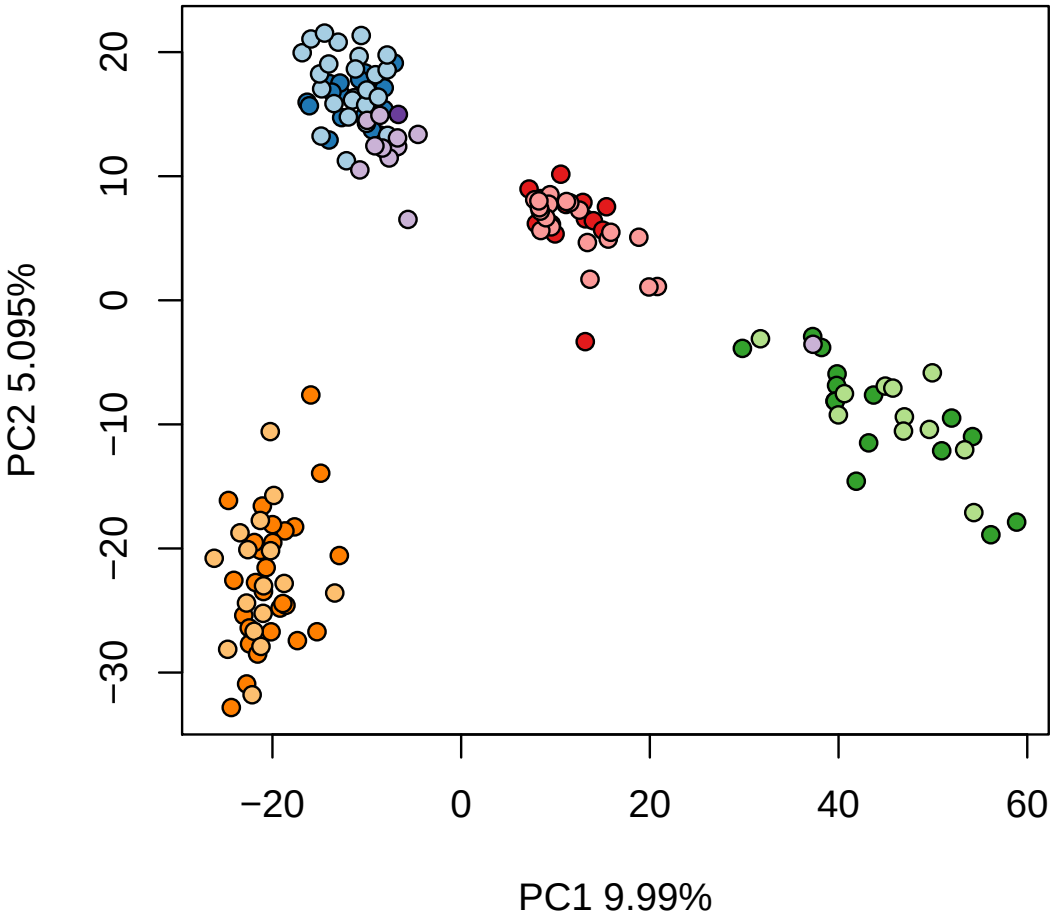

Figure 3.2: Principal component analysis of genotype probabilities estimated from ENTROPY. Dark colors = molariform individuals, light colors= papilliform individuals. Tierra Blanca = blue, Tio Candido = green, Escobedo = red, Juan Santos = orange, Mojarral Este = purple.

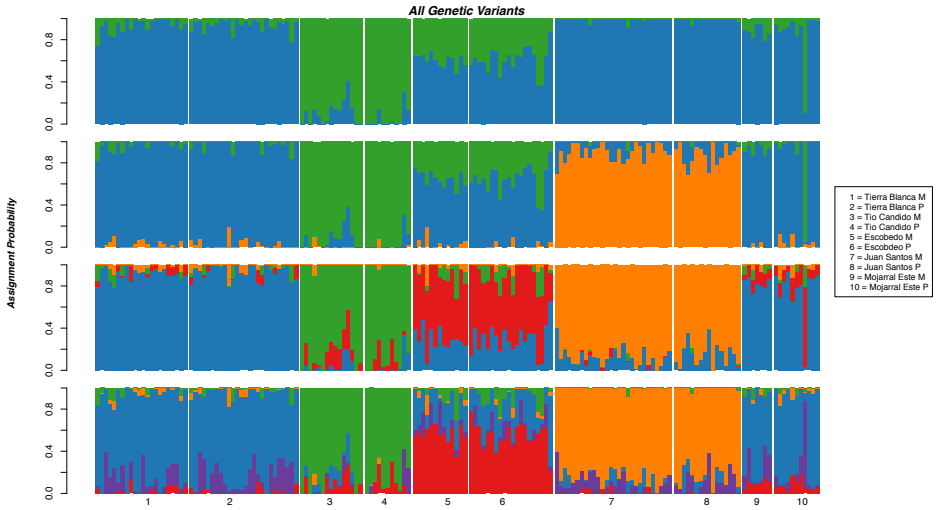

Figure 3.3: Barplot of admixture proportions estimated from ENTROPY for $k = 2$ through $k = 5$.

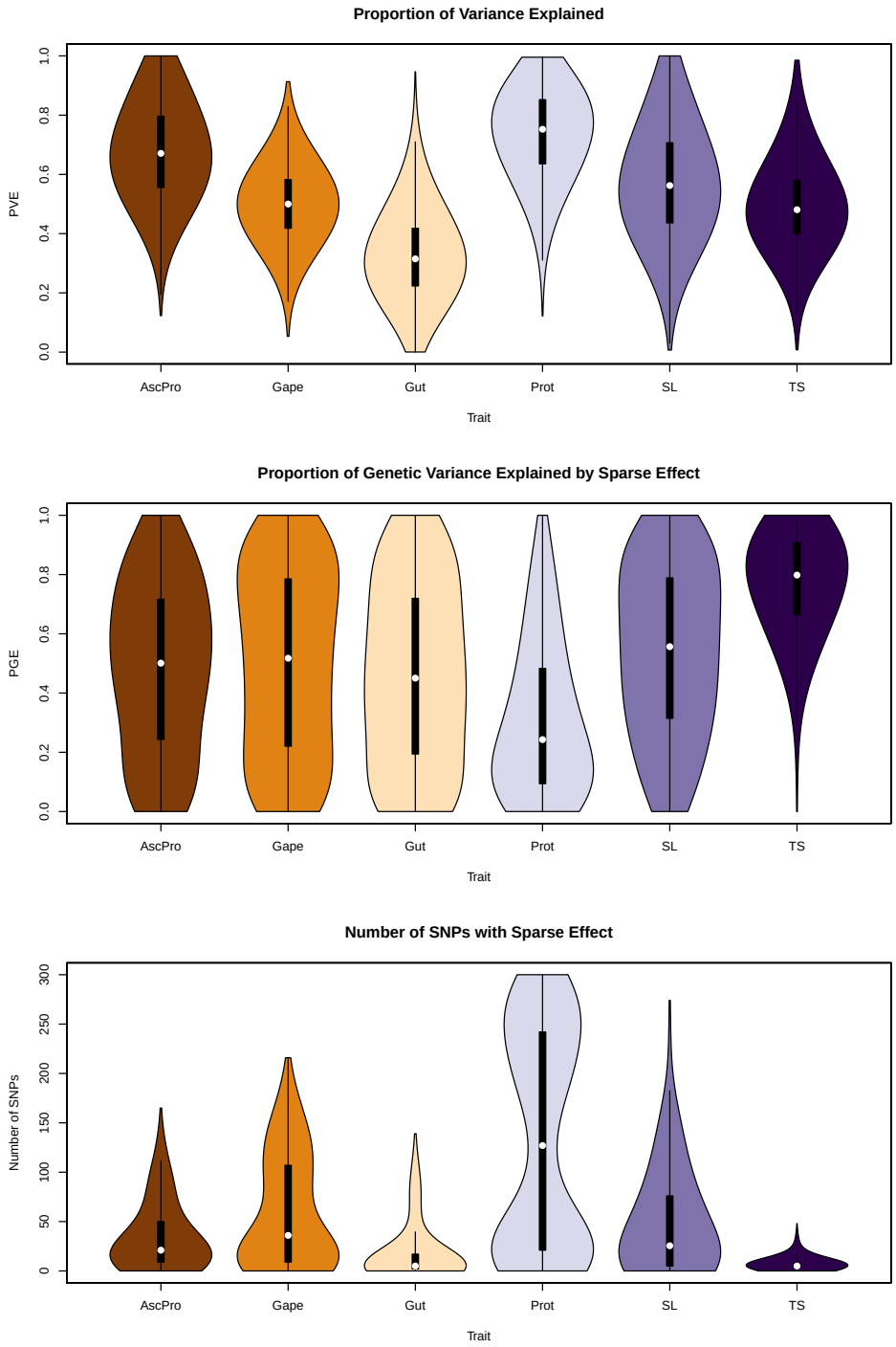

Figure 3.4: Violin plots showing estimates of proportion of variance explained, proportion of genetic variance explained by SNPs with sparse effect, number of SNPs with sparse effect. TS = tooth size, SL = standard length, Prot = jaw protrusion, Gut = gut length, Gape = gape, AscPro = ascending promaxilla length.

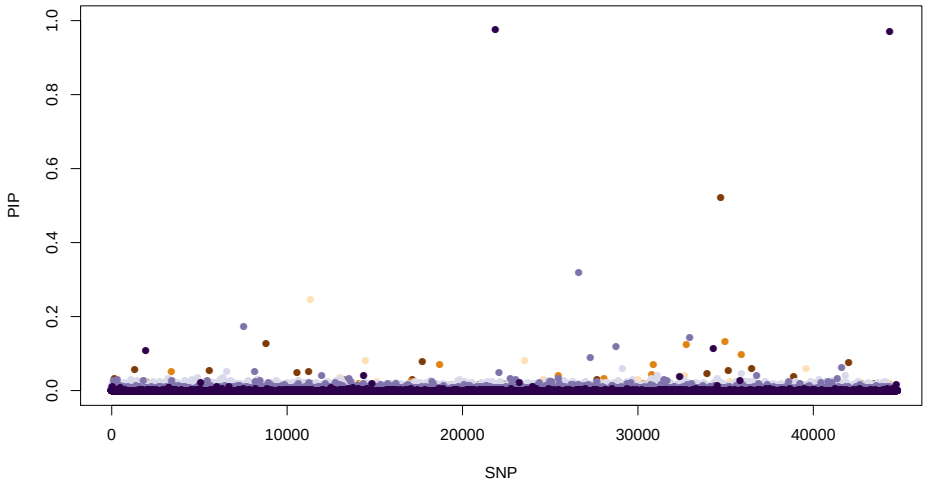

Figure 3.5: Posterior Inclusion Probabilities (PIP's) from BSLMMs conducted using GEMMA for six traits of interest. Ascending premaxilla (AscPro) = dark brown, gape = medium brown, gut length = light brown, protrusion = light purple, standard length = medium purple, standardized tooth size = dark purple.

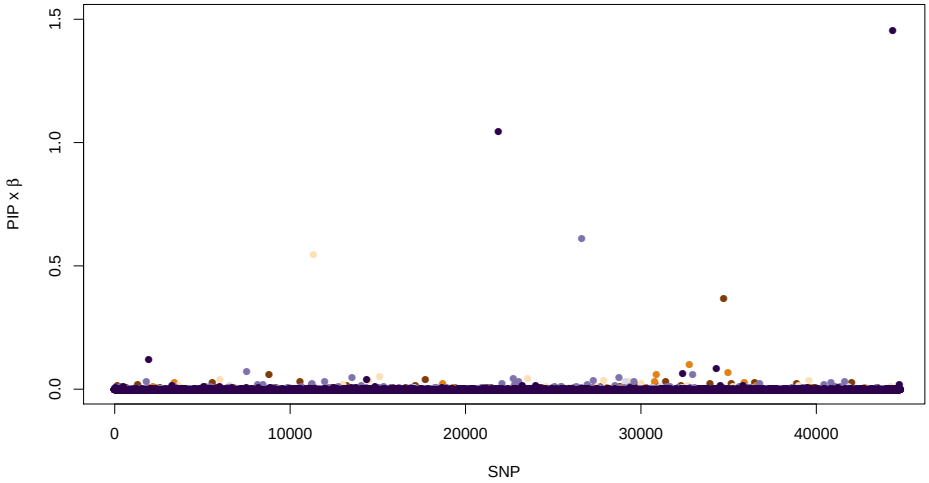

Figure 3.6: Posterior Inclusion Probabilities (PIP's) multiplied by β from BSLMMs conducted using GEMMA for six traits of interest. Ascending premaxilla (AscPro) = dark brown, gape = medium brown, gut length = light brown, protrusion = light purple, standard length = medium purple, standardized tooth size = dark purple.

Table 3.1: Number of *H. minckleyi* Individuals Sampled from Cuatro Ciénegas Valley

Site Location	**Morphotype**	**Sample Size**
Tierra Blanca	Molariform	22
Tierra Blanca	Papilliform	26
Tio Candido	Molariform	12
Tio Candido	Papilliform	11
Tio Candido	Intermediate	3
Escobedo	Molariform	13
Escobedo	Papilliform	20
Juan Santos	Molariform	27
Juan Santos	Papilliform	16
Juan Santos	Intermediate	1
Mojarral Este	Molariform	6
Mojarral Este	Papilliform	11
Mojarral Este	Intermediate	1
Total		**169**

Table 3.2: Hyper Parameter Estimates from GEMMA

Hyper Parameter	**Trait**	**Mean**	**Median**	**2.5%**	**97.5%**
Number of SNPs in Model	AscPro	34.40013	21	1	120
	Gape	5 9.7146	36	1	180
	Gut Length	17.77092	5	0	102
	Jaw Protrusion	133.6163	127	1	296
	Standard Length	47.41639	25.5	1	162
	Tooth Size	7.18823	5	2	26
Proportion of Variance Explained (Sparse Effect)	AscPro	0.4836815	0.5009092	0.001789991	0.962605
	Gape	0.5022337	0.5175771	0.0009515083	0.9800432
	Gut Length	0.4582242	0.4503366	0	0.973079
	Jaw Protrusion	0.3108065	0.2428587	0.001350732	0.9006956
	Standard Length	0.5420668	0.5565053	0.01364948	0.9797276
	Tooth Size	0.7709527	0.798245	0.4066285	0.9913706
Proportion of Variance Explained (Sparse and Polygenic effects)	AscPro	0.676966	0.6710369	0.3705883	0.9859932
	Gape	0.5007443	0.5000596	0.2739701	0.7292987
	Gut Length	0.3275189	0.31492	0.08903117	0.635996
	Jaw Protrusion	0.7349283	0.7523858	0.4121012	0.9576438
	Standard Length	0.5748645	0.5623594	0.231767	0.9610433
	Tooth Size	0.4966001	0.4808556	0.2702815	0.8059733

4. HOW MUCH OF GENOMIC DIFFERENTIATION IS REPEATABLE?: A CONTINENT- AND GENOME-WIDE COMPARISON OF PATTERNS

Introduction

Recent research into the process of speciation has focused on the distribution of genetic variation across the genome. Patterns of differential genomic variation underlie some of the most recent proposed models of speciation, such as ecological speciation (Nosil, 2008; Schluter, 2009). This model proposes that speciation can occur even when only a small proportion of the genome is differentiated (Wu, 2001; Nosil, 2008; Schluter, 2009). If selection acts strongly on regions of the genome associated with fitness and reproductive isolation then the process of speciation may continue even in the face of ongoing gene flow. In order to test this model in natural populations researchers often conduct genome scans in order to identify regions that show elevated levels of differentiation, most frequently this is measured as F_{ST}. Studies that use this approach have demonstrated that differentiation is highly variable across the genome, supporting the model of ecological speciation..

There are, however, criticisms of this method as a way of identifying regions that have been under positive divergent selection. Variation in differentiation across the genome may not necessarily indicate ongoing gene flow with divergent selection leading to increased differentiation. For example, incomplete lineage sorting can commonly lead to patterns misinterpreted as gene flow between

diverging species (Hey, 2006; Noor and Bennett, 2009). Additionally, estimates of F_{ST} represent a relative measure of differentiation, if within population diversity is low due to restricted recombination for example, then this could cause F_{ST} to be elevated (Noor and Bennett, 2009; Cruickshank and Hahn, 2014; Burri, 2017). In order to address these issues recent studies have begun to include additional measures of genomic differentiation such as absolute diversity (D_{XY}), and within population measures of divergence such as average number of nucleotide differences (π) and the number of segregating sites (θ) (Cruickshank and Hahn, 2014; Burri, 2017).

In addition to including a more diverse range of statistics, studies can also use a comparative approach to explore the repeatability of genomic patterns of differentiation across population comparisons (Delmore et al., 2018). If genomic features are conserved across lineages and are significantly involved in the maintenance of genomic patterns of differentiation, then one might expect to find that differentiation is correlated across comparisons. Comparing these relationships across multiple evolutionary scales may reveal temporal differences in the impact of evolutionary processes (Burri, 2017; Delmore et al., 2018). It has been suggested that patterns of genomic differentiation will reflect features of the genomic landscape more at later stages of speciation as evolutionary processes may take time to influence differentiation at these features (Burri, 2017).

We take advantage of the diverse evolutionary relationships present in a butterfly species complex to explore the distribution of genomic patterns of divergence across multiple lineage comparisons. *Lycaeides* represent a diverse genus of butterflies with a complex evolutionary history (Gompert et al., 2008). Nabokov

(1949) revised the classification of *Lycaeides* in North America, recognizing two species; *L. idas* and *L. melissa*, in subsequent years more than 17 subspecies have been described. More recent molecular and morphological analyses suggests in North American there are four nominal species (*L. idas*, *L. melissa*, *L. anna*, and *L. ricei*) and multiple populations of independent hybrid origin occurring in localized regions (e.g. Whites/Sierras, Alpine, Jackson, and Warner) (Gompert et al., 2014). Lineages are both ecologically and morphologically differentiated (Lucas et al., 2018; Forister et al., 2006; Lucas et al., 2014), and vary in host plant use. One entity whose evolutionary relationships to other members of the *Lycaeides* species complex remains unclear is the federally endangered Karner blue butterfly (*L. melissa samuelis*). The Karner blue is found from Minnesota to New Hampshire and its host plant use is restricted to a single species. In the twentieth century habitat for the Karner blue has decreased substantially (Forister et al., 2011). The Karner blue was originally described as a sub-species of *L. melissa* by Nabokov (1949) based on wing pattern and genetalic differences, but he later expressed doubts in this designation (Nabokov, 1975). More recent molecular work found limited evidence of ongoing hybridization between the Karner blue and *L. melissa* and found patterns of genome wide genetic differentiation were similar to that estimated between other nominal species of *Lycaeides* in North America (Gompert et al., 2006; Forister et al., 2011). Herein we first establish the evolutionary relationships between the Karner blue butterfly and the other members of the North American *Lycaeides* species complex. We build upon previous work by sampling extensively throughout the genome, and include samples from all nominal species, and hybrid lineages. Next,

we use the diverse range of lineages present in the *Lycaeides* species complex to explore the distribution and repeatability of genetic differentiation and diversity. Specifically, we ask 1) What is the correlation between genetic differentiation and genetic diversity across the genome? 2) How repeatable are these patterns across different evolutionary scales? We compare these measurements at three scales; between nominal species pairs, between geographically isolated populations of species, and finally between hybrid lineages and their parental species.

Study System

Methods

Molecular Methods

We used GBS data from 1593 individual *Lycaeides* butterflies from 67 localities (Table 4.1, Figure 4.1) that include sequence data from Gompert et al. (2014) and Chaturvedi et al. (2018) as well as new DNA sequence data. The individuals from 16 localities that were sequenced specifically for this study were combined with previous DNA sequence data. As described in more detail below, all sequence data were combined and all data were collectively subjected to assembly, quality filtering, variant calling and downstream analyses. In other words, we used raw sequence data for all individuals and did not rely on previous assemblies or processing of the sequence reads. All samples were collected by hand with nets and genomic DNA was extracted with the Qiagen DNeasy 96 Blood and Tissue kit ((Cat. No. 69581; Qiagen Inc., Valencia, CA, USA). Included among individuals sequenced specifically for this study were samples of the endangered Karner blue butterfly from seven localities that were collected

under permit (USFWS permit PRT842392) (Table 4.1). We also include samples from an additional 9 new localities (Table 4.1). We followed the genotyping-by-sequencing protocol described by of (Gompert et al., 2014) and Parchman et al. (2012) to generate markers. DNA was extracted from tissue excised from the thorax of each butterfly using the Qiagen Blood and Tissue kit. A reduced representation genomic library was produced for each individual by digesting genomic DNA with two restriction enzymes, EcoR1 and Mse1. 8-10bp multiplex identifiers and Illumina adapters were ligated to the restriction fragments and two, independent PCR reactions with the Illumina primers were used to amplify fragments. Amplified fragments were then pooled. This multiplexed genomic library was shipped to the Genome Sequencing and Analysis Facility at the University of Texas, Austin (GSAF) where fragments between 300-450bp were selected with a Blue Pippin (Sage Science) and two lanes of sequencing was performed on the Illumina HiSeq 4000 platform using single-end reads of 100bp in length at the GSAF. These sequence reads were then added to the existing reads from Gompert et al. (2014) and Chaturvedi et al. (2018), which were generated using the same library preparation protocol and sequenced over eight lanes using the Illumina HiSeq 2500 platform generating 100bp, single end reads. Sequence reads from the reduced representation libraries were aligned to the *L. melissa* genome. The genome of *L.melissa* was sequenced by Dovetail Chicago and Hi-C sequencing, and a Hi rise assembly was conducted. The L90/N90 = 21 scaffolds; 11.420 Mb. The expected number of chromosomes for *L. melissa* is 24.

We used the MEM algorithm from BWA 0.7.12-r1039 to align 3.18 billion 100bp

single-end reads to our draft *L. melissa* genome. We called variants using the Genome Analysis Tool Kit (GATK) (McKenna et al., 2010). Sequence alignment/map files were sorted, indexed and converted to binary using PICARD (McKenna et al., 2010). The GATK HaplotypeCaller tool was used to calculate genotype probabilities for variant sites with a minimum base quality of 30 required to consider a site as variable. The HaplotypeCaller tool uses a hidden Markov model to calculate likelihoods based on read data and then calculates the posterior genotype probability for each variable site. We then used the GATK GenotypeGVCFs tool for joint genotyping across the individual Genomic Variant Call Format (GVCF) files prodcued by the HaplotypeCaller. We further filtered variant sites with custom scripts requiring that variant sites be at least 3bp apart, have at least 5216 reads (resulting in an average minimum coverage of 2.6x (i.e. 5216/2004 = 2.6)), and a maximum number of reads to 71196 (which represents the mean coverage depth plus 2 standard deviations), at least 20 reads of the alternative allele, a maximum absolute value of 3 for the base quality rank sum test, maximum absolute value of 2.5 for the mapping quality rank sum test, a maximum absolute value of 2 for the read position rank sum test, a minimum ratio of 2 for quality depth, minimum mapping quality of 30, and a maximum of 401 indivduals (20%) with no data.

Genotypes from GATK were used in the program ENTROPY to obtain estimates of admixture proportions and genotype probabilities. ENTROPY implements a Bayesian hierarchical model similar to the program STRUCTURE but incorporates uncertainty in sequence coverage and sequence error (Gompert et al., 2014). ENTROPY requires the number of ancestral populations to be specified, but does

not incorporate any prior information about which cluster individuals are assigned to. We ran the model for $k = 2$ through $k = 9$. Parameter estimates were obtained using Markov Chain Monte Carlo (MCMC), for each k we ran the model for 2 chains of 15,000 steps, a burn-in of 5,000, saving every 10th step. We checked that models had reached a stable sampling distribution using estimates of effective sample size, and checked chains had reached convergence using Gelman and Rubin's diagnostic (Gelman and Rubin, 1992) in R using the package CODA (Plummer et al., 2006).

Population Genetic Analyses

To address our first question we examined population genomic patterns using SNPs with a minor allele frequency $\geq 5\%$ (6,245). In order to visualize genetic relationships between sampling populations we conducted a principal component analysis (PCA) with genotype probabilities in R using the function PRCOMP. Admixture proportions estimated from ENTROPY were plotted for each value of k in order to explore which cluster individuals were assigned to. We used allele frequencies to estimate genome-wide Wright's F_{ST} using custom R scripts. We calculated credible intervals around these estimates using 10,000 bootstrap simulations.

To explore how genetic differentiation and diversity varied across the genome, and estimate repeatability across evolutionary scales we selected a subset of representative populations. These analyses were conducted on all SNPs that were present in more than one individual, we excluded variable sites found in a single individual as these may be the result of sequencing error. This resulted in 21,156

SNPs. We assigned sampling locations to species subgroups based on admixture proportions estimated from ENTROPY, differentiation found in our PCA, and information from previously published studies (Figures 4.1, 4.2, 4.3, **??**, and Table 4.1). We estimated summary statistics between nominal species pairs using populations of *L. anna* and *L. ricei*, and between the Karner blue edge population (*L. melissa samuelis*) and *L. melissa* east population. For comparisons within species we compared *L. melissa* east population to *L. melissa* west populations, and compared the Karner blue edge (*L. melissa samuelis*) to the Karner blue center (*L. melissa samuelis*). Finally we compared the four hybrid lineages to populations of each putative parental species. For each pairwise comparison we estimated Wright's F_{ST} using a sliding window in order to explore how F_{ST} varies across the genome. Calculations were conducted using custom R scripts, we used a window size of 100kb, with a step size of 10kb, and required a minimum of two SNPs per window. We estimated π, the average number of nucleotide differences, and Watterson's θ, the number of segregating sites, for each population using the same sliding window sizes using BCFTOOLS and SAMTOOLS. We plotted summary statistics across the genome for each lineage comparison. We also calculated the correlation coefficient between summary statistics for each linkage group, and compared these across lineages. Calculations of π and θ were carried out on all sequence reads.

Results

Our genotype-by-sequencing approach resulted in 21,166 SNPs distributed throughout the genome. In order to explore the relationship of the Karner blue

butterfly (*L. melissa samuelis*) with the other lineages in the *Lycaeides* species complex (Figure 4.1) we used a PCA to compare genotype probabilities (for SNPs with minor allele frequency $\geq 0.5\%$, 6,245 SNPs) across all individuals (Figure 4.2). Karner blue butterfly samples were clearly differentiated from the rest of North American *Lycaeides*. The first PCA axis divided the two Karner blue sample groups from all other groups and explained 19.27% of the variation in genotype probabilities. The PC2 axis showed a less clear distinction but appears to separate all three *L. melissa* groups (green), from the four hybrid lineages in the center (blue), followed by the other nominal species *L. idas* (pink), *L. anna* (orange), and *L. ricei* (red). We plotted admixture proportions estimated from ENTROPY for $k = 2$ through $k = 9$ (Figure 4.3 and 4.4). For the $k = 2$ model we found that the Karners sampling group formed its own cluster, while all other lineages clustered together. We found that the $k = 7$ appeared to provide the clearest resolution in clusters, the $k = 8$ did not add any additional distinctive groups. Our calculations of genome-wide F_{ST} were overall low, even between nominal species groups (Figure 4.5). Curiously the highest F_{ST} estimate was between the two Karner populations, but this was still overall quite low (0.05). We also estimated credible intervals around these estimates using 10,000 bootstrap simulations. We found CI's showed a relatively high level of certainty in the F_{ST} estimates, and there were no CI's that overlapped zero.

We calculated summary statistics for several lineage pairs at varying evolutionary scales. First between nominal species, between geographically isolated populations within species, and finally between hybrid lineages and their putative parental species. We first calculated correlations for each lineage

between F_{ST}, θ, and π for each linkage group. We found that 30 comparisons showed significant correlations out of 72 comparisons total (Figure 4.6). There tended to be more positive correlations between F_{ST} and θ than negative. For the comparison between the two Karner populations we found significantly positive correlations for linkage group 11 and linkage group 503, for both populations, indicating that regions within these linkage groups with higher values of θ also showed higher estimates of F_{ST}. We also identified significant positive correlations for the Karner Center population on linkage group 270, and for the Karners Edge population on linkage groups 309 and 588. For linkage group 833 in the Karners Center we found a significantly negative correlation. For the comparison between Melissa East and Karners Edge populations we found a significantly positive correlation for linkage group 1095 in both populations, and found 5 other positive correlations that had lower correlation coefficients. Overall there were limited repeated patterns outside of population comparisons. For example, for linkage group 503 we identified four significantly negative correlations, and four significantly positive correlations across populations.
For comparisons between F_{ST} and π we found 27 significant correlations out of 72 (Figure 4.7). Across linkage group 11 we found some repeatability, where 6 out of 8 comparisons showed significantly negative relationships. However for linkage group 833 we did not find repeatability, four comparisons had significantly positive correlations while 2 had significantly negative ones. The comparison between Melissa East and Melissa Rockies showed the most significant correlations across linkage groups with 6 linkage groups showing significant correlations between F_{ST} and π.

As might be expected we found that all comparisons between θ and π were significantly positively correlated (Figure 4.8). While all correlations were significant, there was also variation in correlation coefficients. Linkage group 1095 had consistently very high correlations, indicating that there is a lower proportion of regions within this linkage group where θ and π deviated from one another. Overall the two Karner blue populations had the lowest correlation coefficients across linkage groups. This could indicate that these populations have a high proportion of regions where θ and π deviate.

Discussion

Our research used extensive geographic and genomic sampling to address two broad questions, first to place the endangered Karner blue butterfly into its evolutionary context among the *Lycaeides* species complex, and second to explore the distribution and repeatability of differentiation across the genome (Table 4.1 and Figure 4.1). Understanding how differentiation varies across the genome, and the mechanisms that underlie these differences remains a key challenge in the current research into the process of speciation.

In order to examine the patterns of genome wide differentiation between the Karner blue and other *Lycaeides* lineages we first used a PCA on genotype probabilities to visualize relationships. We found the PC1 axis, explaining 19.27% of variation in genotype probabilities, divided the two Karner blue populations from all other *Lycaeides* lineages (Figure 4.2). This indicates that based on genotypes probabilities the Karner blues are the most distinct lineage in the *Lycaeides* species complex. Next we estimated admixture proportions

using a hierarchical Bayesian approach implemented in ENTROPY for $k = 2$ through $k = 9$ (Figures 4.3 and 4.4). When we examined cluster assignment for the $k = 2$ model it showed the two Karner blue populations assigning to their own cluster, while all other populations assigned to cluster two (Figure 4.3). This is in keeping with the results we found from the PCA of genotype probabilities. At higher values of k the Karner blues maintain their own cluster, but across k's there appears to be divergence between the two populations with the Karner edge population showing higher levels of admixture than the Karners center population. One individual for the Karner Center group assigns to the same cluster of *L. melissa* Rockies and east populations, while all individuals from the Karners edge group show admixture from all three *L. melissa* populations. Our calculations of genome wide F_{ST} were overall low, but all had credible intervals that did not overlap zero. Surprisingly we found the highest F_{ST} was between the two Karner blue populations. In general, patterns of F_{ST} did not entirely reflect the patterns that were found for our two other analyses. Given how low values of F_{ST} were it may be that there is not enough resolution between genome wide estimates to clearly delineate evolutionary lineages. Overall, these results are consistent with previous research which has shown low levels of gene flow between the Karner blues and *L. melissa* (Forister et al., 2011). However, in contrast to previous research, we detected genetic differences within the Karner blues, between two populations found in the center of the range, and populations found at the edge. It may be that populations at the edge of the range may exhibit higher patterns of gene flow between neighboring species such as *L. melissa* populations or *L. idas* populations to the North.

Overall we found variation in correlations between summary statistics both across linkage groups, and across population comparisons. We found that comparisons of θ and π were the most consistent, and all comparisons had a significantly positive correlation (Figure 4.8). In a neutral population, being affected primarily by mutation and drift, the expectation would be that θ and π are very similar. Although all correlations were significantly positive, there was variation in correlation coefficients. We found that across populations linkage group 1095 had the highest correlation coefficients, this indicates that regions within this linkage group show very little deviation between θ and π and could have fewer regions experiencing selection relative to other linkage groups. We found that the two Karner blue populations consistently had the lowest correlation coefficients across linkage groups (with the exception of 1095) indicating that they had a high proportion of regions where there were deviations between θ and π. It has been documented that the Karner blue is experiencing habitat loss and changes in population size are known to cause differences between estimates of θ and π (Andow et al., 1994). Alternatively these patterns could be associated with selection.

For comparisons of F_{ST} to θ and π we found low repeatability both across linkage groups and across populations, with a small number of exceptions. For comparisons of F_{ST} and θ we found 30 significant correlations (Figure 4.6). We found that there tended to be a higher number of significantly positive correlations. Previous research has suggested that if within population diversity is low this could lead to inflation of estimates of F_{ST} (Cruickshank and Hahn, 2014), in regions where F_{ST} and θ are positively correlated it seems unlikely that

F_{ST} estimates are being inflated by low within population diversity. We found that comparisons between the Karner populations (either pairwise comparisons between the two populations, or between Karners Edge and Melissa East) tended to have the highest correlation coefficients. If the two Karner populations have undergone a reduction in population size then the impact of drift could be more influential and could cause increased F_{ST}. Where we identified significantly negative correlations between F_{ST} and θ we tended to identify these in both populations in the pairwise comparison. However, for the pairwise comparison between the two Karner populations we found that the Karners Center population had negative correlation for linkage group 833, but did not find a significant correlation for Karners Edge. It is possible that for this linkage group there was low genetic diversity in the Karners Center population which could lead to artificially inflated F_{ST} estimates between the two species for these regions. When we compared F_{ST} to π we found that 27 comparisons has significant correlations (Figure 4.7). For linkage group 11 we found some repeatability, with six out of eight comparisons showing negative correlations. This is in contrast to what we found for comparisons of F_{ST} and θ indicating that there may be some divergence in estimates of θ and π, even though they are significantly correlated with one and other.

To summarize, we found that the Karner blue butterfly (*L. melissa samuelis*) appears to be a distinct evolutionary lineage, and not a sub species of *L. melissa*. We found that the Karner blue shows the largest genetic differences compared to other *Lycaeides* lineages in the North American species complex (Figures 4.2, 4.3, and 4.5). When we compared patterns of differentiation and diversity across

the genome, and across population comparisons we found that comparisons between θ and π were the most repeatable although they still varied in scale (Figure 4.8). We found patterns that support the conclusion that the Karner blues have been through changes in population size, their populations had the lowest correlation coefficients between θ and π. Correlations between F_{ST} and θ and π were variable, but we did find some examples of repeatability such as the significantly negative correlations for 6 out of 8 populations, for comparisons of F_{ST} and π for linkage group 11 (Figure 4.6 and 4.7). In the future the underlying causes of these patterns could be investigated further by estimating gene density, and by using high coverage sequence data to estimate D_{XY}.

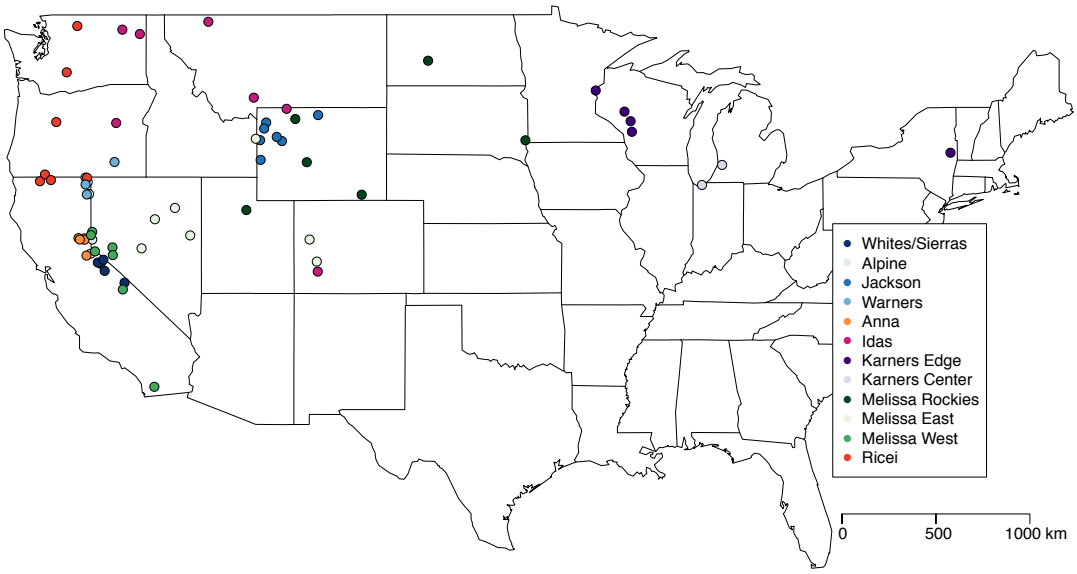

Figure 4.1: Map of sampling locations across North America. Full details for localities can be found in Table 4.1.

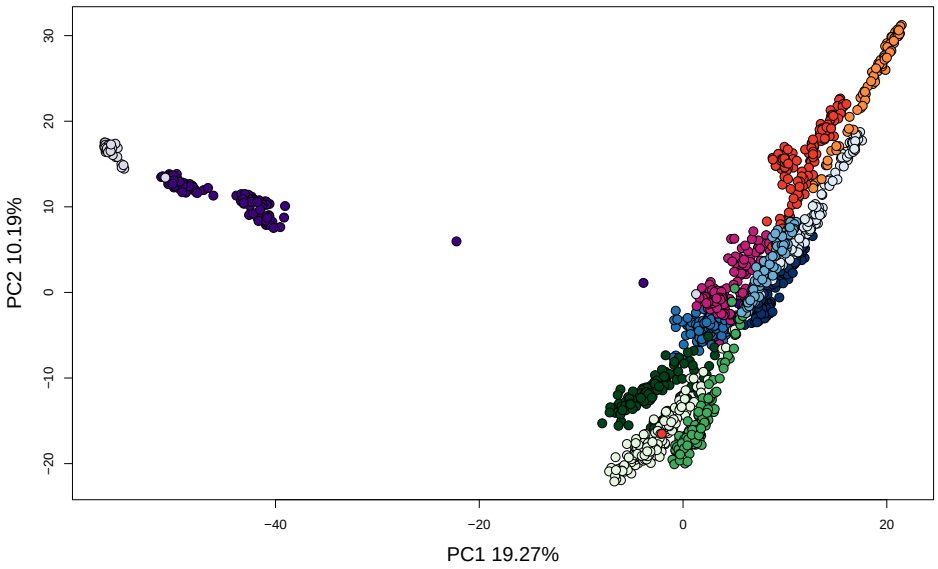

Figure 4.2: Principal Component Analysis (PCA) on genotype probabilities across 21,166 SNPs. Each data point represents one individuals genotype probabilities across all 21,166 SNPs. Data points are color coded by their *Lycaeides* lineage. Light purple = Karners center, dark purple = Karners edge, darkest blue = Whites/Sierra hybrids, blue = Jackson hybrids, light blue = Warners hybrids, lightest blue = Alpines, orange = *L. anna*, pink = *L. idas*, dark green = *L. melissa* Rockies, green = *L. melissa* west, light green = *L. melissa* east, red = *L. ricei*.

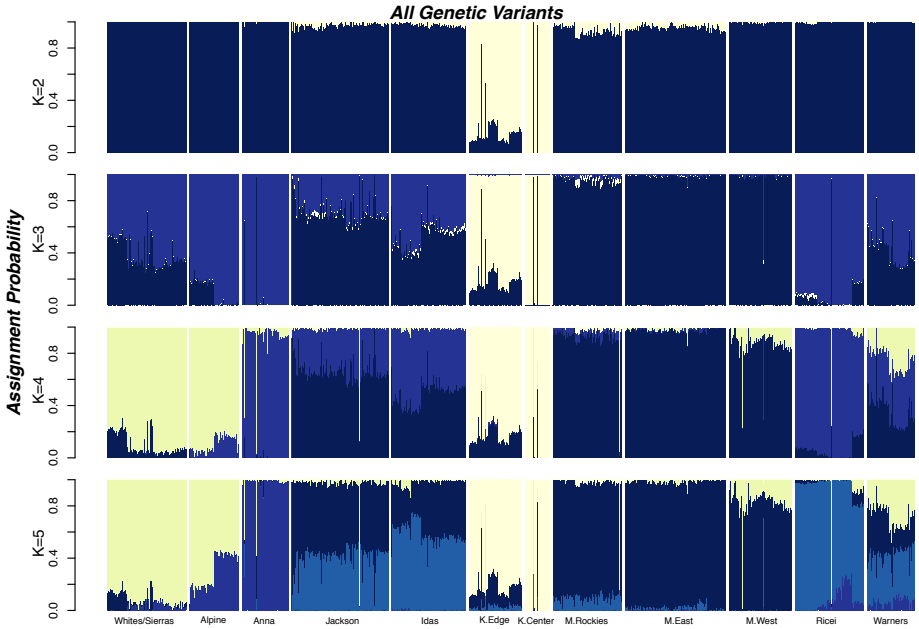

Figure 4.3: Plot of admixture proportions for k = 2 through k=5.

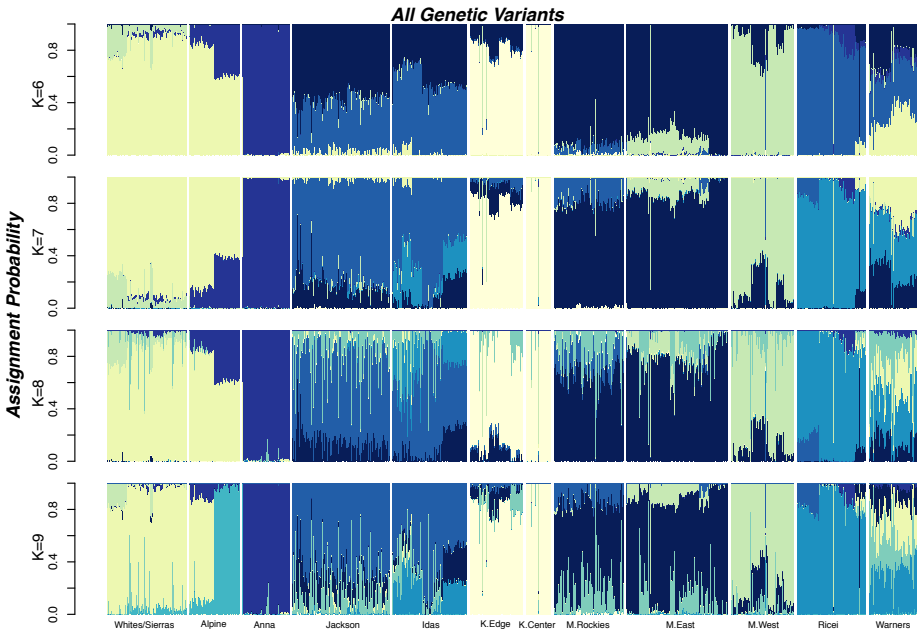

Figure 4.4: Plot of admixture proportions for $k = 6$ through k=9.

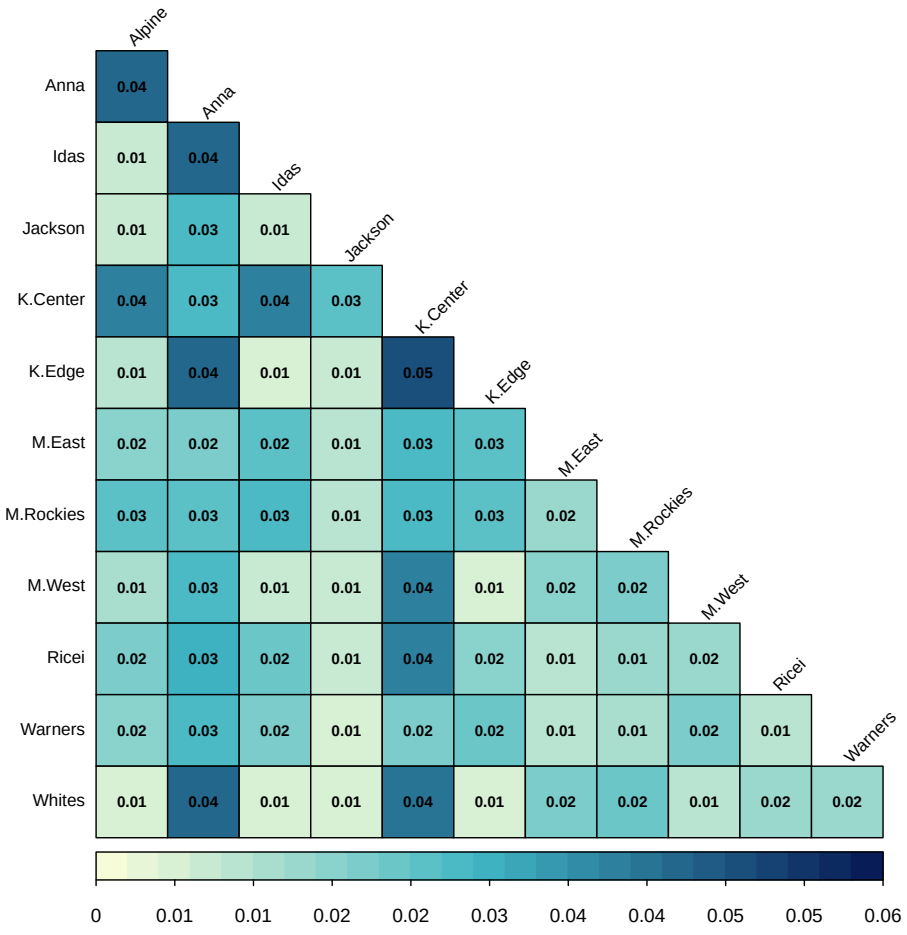

Figure 4.5: Estimates of Genome-Wide F_{ST} for pairwise comparisons between all evolutionary lineages, across all SNPs with minor allele frequency $\geq 0.5\%$ (6,245).

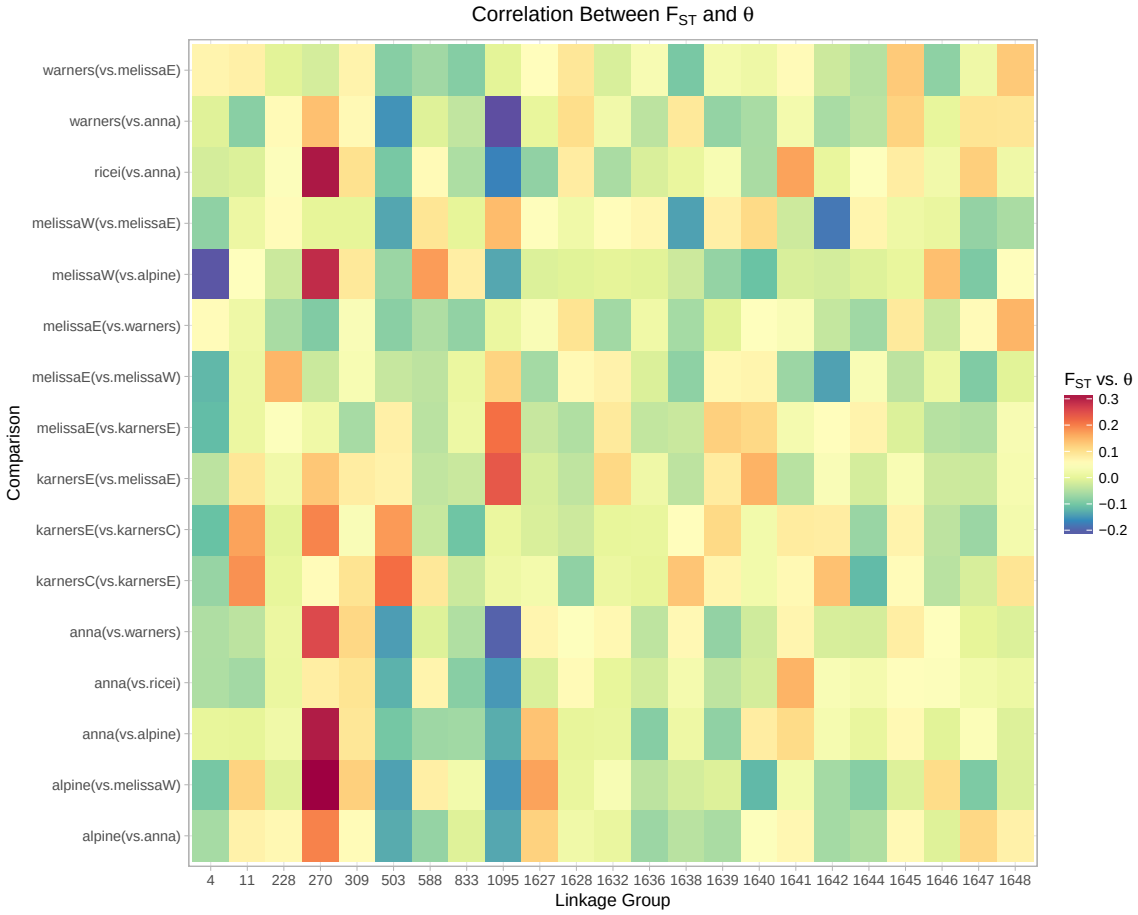

Figure 4.6: Correlation between pairwise F_{ST} and population estimates of θ.

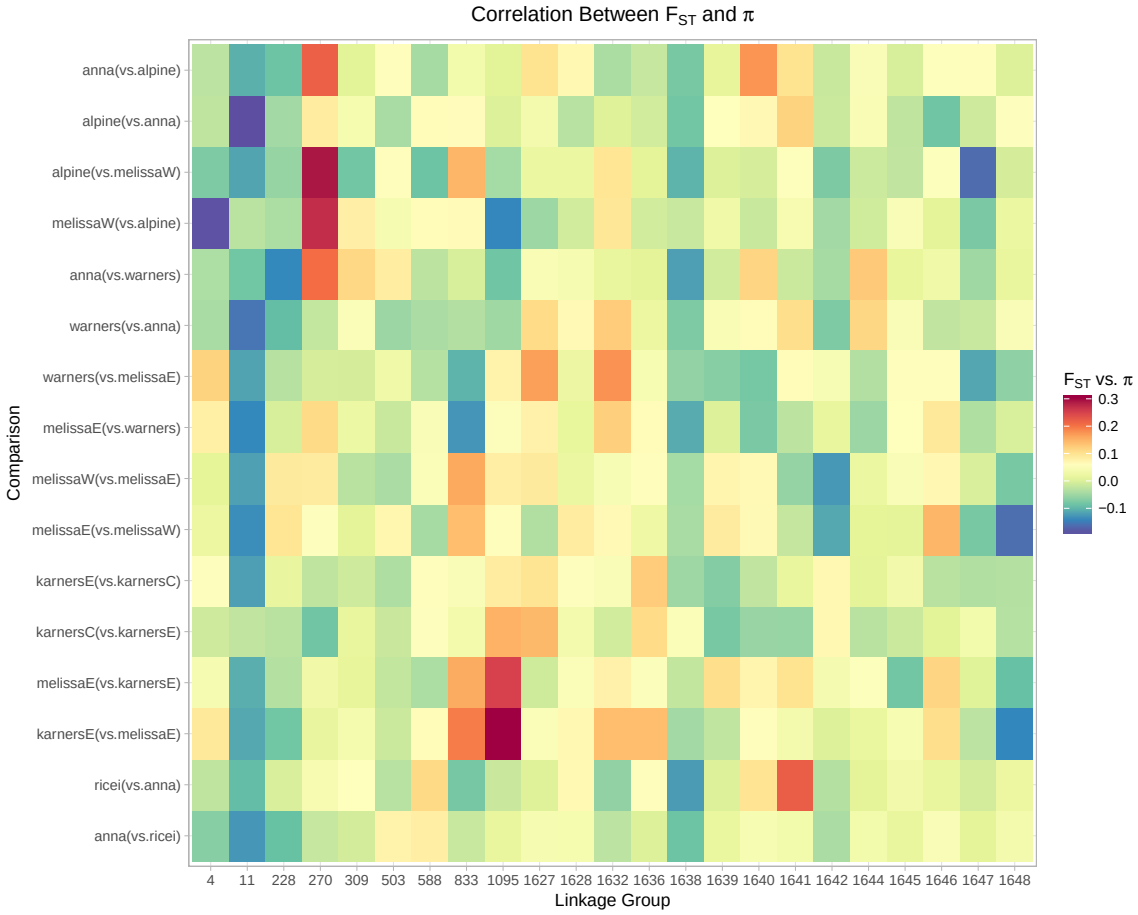

Figure 4.7: Correlation between pairwise F_{ST} and population estimates of π for each linkage group.

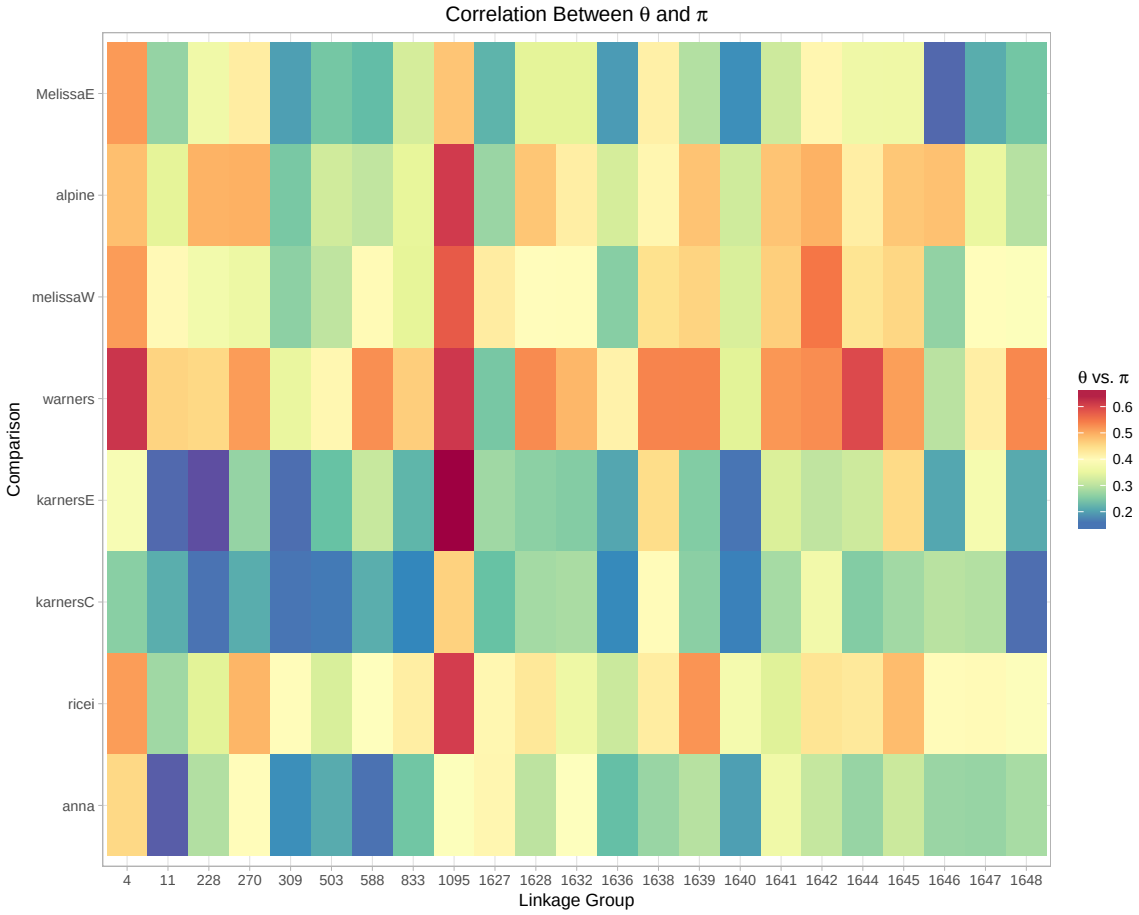

Figure 4.8: Correlation between population estimates of θ and π for each linkage group.

Table 4.1: Sample information for 67 *Lycaeides* collection localities. Locality numbers, names, nominal species designations, Subgroup assignment (see text for details) latitude, longitude, and number of individuals are provided. The last column indicates previously published sequence data (G&C) (Gompert et al., 2014; Chaturvedi et al., 2018) or sequence data presented here for the first time (Present).

#	**Locality**	**Nominal Species**	**Subgroup**	**Lat. N**	**Long. W**	**# Ind.s**	**Data**
1	County Line Hill, CA	*hybrid*	Whites/Sierra	37.46	118.19	40	G&C
2	Lake Emma, CA	*hybrid*	Whites/Sierra	38.28	119.48	33	G&C
3	Sonora Pass, CA	*hybrid*	Whites/Sierra	38.33	119.63	40	G&C
4	Sweetwater Mtns., CA	*hybrid*	Whites/Sierra	38.45	119.33	23	G&C
5	Tioga Crest, CA	*hybrid*	Whites/Sierra	37.97	119.26	38	G&C
6	Carson Pass, CA	*hybrid*	Alpine	38.71	120.02	50	G&C
7	Mt. Rose, NV	*hybrid*	Alpine	39.32	119.93	52	G&C
8	Castle Peak, CA	*L. anna*	Anna	39.37	120.35	18	G&C
9	Donner Pass, CA	*L. anna*	Anna	39.31	120.35	18	G&C
10	Fall Creek, CA	*L. anna*	Anna	39.38	120.67	20	G&C
11	Yuba Gap, CA	*L. anna*	Anna	39.32	120.6	20	G&C
12	Leek Springs, CA	*L. anna*	Anna	38.63	120.24	20	G&C
13	Dubois, WY	*hybrid*	Jackson	43.56	109.7	41	G&C
14	Hunt Mt., WY	*hybrid*	Jackson	44.68	107.75	30	G&C

Table 4.1 - *Continued from previous page*

#	**Locality**	**Nominal Species**	**Subgroup**	**Lat. N**	**Long. W**	**# Ind.s**	**Data**
15	Pinnacles Butte, WY	*hybrid*	Jackson	43.74	109.98	20	G&C
16	Periodic Spring, WY	*hybrid*	Jackson	42.75	110.85	20	G&C
17	Riddle Lake, WY	*hybrid*	Jackson	44.36	110.55	30	G&C
18	Rendezvous Mt., WY	*hybrid*	Jackson	43.6	110.88	32	G&C
19	Sheffield Creek, WY	*hybrid*	Jackson	44.1	110.66	26	G&C
20	Garnet Peak, MT	*L. idas*	Idas	45.43	111.22	16	G&C
21	Strawberry Mt.s, OR	*L. idas*	Idas	44.34	118.64	20	G&C
22	Siyeh Creek, MT	*L. idas*	Idas	48.7	113.67	20	G&C
23	Tomboy Road, CO	*L. idas*	Idas	37.94	107.77	24	G&C
24	Tibbs Butte, MT	*L. idas*	Idas	44.95	109.45	20	G&C
25	Cotton Wood Divide, WA	*L. idas*	Idas	48.173	12.362	25	Present
26	White Mt. Fire Overlook, WA	*L. idas*	Idas	48.36	118.31	24	Present
27	Black River State Forest, WI	*L. m. samuelis*	Karners Edge	44.42	90.90	17	Present
28	Eau Claire State Forest, WI	*L. m. samuelis*	Karners Edge	44.83	91.23	22	Present
29	Fish Lake, WI	*L. m. samuelis*	Karners Edge	45.74	92.78	20	Present
30	Fort McCoy, WI	*L. m. samuelis*	Karners Edge	43.96	90.83	23	Present

Table 4.1 - *Continued from previous page*

#	**Locality**	**Nominal Species**	**Subgroup**	**Lat. N**	**Long. W**	**# Ind.s**	**Data**
31	Saratoga Springs, NY	*L. m. samuelis*	Karners Edge	43.06	73.65	27	Present
32/33	Allegan, MI	*L. m. samuelis*	Karners Center	42.53	85.97	30	Present
34	Indiana Dunes, IN	*L. m. samuelis*	Karners Center	41.67	87.05	21	Present
35	Albion Meadow, UT	*L. melissa*	M. Rockies	40.59	111.62	46	G&C
36	Beulah, ND	*L. melissa*	M. Rockies	47.02	101.82	10	Present
37	Brandon, SD	*L. melissa*	M. Rockies	43.59	96.57	20	G&C
38	Cody, WY	*L. melissa*	M. Rockies	44.51	108.98	23	G&C
39	Lander, WY	*L. melissa*	M. Rockies	42.65	108.36	24	G&C
40	Yellow Pine CG, WY	*L. melissa*	M. Rockies	41.25	105.4	20	G&C
41	De Beque, CO	*L. melissa*	M. East	39.32	108.21	20	G&C
42	Montrose, CO	*L. melissa*	M. East	38.37	107.82	20	G&C
43	East Creek CG, NV	*L. melissa*	M. East	39.50	114.65	25	Present
44	Goose Lake, CA	*L. melissa*	M. East	41.99	120.29	30	G&C
45	Lamoille Canyon, NV	*L. melissa*	M. East	40.68	115.47	20	G&C
46	Mill Creek, NV	*L. melissa*	M. East	40.19	116.55	24	Present
47	Ophir City, NV	*L. melissa*	M. East	38.94	117.27	19	G&C
48	Surprise Valley, CA	*L. melissa*	M. East	41.28	120.1	20	G&C

Table 4.1 - *Continued from previous page*

#	**Locality**	**Nominal Species**	**Subgroup**	**Lat. N**	**Long. W**	**# Ind.s**	**Data**
49	Upper Alkali Lake, CA	*L. melissa*	M. East	41.79	120.17	20	Present
50	Victor, ID	*L. melissa*	M. East	43.66	111.11	20	G&C
51	Bishop, CA	*L. melissa*	M. West	37.17	118.28	20	G&C
52	Gardnerville, NV	*L. melissa*	M. West	38.81	119.78	18	G&C
53	Red Earth Way, NV	*L. melissa*	M. West	38.98	118.84	20	G&C
54	Silver Lake, NV	*L. melissa*	M. West	39.65	119.93	18	G&C
55	Trout Pond Trailhead, CA	*L. melissa*	M. West	32.98	116.58	13	Present
56	Verdi Crystal, NV	*L. melissa*	M. West	39.51	120	20	G&C
57	Washoe Lake, NV	*L. melissa*	M. West	38.65	118.82	20	G&C
58	Chinook Pass, WA	*L. ricei*	Ricei	46.52	121.31	25	Present
59	Rainy Pass, WA	*L. ricei*	Ricei	48.517	120.736	20	Present
60	Big Lake, OR	*L. ricei*	Ricei	44.38	121.87	20	G&C
61	Soda Mt., OR	*L. ricei*	Ricei	42.12	122.48	20	G&C
62	Marble Mts., CA	*L. ricei*	Ricei	41.83	122.75	12	G&C
63	Shovel Creek, CA	*L. ricei*	Ricei	41.88	122.16	20	G&C
64	Cave Lake, CA	*L. ricei*	Ricei	41.98	120.21	24	G&C
65	Buck Mt., CA	*hybrid*	Warners	41.69	120.29	44	G&C

Table 4.1 - *Continued from previous page*

#	**Locality**	**Nominal Species**	**Subgroup**	**Lat. N**	**Long. W**	**# Ind.s**	**Data**
66	Eagle Peak, CA	*hybrid*	Warners	41.26	120.22	40	G&C
67	Steens Mountain, OR	*hybrid*	Warners	42.66	118.73	13	G&C

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., et al. (2013). Hybridization and speciation. *Journal of evolutionary biology*, 26(2):229–246.

Abbott, R. J. (2017). Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *Journal of Systematics and Evolution*, 55(4):238–258.

Albertson, R. C., Streelman, J. T., and Kocher, T. D. (2003). Genetic basis of adaptive shape differences in the cichlid head. *Journal of Heredity*, 94(4):291–301.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Andow, D. A., Baker, R. J., and Lane, C. P. (1994). *Karner blue butterfly: a symbol of a vanishing landscape*. Minnesota Agricultural Experiment Station.

Baack, E. J. and Rieseberg, L. H. (2007). A genomic view of introgression and hybrid speciation. *Current opinion in genetics & development*, 17(6):513–518.

Barth, J. M., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., Jakobsen, K. S., Johannesson, K., Jorde, P. E., Knutsen, H., et al. (2017). Genome architecture enables local adaptation of atlantic cod despite high connectivity. *Molecular ecology*.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*, 16(1):113–148.

Bolnick, D. I. and Fitzpatrick, B. M. (2007). Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, pages 459–487.

Borge, T., Lindroos, K., Nadvornik, P., Syvänen, A.-C., and Sætre, G.-P. (2005). Amount of introgression in flycatcher hybrid zones reflects regional differences in pre and post-zygotic barriers to gene exchange. *Journal of evolutionary biology*, 18(6):1416–1424.

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., Zhao, K., Brisbin, A., Parker, H. G., Cargill, M., et al. (2010). A simple genetic architecture underlies morphological variation in dogs. *PLoS biology*, 8(8):e1000451.

Buerkle, C. A. and Lexer, C. (2008). Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, 23(12):686–694.

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 1(3):118–131.

Bush, G. L. (1994). Sympatric speciation in animals: new wine in old bottles. *Trends in Ecology & Evolution*, 9(8):285–288.

Chaturvedi, S., Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., and Gompert, Z. (2018). The predictability of genomic changes underlying a recent host shift in melissa blue butterflies. *Molecular ecology*.

Chaves-Campos, J., Johnson, S. G., De León, F. J. G., and Hulsey, C. D. (2011a). Phylogeography, genetic structure, and gene flow in the endemic freshwater shrimp palaemonetes suttkusi from cuatro ciénegas, mexico. *Conservation Genetics*, 12(2):557–567.

Chaves-Campos, J., Johnson, S. G., and Hulsey, C. D. (2011b). Spatial geographic mosaic in an aquatic predator-prey network. *PloS one*, 6(7):e22472.

Coghill, L. M., Hulsey, C. D., Chaves-Campos, J., de Leon, F. J. G., and Johnson, S. G. (2013). Phylogeography and conservation genetics of a distinct lineage of sunfish in the cuatro ciénegas valley of mexico. *PloS one*, 8(10):e77013.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*, volume 37. Sinauer Associates Sunderland, MA.

Crawford, J. E., Riehle, M. M., Guelbeogo, W. M., Gneme, A., Sagnon, N., Vernick, K. D., Nielsen, R., and Lazzaro, B. P. (2015). Reticulate speciation and barriers to introgression in the anopheles gambiae species complex. *Genome biology and evolution*, 7(11):3116–3131.

Cruickshank, T. E. and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology*, 23(13):3133–3157.

Darwin, C. (1859). On the origin of species by means of natural selection. *Murray, London*.

Delmore, K. E., Lugo Ramos, J. S., Van Doren, B. M., Lundberg, M., Bensch, S., Irwin, D. E., and Liedvogel, M. (2018). Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evolution Letters*, 2(2):76–87.

Dupuis, J. R. and Sperling, F. A. (2015). Repeated reticulate evolution in north american papilio machaon group swallowtail butterflies. *PloS one*, 10(10):e0141882.

Ellison, A. M., Butler, E. D., Hicks, E. J., Naczi, R. F., Calie, P. J., Bell, C. D., and Davis, C. C. (2012). Phylogeny and biogeography of the carnivorous plant family sarraceniaceae. *PLoS One*, 7(6):e39291.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Feder, J. L., Egan, S. P., and Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7):342–350.

Forister, M. L., Fordyce, J. A., McCall, A. C., and Shapiro, A. M. (2011). A complete record from colonization to extinction reveals density dependence and the importance of winter conditions for a population of the silvery blue, glaucopsyche lygdamus. *Journal of Insect Science*, 11.

Forister, M. L., Fordyce, J. A., Nice, C. C., Gompert, Z., and Shapiro, A. M. (2006). Egg morphology varies among populations and habitats along a suture zone in the lycaeides idas-melissa species complex (lepidoptera: Lycaenidae). *Annals of the Entomological Society of America*, 99(5):933–937.

Fraser, G. J., Hulsey, C. D., Bloomquist, R. F., Uyesugi, K., Manley, N. R., and Streelman, J. T. (2009). An ancient gene network is co-opted for teeth on old and new jaws. *PLoS biology*, 7(2):e1000031.

Gavrilets, S., Vose, A., Barluenga, M., Salzburger, W., and Meyer, A. (2007). Case studies and mathematical models of ecological speciation. 1. cichlids in a crater lake. *Molecular Ecology*, 16(14):2893–2909.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Gompert, Z. and Buerkle, C. A. (2011). A hierarchical bayesian model for next-generation population genomics. *Genetics*, 187(3):903–917.

Gompert, Z. and Buerkle, C. A. (2016). What, if anything, are hybrids: enduring truths and challenges associated with population structure and gene flow. *Evolutionary applications*, 9(7):909–923.

Gompert, Z., Fordyce, J. A., Forister, M. L., and Nice, C. C. (2008). Recent colonization and radiation of north american lycaeides (plebejus) inferred from mtdna. *Molecular phylogenetics and evolution*, 48(2):481–490.

Gompert, Z., Lucas, L. K., Buerkle, C. A., Forister, M. L., Fordyce, J. A., and 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*, 23:4555–4573.

Gompert, Z., Lucas, L. K., Nice, C. C., and Buerkle, C. A. (2013). Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides idas* and *L. melissa*. *Evolution*, 67(9):2498–2514.

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

Gompert, Z., Mandeville, E. G., and Buerkle, C. A. (2017). Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, 48.

Gompert, Z., Nice, C. C., Fordyce, J. A., Forister, M. L., and Shapiro, A. M. (2006). Identifying units for conservation using molecular systematics: the cautionary tale of the karner blue butterfly. *Molecular ecology*, 15(7):1759–1768.

Gompert, Z., Parchman, T. L., and Buerkle, C. A. (2012b). Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):439–450.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, pages genetics–112.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Bustamante, C. D., Altshuler, D. L., et al. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988.

Gunter, H. M., Schneider, R. F., Karner, I., Sturmbauer, C., and Meyer, A. (2017). Molecular investigation of genetic assimilation during the rapid adaptive radiations of east african cichlid fishes. *Molecular ecology*, 26(23):6634–6653.

Harrison, R. G. and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1):795–809.

Hendry, A. P., Huber, S. K., De Leon, L. F., Herrel, A., and Podos, J. (2009). Disruptive selection in a bimodal population of darwin's finches. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):753–759.

Hey, J. (2006). Recent advances in assessing gene flow between diverging populations and species. *Current opinion in genetics & development*, 16(6):592–596.

Hulsey, C., Marks, J., Hendrickson, D., Williamson, C., Cohen, A., and Stephens, M. (2006). Feeding specialization in *Herichthys minckleyi*: a trophically polymorphic fish. *Journal of Fish Biology*, 68(5):1399–1410.

Hulsey, C. D., Bell, K. L., García-de León, F. J., Nice, C. C., and Meyer, A. (2016). Do relaxed selection and habitat temperature facilitate biased mitogenomic introgression in a narrowly endemic fish? *Ecology and evolution*, 6(11):3684–3698.

Hulsey, C. D. and García-de León, F. J. (2013). Introgressive hybridization in a trophically polymorphic cichlid. *Ecology and evolution*, 3(13):4536–4547.

Hulsey, C. D., Hendrickson, D. A., and de León, F. G. (2005). Trophic morphology, feeding performance and prey use in the polymorphic fish *Herichthys minckleyi*. *Evolutionary Ecology Research*, 7(2):303–324.

Hvala, J. A., Frayer, M. E., and Payseur, B. A. (2018). Signatures of hybridization and speciation in genomic patterns of ancestry. *Evolution*.

Johnson, S. G., Hulsey, C. D., and De León, F. J. G. (2007). Spatial mosaic evolution of snail defensive traits. *BMC Evolutionary Biology*, 7(1):50.

Kopp, M. and Hermisson, J. (2006). The evolution of genetic architecture under frequency-dependent disruptive selection. *Evolution*, 60(8):1537–1550.

Kornfield, I., Smith, D. C., Gagnon, P., and Taylor, J. N. (1982). The cichlid fish of Cuatro Cienegas, Mexico: direct evidence of conspecificity among distinct trophic morphs. *Evolution*, pages 658–664.

Kornfield, I. and Taylor, J. N. (1983). A new species of polymorphic fish, *Cichlasoma minckleyi*, from Cuatro Cienegas, Mexico (teleostei: Cichlidae). *Proc. Biol. Soc. Wash*, 96(2):253–269.

Kornfield, I. L. and Koehn, R. K. (1975). Genetic variation and speciation in new world cichlids. *Evolution*, pages 427–437.

Langmead, B. (2010). Aligning short sequencing reads with bowtie. *Current protocols in bioinformatics*, 32(1):11–7.

Legendre, P., Oksanen, J., and ter Braak, C. J. (2011). Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution*, 2(3):269–277.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Liem, KF, K. L. (1984). Intraspecific macroevolution: Functional biology of the polymorphic cichlid species *Cichlasoma minckleyi*. In Echelle A.A., K. I., editor, *Evolution of Fish Species Flocks*, pages 203–215. Orono, ME: University of Maine Press.

Loh, Y.-H. E., Katz, L. S., Mims, M. C., Kocher, T. D., Soojin, V. Y., and Streelman, J. T. (2008). Comparative analysis reveals signatures of differentiation amid genomic polymorphism in lake malawi cichlids. *Genome biology*, 9(7):R113.

Lucas, L., Fordyce, J., and Nice, C. (2014). Patterns of genitalic morphology around suture zones in north american lycaeides (lepidoptera: Lycaenidae): implications for taxonomy and historical biogeography. *Annals of the Entomological Society of America*, 101(1):172–180.

Lucas, L. K., Nice, C. C., and Gompert, Z. (2018). Genetic constraints on wing pattern variation in lycaeides butterflies: A case study on mapping complex, multifaceted traits in structured populations. *Molecular ecology resources*.

Magalhaes, I. S., Ornelas-Garcıa, C. P., Leal-Cardin, M., Ramírez, T., and Barluenga, M. (2015). Untangling the evolutionary history of a highly polymorphic species: introgressive hybridization and high genetic structure in the desert cichlid fish *Herichtys minckleyi*. *Molecular Ecology*.

Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133):279.

Mandeville, E. G., Parchman, T. L., McDonald, D. B., and Buerkle, C. A. (2015). Highly variable reproductive isolation among pairs of *Catostomus* species. *Molecular ecology*, 24(8):1856–1872.

Marques, D. A., Lucek, K., Meier, J. I., Mwaiko, S., Wagner, C. E., Excoffier, L., and Seehausen, O. (2016). Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS genetics*, 12(2):e1005887.

Mayr, E. (1963). *Animal species and evolution.* Cambridge, Belknap Press of Harvard University Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*.

McPherson, S. (2007). Pitcher plants of the americas. *Blacksburg, Va.: McDonald and Woodward 320p. ISBN*, 939923750.

Muschick, M., Barluenga, M., Salzburger, W., and Meyer, A. (2011). Adaptive phenotypic plasticity in the midas cichlid fish pharyngeal jaw and its relevance in adaptive radiation. *BMC Evolutionary Biology*, 11(1):116.

Nabokov, V. (1975). *Nabokov's butterflies*, chapter From letter to Robert Dirig, pages 713–714. Beacon Press.

Nabokov, V. V. (1949). *The nearctic members of the genus Lycaeides Hübner (Lycaenidae, Lepidoptera).* Museum of Comparative Zoology.

Noor, M. A. and Bennett, S. M. (2009). Islands of speciation or mirages in the desert? examining the role of restricted recombination in maintaining species. *Heredity*, 103(6):439.

Nosil, P. (2008). Speciation with gene flow could be common. *Molecular Ecology*, 17(9):2103–2106.

Nosil, P. (2012). *Ecological speciation*. Oxford University Press.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., et al. (2013). Package 'vegan'. *Community ecology package, version*, 2(9).

Oppenheim, S. J., Gould, F., and Hopper, K. R. (2018). The genetic architecture of ecological adaptation: intraspecific variation in host plant use by the lepidopteran crop pest chloridea virescens. *Heredity*, 120(3):234.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., and Buerkle, C. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular ecology*, 21(12):2991–3005.

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

Peres-Neto, P. R., Legendre, P., Dray, S., and Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, 87(10):2614–2625.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R news*, 6(1):7–11.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014a). ddocent: a radseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2:e431.

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., and Bird, C. E. (2014b). Demystifying the rad fad. *Molecular ecology*, 23(24):5937–5942.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rhymer, J. M. and Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual review of ecology and systematics*, 27(1):83–109.

Rueffler, C., Van Dooren, T. J., Leimar, O., and Abrams, P. A. (2006). Disruptive selection and then what? *Trends in Ecology & Evolution*, 21(5):238–245.

Sage, R. D. and Selander, R. K. (1975). Trophic radiation through polymorphism in cichlid fishes. *Proceedings of the National Academy of Sciences*, 72(11):4669–4673.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323(5915):737–741.

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, 9(7):671–675.

Servedio, M. R. and Noor, M. A. (2003). The role of reinforcement in speciation: theory and data. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):339–364.

Skulason, S. and Smith, T. B. (1995). Resource polymorphisms in vertebrates. *Trends in ecology & evolution*, 10(9):366–370.

Smith, T. B. and Skúlason, S. (1996). Evolutionary significance of resource polymorphisms in fishes, amphibians, and birds. *Annual Review of Ecology and Systematics*, pages 111–133.

Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., and Malmberg, R. L. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus sarracenia using target enrichment. *Molecular Phylogenetics and Evolution*, 85:76–87.

Stephens, M. J. and Hendrickson, A. (2001). Larval development of the cuatro cienegas cichlid, *Cichlasoma minckleyi*. *The Southwestern Naturalist*, pages 16–22.

Streelman, J. T. and Albertson, R. C. (2006). Evolution of novelty in the cichlid dentition. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306(3):216–226.

Trapani, J. (2003). Morphological variability in the Cuatro Cienegas cichlid, *Cichlasoma minckleyi*. *Journal of Fish Biology*, 62(2):276–298.

Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219.

Via, S. (2001). Sympatric speciation in animals: the ugly duckling grows up. *Trends in Ecology & Evolution*, 16(7):381–390.

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587):451–460.

Wen, D., Yu, Y., Hahn, M. W., and Nakhleh, L. (2016). Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular ecology*, 25(11):2361–2372.

West-Eberhard, M. J. (2005). Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences*, 102(suppl 1):6543–6549.

Whitney, K. D., Randell, R. A., and Rieseberg, L. H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower helianthus annuus. *New Phytologist*, 187(1):230–239.

Wimberger, P. H. (1994). Trophic polymorphisms, plasticity, and speciation in vertebrates. *Theory and application in fish feeding ecology. University of South Carolina Press, Columbia*, pages 19–43.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6):851–865.

Yeaman, S. and Whitlock, M. C. (2011). The genetic architecture of adaptation under migration–selection balance. *Evolution*, 65(7):1897–1911.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264.