GENETIC VARIATION AND ADMIXTURE IN THE PALLID-DOTTED BLUE BUTTERFLY (*Euphilotes pallescens*) WITHIN THE GREAT BASIN OF THE WESTERN UNITED STATES

by

Sarah Beth Bialik

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Population and Conservation Biology
August 2017

Committee members:

Chris Nice, Chair

Noland Martin

David Rodriguez

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplicate Permission

ACKNOWLEDGEMENTS

I would like to thank Chris Nice for all of allowing me to work in his lab and pursue this research and supporting me along the way. I would also like to thank my committee for their helpful feedback on my thesis. In addition, thanks go out to my friends and family for their support throughout this process.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation **Description**

SNP- Single-Nucleotide Polymorphism

MAF- Minor Allele Frequency

MID- Muliplex Identifier

GSAF- University of Texas Austin Genomic and Sequencing Analysis Facility

BWA- Burrows-Wheeler Aligner

MCMC- Markov Chain Monte Carlo

DIC- Deviance Information Criterion

ESS- Effective Sample Size

PCA- Principal Components Analysis

NMDS- Non-metric Multidimension Scaling

ABSTRACT

Quantifying genomic variation provides information that can be used to understand the evolutionary history of populations. Here, I examined populations of *Euphilotes pallescens*, a species of butterfly that lives within the Great Basin of western North America. I genotyped 376 butterflies at over 90,000 loci to address questions surrounding gene exchange among lineages of *E. pallescens* and other geographically proximate *Euphilotes* species. I also investigated what relative contributions of historical and contemporary admixture to the patterns I saw. I stratified loci into “common” and “rare” loci based on minor allele frequencies to investigate historical and contemporary genetic structure, respectively. I used a Bayesian hierarchical model to visualize and quantify genetic variation in two analyses: one included only *E. pallescens* populations, while the second analysis was performed at the genus-level. I found evidence of both historical and contemporary gene exchange among subspecies within *E. pallescens* and among *Euphilotes* species. However, there was little evidence of a history of admixture between the Great Basin populations of *E. pallescens* and other *Euphilotes* species. I also found conflict between the patterns of genomic differentiation in these butterflies and their nominal taxonomy. My investigation of the evolutionary history of these butterflies revealed complex relationships and patterns of gene exchange between lineages that suggest the organization of biological diversity is not always strictly hierarchical and the history of divergence is not always strictly bifurcating

I. INTRODUCTION

The quantification of genomic variation can be used to understand the evolution of populations and lineages. Investigations of the mechanisms generating biodiversity begin by delineating boundaries between independent lineages (Gompert *et al.*, 2014; Mandeville *et al.*, 2015; Munshi-South, Zolnik, and Harris, 2015; Underwood, Mandeville, & Walters, 2015; Galaska *et al*., 2016; Parchman *et al.*, 2016). This quantification of genomic variation can also provide the basis for generating hypotheses about the evolution of reproductive isolation within and among lineages (Bigelow, 1965). Recently, some analyses examining the structure of genomic variation among wild populations (Gompert *et al.*, 2014; Whitney *et al*., 2015; Underwood *et al*., 2015) have revealed complex relationships and patterns of gene exchange between lineages that suggest the organization of biological diversity is not always strictly hierarchical. Here, I focus on a polytypic species complex that offers opportunities to quantify historical and contemporary patterns of gene exchange and allows for understanding the history of diversification.

I examined the *Euphilotes pallescens* (Lepidoptera: Lycaenidae) species complex of butterflies, which live in isolated regions of the Great Basin of western North America (Wilson *et al.*, 2013). Commonly known as the pallid dotted-blue butterfly, this polytypic species is currently composed of eight subspecies (Pratt and Emmel, 1998; Austin and Leary, 2008). These include *E. pallescens pallescens* (Tilden and Downey, 1955), *E. p. arenamontana* (Austin, 1998a), *E. p. ricei* (Austin, 1998c), *E. p. calneva* (Emmel and Emmel, 1998), *E. p. confusa* (Pratt and Emmel, 1998), *E. p. emmeli* (Shields, 1975), and *E. p.mattonii* (Shields, 1975), along with *E. p. elvirae* which is the only subspecies not found in the Great Basin (Mattoni, 1966; Austin, 1998a). *Euphilotes pallescens* are typically restricted to low-elevation habitats and use buckwheats in the genus *Eriogonum* as larval hosts (Austin, 1998b, c; Emmel and Emmel, 1998; Pratt and Emmel, 1998; Brock and Kaufman, 2003; Austin and Leary, 2008). Host plant declines due to habitat loss have led to butterfly population declines and motivated conservation efforts (Nevada Natural Heritage Program, 2010; Nevada Fish and Wildlife Service, 2010).

Table 1: Species designations for *Euphilotes* sampling localities sample sizes from each (n).

Locality Number	Nominal Species	Nominal Subspecies	Locality Name	n
1	*pallescens*	*arenamontana*	Sand Mountain	26
2	*pallescens*	*calneva*	Sand Pass	20
3	*pallescens*	*calneva*	Turtle Mountain	27
4	*pallescens*	*ricei*	Silver State Dunes	30
5	*pallescens*	*confusa*	Hot Springs Mountain	28
6	*pallescens*	*confusa*	Mono Lake	20
7	*pallescens*	*confusa*	Marietta	23
8	*pallescens*	*confusa*	Esmeralda	28
9	*pallescens*	*pallescens*	Railroad Valley	28
10	*pallescens*	*pallescens*	Sunnyside	30
11	*pallescens*	*emmeli*	Panaca	29
12	*ancilla*	-	Bull Creek	14
13	*ancillas*	-	Steens Mountain	9
14	*ancilla*	-	Swift Creek	9
15	*battoides*	-	Drake Peak	11
16	*enoptes*	-	Shadow Mountain	15
17	*enoptes*	-	Cave Lake	7
18	*glaucon*	-	Mount Ashland	14
19	*glaucon*	-	Soda Mountain Road	8

Wilson *et al.* (2013), used mtDNA, nuclear loci, and morphological data to examine variation in six subspecies of these butterflies (Table 1).

Morphological and molecular data were only partially concordant and discordance among molecular markers suggested a complex history with periods of potential gene exchange among lineages within *E. pallescens* and between *E. pallescens* and other *Euphilotes*. I used next-generation sequencing techniques to produce high resolution single-nucleotide polymorphism marker (SNP) data and examined genome-wide variation and genetic structure in the same six subspecies. I also examined variation in congeneric species that might also have complex histories with *E. pallescens*. In my analyses, I categorized variable sites according to minor allele frequencies (MAFs) and examined patterns observed for SNPs with low MAFs separate from those with higher MAFs. This approach allowed me to distinguish historical admixture from more contemporary gene exchange affecting the patterns of ancestry across the genome in these butterflies.

I quantified patterns of genetic variation across the genome of *E. pallescens* and four congeners using SNP data to investigate the history of evolutionary diversification in this complex group of butterflies. Specifically, I asked: 1. Has gene exchange occurred among lineages of *E. pallescens* or between *E. pallescens* and other geographically proximate *Euphilotes* species? and 2. If there is evidence of admixture, what are the relative contributions of historical and contemporary gene exchange in the history of these butterflies? The answers to these questions provide a foundation for understanding the patterns of gene flow and evolution in this notoriously complex group of butterflies and will inform management strategies for future conservation efforts.

II. MATERIALS AND METHODS

DNA sequencing and data collection

I sampled *Euphilotes* (n=376) from locations across Nevada, Oregon, Wyoming, and California (Table1 & Figure 1; Kahle & Wickham, 2013);

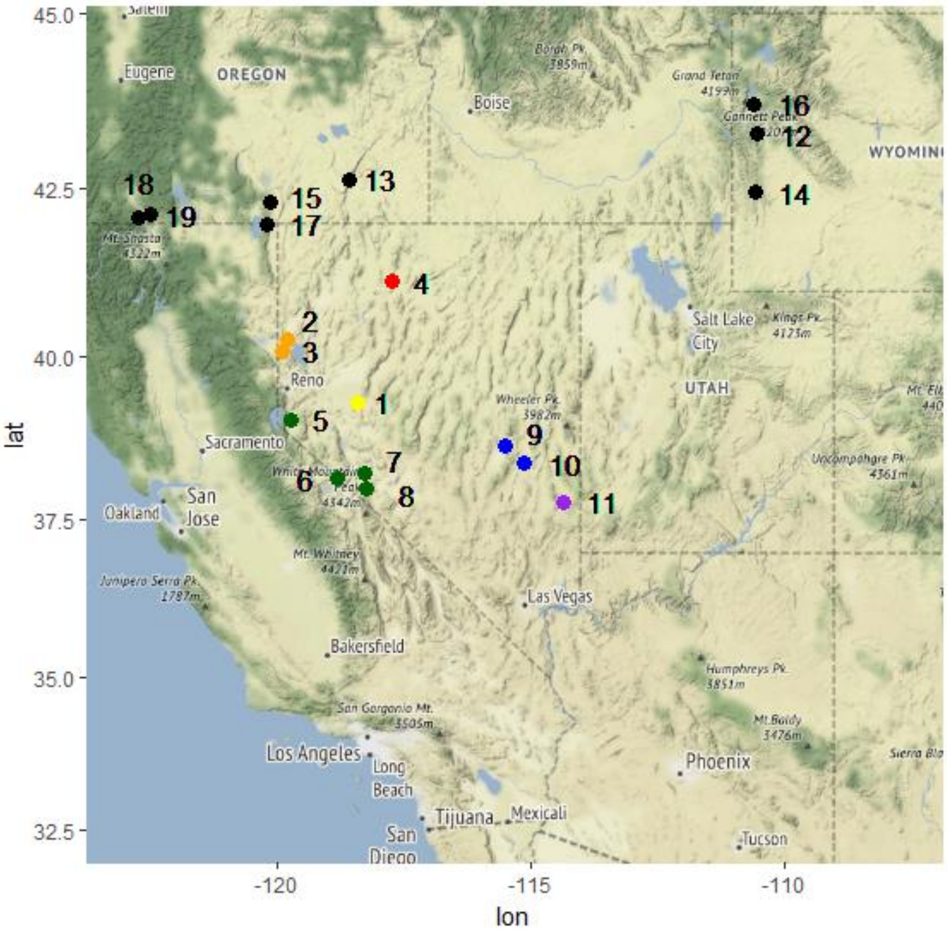


Figure 1: Sampling sites of *Euphilotes* in western North America. Numbers correspond to localities in Table 1. For *E. pallescens*, eleven sites were sampled with at least 20 individuals from each location. These sites are color-coded according to the information presented in Wilson et al. (2013) and represent different subspecies. Red represents *E. p. ricei*, orange is *E. p. calneva*, yellow is *E. p..arenamontana*, green is *E. p. confusa*, blue is *E. p. pallescens*, and purple is *E. p. emmeli*. The black dots show collection sites for the other *Euphilotes* species.

289 were *E. pallescens* butterflies collected in Nevada from within the Great Basin region. The remaining samples were individuals from other *Euphilotes* species (*E. ancilla*, *E. battoides*, *E. enoptes*, *E. glaucon*) and were collected to provide a basis for comparison of divergence within *E. pallescens* and to investigate potential patterns of

gene exchange among species (Wilson *et al.*, 2013). Each specimen was placed into a glassine envelope and stored in a -80°C freezer until DNA extraction took place. The DNA was isolated from butterfly thoracic tissue using the DNeasy Blood and Tissue Kit (Qiagen Inc., Alameda, CA, USA; Wilson *et al.*, 2013).

I used a genotyping-by-sequencing approach for generating genomic-level population genetics data. I prepared reduced representation genomic libraries for each individual following the protocol of Parchman *et al*. (2012) and Gompert *et al.* (2012). Briefly, genomic DNA was digested with two restriction enzymes: EcoR1 and Mse1. Illumina sequencing adaptors and 8-10bp individual multiplex identifier (MID) sequences were ligated to fragments and then amplified. The PCR product was pooled and size-selection of fragments between 300bp and 450bp was performed with BluePippin technology (Sage Science Inc., Beverly, MA, USA). These fragments from all individuals were then sequenced on a single lane using the Illumina HiSeq 2500 platform at the University of Texas Austin Genomic Sequencing and Analysis Facility (GSAF; Austin, Texas, U.S.A.).

EcoR1 restriction sites and MIDs were removed from sequence reads and replaced with individual identifiers using a custom PERL script. Because I lack a reference genome for *Euphilotes* butterflies, a random subset of reads were used for an initial *de novo* assembly. SeqMan NGEN ver. 12.2.0 build 86 software (DNASTAR, Inc.) was used for *de novo* assembly of a subset of the reads (25 million; Mandeville *et al.*, 2015). The majority-rule consensus sequences from each contig were pruned to eliminate excessively long or short sequences and then aligned to each other to filter out potentially paralogous sequences. Each of the remaining consensus sequences were then used as scaffolds for the reference-based assembly, where all sequence reads were included. This alignment of sequences was performed using Burrows-Wheeler Aligner version 0.7.13 (BWA; Li and Durbin, 2009) allowing up to 4-bp mismatches. SAMTOOLS version 0.1.18 was used to sort and compress individual alignments (Li and Durbin, 2009).

Variable genetic sites, or SNPs, were identified using SAMTOOLS and BCFTOOLS version 0.1.18 (Li and Durbin, 2009; Mandeville *et al.*, 2015). For each particular SNP, at least 50% of individuals were required to have data at that SNP site to be considered in these analyses**.** Individuals that had low coverage (<0.5x mean coverage across loci) were removed from all analyses. Only bi-allelic SNP's were retained for my analyses to reduce the potential of including paralogous loci. To ensure independence among loci, one SNP per contig was chosen randomly to minimize the potential for linkage disequilibrium. Global allele frequencies estimated from genotype likelihoods were used to categorize loci into "common" and "rare" minor allele frequency classes.

Categorizing alleles in this manner allows for a novel way of investigating geographic population boundaries that have evolved over time. Loci whose MAFs are greater than 5% are designated as common, whereas rare loci are those with MAFs less than <5%. Though 5% is an arbitrary threshold, coalescent theory predicts that most rare alleles (minor alleles at rare loci) are a result of recent mutations in the population being examined and have had little time to spread across lineages (spatial isolation) or can consequently be representative of recent gene flow between populations (Gravel *et al*., 2011; Gompert *et al*., 2014). Rare alleles that are observed in more than one population should therefore demonstrate contemporary (or relatively recent) gene exchange between populations; common (older) alleles (MAF greater than 5%) have had time to persist in, and spread among, populations (Gravel *et al*., 2011; Gompert *et al*., 2014). Common alleles are more likely to have persisted for longer than rare alleles and can be informative about historical gene flow. Examining patterns of variation at loci by classifying them as rare (MAF <5%) and common (MAF >5%) facilitates the examination of historical and contemporary gene exchange, represented by common and rare alleles, respectively.

Population Genetic Analyses

To estimate population genetic parameters, I used the program ENTROPY to quantify genetic variation using Bayesian clustering. ENTROPY is similar to STRUCTURE in that it uses a clustering algorithm to determine admixture proportions without *a priori* information (Pritchard, Stephens, & Donnelly, 2000). Unlike STRUCTURE, ENTROPY uses genotype likelihoods as input which allows for the calculation of posterior probability estimates for individual genotypes and credible intervals, taking into account sequencing or alignment errors that may be present and allowing for the incorporation of uncertainty (Gompert *et al*., 2013). Using these genotype estimates, allele frequencies were calculated. Admixture proportions (*q*) were calculated to indicate the proportion of each individual's genome derived from each of *k* source ancestral populations. This hierarchical Bayesian clustering requires input from the user for the number of clusters (*k*) to be determined with the data (Gompert *et al.*, 2014). For the first analyses, which included only my focal species *E. pallescens* (n=289), ENTROPY was run for *k*=2 to *k*=10 allowing clustering beyond the number of nominal subspecies (*k*=6; see Wilson *et al.*, 2013) for both the common and rare sets of loci. After examination of model parameters, it was shown that higher *k*'s (*k*=8+) were poorly fit to the common loci data and were not focused upon in this study. To provide a basis of comparison for differentiation within *E. pallescens* and to examine the possibility of gene exchange with other *Euphilotes* species, I performed clustering analyses that included select *E. pallescens* populations (Locilities 1, 5, 9, and 11; n=112; see Table 1) and four other nominal *Euphilotes* species— *E. ancilla* (n=32), *E. battoides* (n=11), *E. enoptes* (n=22), and *E. glaucon* (n=22)—for a total of 199 individuals. Because Principal Component analyses and clustering analyses like STRUCTURE and ENTROPY can be less accurate with unbalanced sample sizes (see McVean, 2009; Onogi, Nurimoto, & Morita, 2011), I restricted the samples of *E. pallescens* to four sampling localities representing the major patterns of genetic structure from the first analysis (see above and Results). This allowed sample sizes of *E. pallescens* and the other *Euphilotes* species to be more even. To compare the results of ENTROPY for different numbers of *k*, ENTROPY was run for *k*=2 to *k*=10 for the *E. pallescens*-only analyses of the common and rare sets of loci and *k*=2 to *k*=7 for the genus-level analyses of both common and rare sets of loci.

For all analyses, a total of 10 chains were run with 100 000 MCMC steps and 5000 as burn-in and thinning every 10th step. DIC scores were obtained and averaged across all chains to compare models. Generally, models with lower DIC values are those that fit the data best. Gelman-Rubin diagnostics were run to ensure chains were mixing properly and probable convergence to the posterior distribution (Gelman & Rubin, 1992). Average effective sample size (ESS) was calculated for admixture proportions for all individuals to verify a sufficient number of steps for MCMC chains had been retained for analyses using *coda* in R (Plummer *et al*., 2006; R Core Team, 2015). Mean assignments for individuals were calculated across all chains. Mean genotype posterior probabilities were calculated from one chain from all models (i.e. across *k*'s) to display the average results across chains. Using these probabilities, a principal components analysis (PCA) was performed to visualize the patterns of genetic differentiation among populations (see Gompert *et al*., 2014). Allele frequencies were calculated from the mean genotype posterior probabilities and used to estimate pairwise G_{st} (Nei, 1987) for both common analyses, quantifying genetic differentiation among sampling localities (Wright, 1950; Gaggiotti and Foll, 2010). Population differentiation as measured by pairwise G_{st} values was illustrated using non-metric multidimensional scaling (NMDS) using the R packages *vegan* and MASS (Venables & Ripley, 2002; Oksanen *et al*., 2017) for 1000 permutations.

III. RESULTS

Sequencing for all individuals (original n=404) resulted in 201 145 679 short sequence reads. A total of n=376 individuals were used in the *E. pallescens*-only analyses and n=199 for genus-level analyses. For both analyses, I also consider common and rare loci separately with the express purpose of comparing patterns of historical and contemporary admixture, respectively. Inspection of ESS's confirmed that a sufficient number of steps had been retained per individual. Examination of MCMC chains showed sufficient chain mixing and convergence of posterior distributions.

For the *E. pallescens*-only analysis, I identified fewer (34, 940) common SNP loci than (63, 339) rare SNP loci, as might be expected for geographically isolated populations. The major structure within *E. pallescens* is a north-south division. This is clearly evident in the PCA plot (Figures S1, S2, Supporting

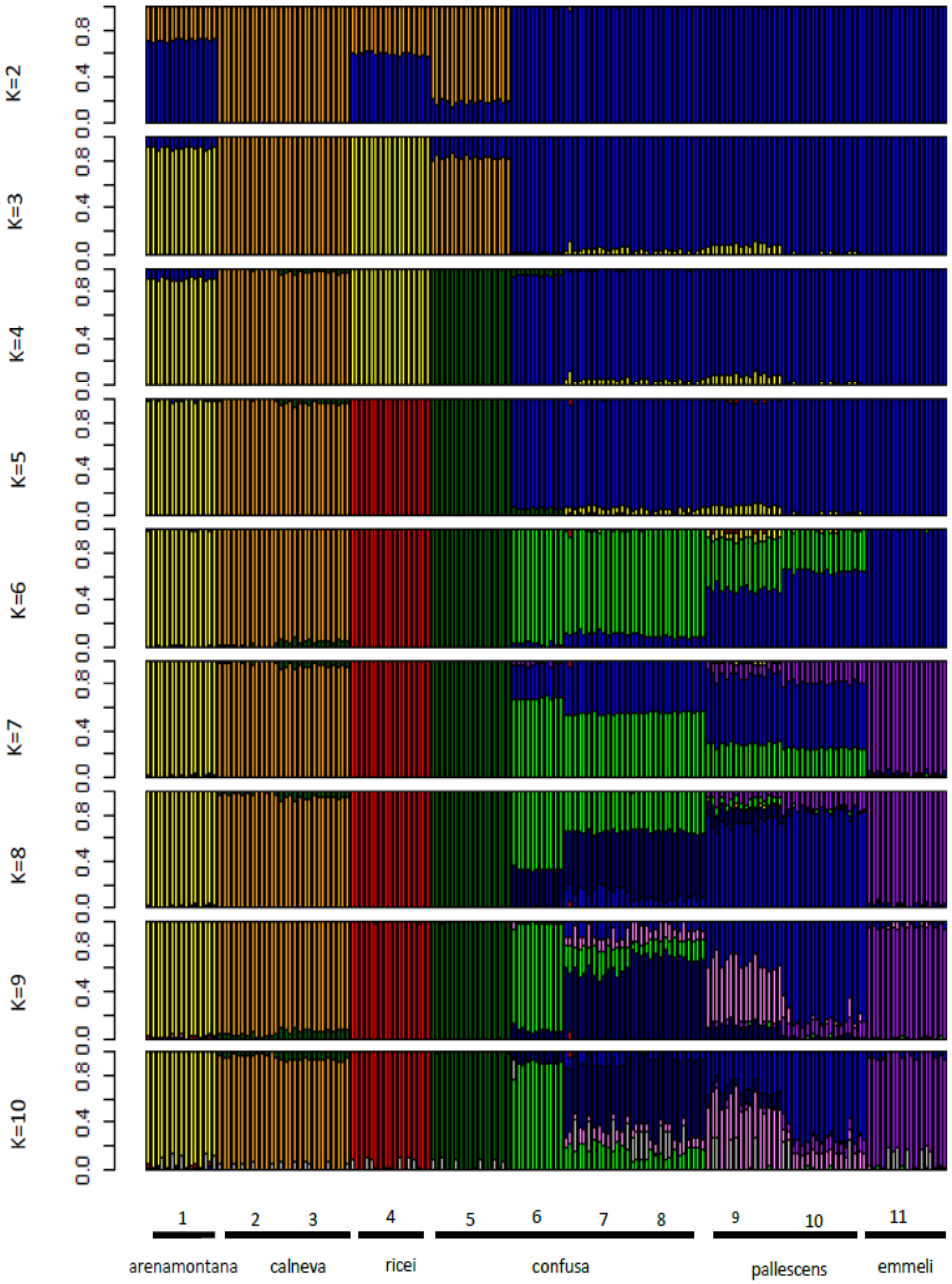


Figure 2: Admixture proportions (*q*) based on common loci for *E. pallescens*-only analysis. The x-axis shows each individual (n=289) as a separate vertical line while the y-axis is assignment probability. Each color represents a different cluster (*k*). The numbers below the plot correspond to the locality numbers in Table 1, and the bars below identify the nominal subspecies designations for localities. Individuals from locality 5 are differentiated from other *E. p. confusa* localities and form their own cluster. Localities 6-11 cluster together and show some admixture at higher *k*'s, although it nominally represents three subspecies.

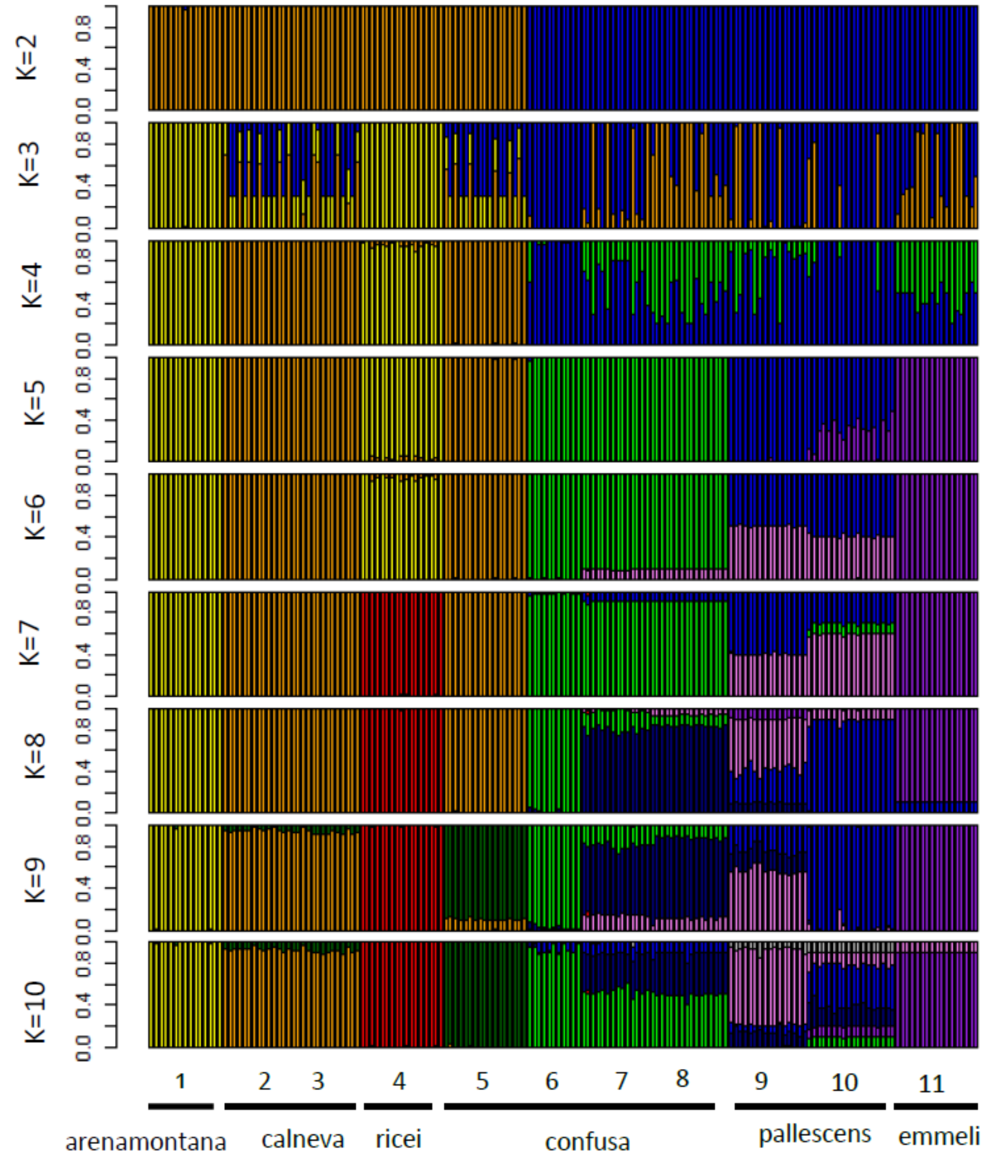


Figure 3: Admixture proportions (*q*) based on rare loci for *E. pallescens*-only analysis. The x-axis shows each individual (n=289) as a separate vertical line while the y-axis is assignment probability. Each color represents a different cluster. The numbers below the plot correspond to the locality numbers in Table 1 and the bars below identify the nominal subspecies designations. *E. p. areanamontana* (locality 1), *E. p. ricei* (locality 4), and *E. p. emmeli* (locality 11) all cluster independently. *Euphilotes p. pallescens* (localities 9 and 10) form a single cluster but demonstrate minor admixture with the *E. p. confusa* (localities 6-8) populations. The only *E. p. confusa* populations that does not group according to this designation is locality 5, which groups with *E. p. calneva* populations at lower *k*'s.

Information) of individuals based on their posterior genotype probabilities and also in the clustering estimated by ENTROPY (Figures 2,3). The northern localities (localities 1-5, Table 1, Figure 1) are clearly differentiated from the southern localities, and the northern localities cluster according to their nominal subspecific designations. The southern localities are less clearly differentiated and show more evidence of admixture in both the common and rare loci. Further, the nominal subspecies *E. p. confusa* (localities 5-8) is divided with the northernmost locality of this subspecies (locality 5) which forms its own cluster (Figures 2, 3). Pairwise G_{st} values show the same patterns of differentiation, particularly the north-south division (Table S2) and is further supported in my NMDS analysis (Figure S3).

To provide a basis of comparison for differentiation within *E. pallescens* and to investigate any potential admixture between species, the second analysis included select *E. pallescens* populations (localities 1, 5, 9, and 11, see Table 1) and congeners and will be called my genus-level analyses. Here, I identified 31, 855 common and 40, 184 SNP loci. The major structure in both of these data sets separates *E. pallescens* from its congeners in both the common and rare loci (Figures 4, 5). The clustering pattern within *E. pallescens* is similar to the *E. pallescens*-only analyses, but with more evidence of admixture. The nominal taxonomy of the four other *Euphilotes* species conflicts with the patterns of differentiation observed in both the common and rare loci. Both PCA (Figure 6) of posterior genotype

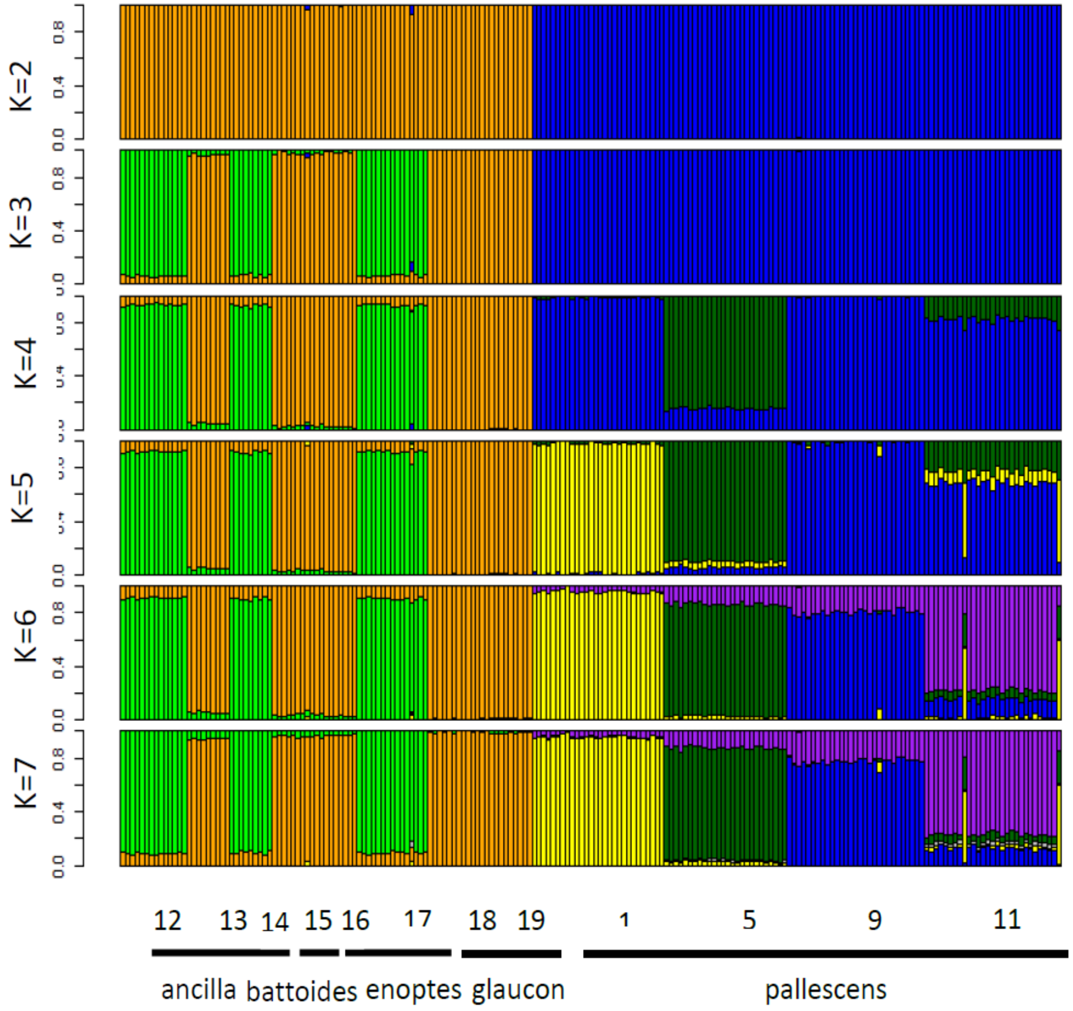


Figure 4: Admixture proportions (*q*) based on common loci for genus-level analysis. The x-axis shows each individual (n=199) as a separate vertical line while the y-axis is assignment probability. Each color represents a different cluster. The numbers below the plot correspond to the locality numbers in Table 1, and the bars below identify the nominal subspecies designations. Two nominal species of *Euphilotes* grouped together in one cluster (*E. ancilla* and *E. enoptes*). The other nominal species clustered in another separate group (*E. ancilla*, *E. battoides*, *E. enoptes*, and *E. glaucon*). All *E. pallescens* populations clustered independently at *k*=7.

probabilities and the clustering from ENTROPY show two distinct groups (Figures 4, 5). One of the groups consists of part of the *E. ancilla* populations (locality 13), plus *E. battoides* (locality 15), part of *E. enoptes* (locality 16)

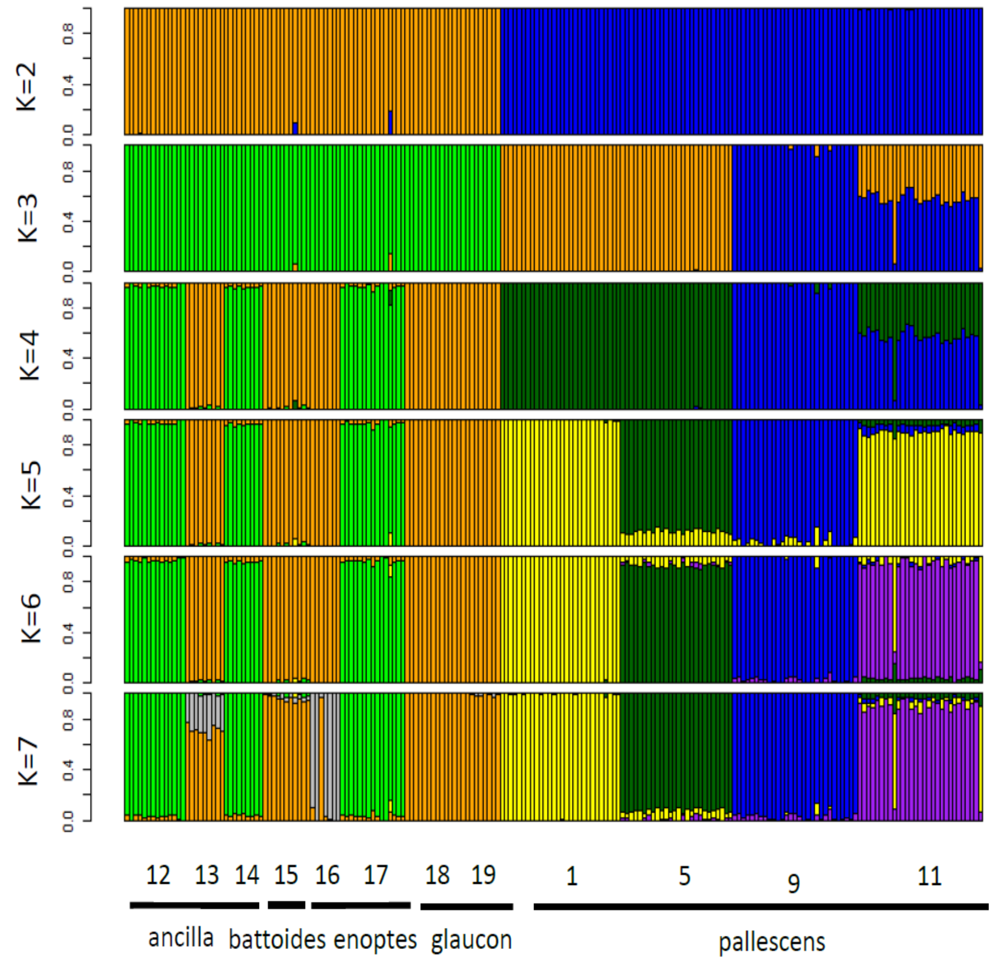
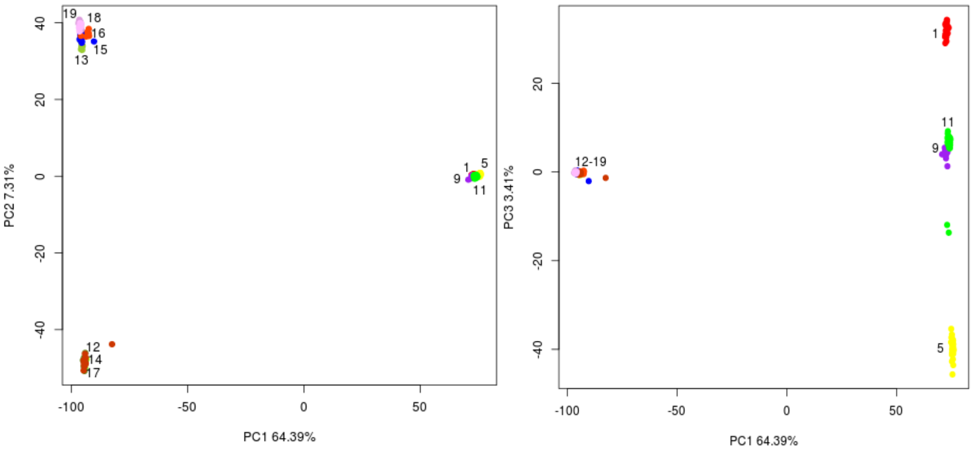


Figure 5: Admixture proportions (q) for the genus-level based on the analysis of rare loci. The x-axis shows each individual (n=199) as a separate vertical line while the y-axis is assignment probability. Each color represents a different cluster. The numbers below the plot correspond to the locality numbers in Table 1, and the bars below identify the nominal subspecies designations. The patterns of genetic differentiation among and between populations remains consistent with the common loci analysis. At *k*=7, an *E. enoptes* locality is more distinct than the other *Euphilotes* localities.

and *E. glaucon* (localities 18 and 19). The other group is comprised of parts of *E. ancilla* and *E. enoptes* (localities 12 and 14, 17). However, these two groups also show evidence of admixture in both the common and rare loci, indicating both historical and contemporary gene exchange. There is very little evidence of admixture between *E. pallescens* and these other species. These patterns of genetic structure are also clearly evident in analysis of pairwise G_{st} values which are further supported by the NMDS analysis (Table S3, Figure S5).

Figure 6: PCA plot for the genus-level analysis of common loci. PC1 splits the *E. pallescens* lineages from the congener lineages (while explaining 63.61% of the variation). PC2 is equal to 7.52% of the variation and divides the two clusters of nominal *Euphilotes* species.



IV. DISCUSSION

I found that the evolutionary history of the *Euphilotes* complex is complicated and did not follow a strictly bifurcating pattern. Instead, I saw differences in the patterns observed for common and rare loci both within *E. pallescens* and among other *Euphilotes* species. In this study, I asked if gene exchange had occurred among lineages of *E. pallescens* or congeners as predicted by Wilson et al. (2013) and, if there was evidence of admixture, were the patterns different when comparing historical and contemporary gene flow? I found that patterns of admixture were present within *E. pallescens* and among congeners. Other patterns show unexpected divergence of some *Euphilotes* populations along with gene exchange in some *E. pallescens* populations. These gene flow events were more often observed when examining the rare (contemporary) loci, demonstrating more recent gene exchange between populations.

In the *E. pallescens*-only analysis, I found that using SNPs classified as common (34, 940 loci), *k*=5 was most likely the best fit for the data. At *k*=5 (Figure 2), I see a clear division between the southern populations and the northern populations. Almost all of the northern populations (localities 1, 2 and 3, 4, and 5) split into their own distinct clusters that represent their respective nominal subspecies designations (Figure 2). The southern populations cluster together as one, although nominally they belong to three separate subspecies. Wilson et al. (2013) found that six groups (subspecies, see Figure 1) best fit their genetic data. For mydata, I found the overall groupings were similar, but differed slightly. As in Wilson et al. (2013), *E. p. ricei* (locality 1) and *E. p. arenamontana* (locality 4) remained distinct, along with *E. p. calneva* (localities 2 and 3) clustering together. *Euphilotes p. confusa*, specifically locality 5, was different from other

E. p. confusa populations and clustered alone in my analysis; in Wilson et al. (2013), this population clustered with localities 6-8. The rare loci showed minor amounts of contemporary admixture among *E. p. confusa* (localities 6-8) and *E. p. pallescens* (localities 9 and 10) while maintaining that these were separate clusters, which is more representative of the subspecies designations of current taxonomy (Figure 3). This difference between patterns observed for common and rare loci demonstrates that these populations have been isolated until relatively recently.

In my second, genus-level analyses, I found that the patterns displayed in *k*=3 for common (31, 855; Figure 4) and rare (40, 184; Figure 5) loci were very different and can represent recent divergence. The contrast in this comparison was very clear—divergence patterns varied significantly depending on which set of loci are examined. This is a powerful way to identify the distinctiveness of populations and to demonstrate divergence among populations. Depending on the loci being examined, common (historical) or rare (contemporary), different patterns may appear. In the common loci, I saw at *k*=3, the third cluster that formed was a group of the congener species (Figures 4, 6). *Euphilotes ancilla* and *E.enoptes* clustered in one group (localities 12, 14 and 17) while the rest of the nominal *Euphilotes* species grouped in another cluster. In comparison, the rare loci showed an *E. pallescens* population as more distinct (locality 9) than any of the sampled *Euphilotes* species while all of the congeners clustered together (Figures 5, S4). There was also some contemporary admixture displayed among the *E. pallescens* populations, indicating potential gene flow.

Although rare variants can be recent mutations unique to their populations due to having limited time to spread among populations, they also can be indicative of older variants that have reached an equilibrium and have remained in the population (Slatkin 1985). Regardless of their age, rare alleles should reveal fine-scale genetic structure in spatially restricted populations (Barton & Slatkin, 1986). Bi-allelic loci with minor alleles classified as rare provide a solid foundation for determining divergence and evolution of populations. Patterns of ancestry and levels of gene flow can also be investigated with these techniques, allowing me to better understand the history of lineages not following a strictly bifurcating pattern of divergence (McDonald et al., 2008; Foote & Morin, 2016; Novikova et al., 2016).

Understanding the history of populations and understanding their genetic distinctiveness can help conservation priorities. In this data set in particular, the Sand Mountain blue butterfly (*E. p. arenamontana*) had been previously petitioned for protection by several conservation groups (Nevada Fish & Wildlife Service, 2010). This petition failed and the population was determined to be stable and did not require federal protection. The analyses presented here demonstrate the distinctiveness of this population (locality 4; see Figures 2, 3, S1) and the importance for protection due to their genetic isolation from other populations. The Nevada Natural Heritage Program (NNHP) also considered three other subspecies (two of which were included in this analysis: *E. p. ricei* (represented by the locality 4) for conservation implications previously. *Euphilotes. p. ricei* (locality 4) formed a separate cluster in both analysis of common and rare loci and was genetically distinct from other populations indicating isolation and the need for protection (Figures 2, 3, S1).

In conclusion, this study offers new insights into the patterns of evolution and the history of divergence in the species *Euphilotes pallescens*. The differences seen between the analysis of common and rare loci are informative on the process of divergence and speciation in geographically isolated populations. Some evidence of admixture between populations indicated that these butterfly lineages did not follow a strictly bifurcating pattern. Interesting patterns were displayed across the barplots and PCAs that demonstrated that nominal species names do not always follow the patterns of genetic differentiation in populations. This study provides a foundation for understanding the patterns of gene flow and evolution in this notoriously complex group of butterflies and will inform management strategies for future conservation efforts. Future work could focus upon congener relationships independently of *E. pallescens* or could include *E. rita* to examine the relationship of divergence of *E. pallescens* from other *Euphilotes*.

APPENDIX SECTION

Table S1: Deviance Information Criterion (DIC) estimates for ENTROPY models run for each analysis. Lower estimates of DIC generally reflect better model fit to the data

	***E. pallescens*-only, common**	***E. pallescens*-only, rare**	**Genus-level, common**	**Genus-level, rare**
K2	42097038	35350839	22109049	16807399
K3	41365066.32	35398126.2	21323327.05	16701600.12
K4	40707523.84	34859586.34	20485706.95	16440098.37
K5	40091610.44	34543971.73	19711188.09	16334231.35
K6	39257121.12	34312772.54	19232662.79	16235698.48
K7	38696416.53	34142224.86	19155434.48	16056615.89
K8	38299775.66	33974743.8	--	--
K9	37980757.67	33861389.42	--	--
K10	37277469.39	33877036.94	--	--

Table S2: Pairwise G_{st} values with lower (bottom triangle) and upper (top triangle) credible intervals for *E. pallescens*-only, common analysis at *k*=5 and all localities. Epe represents Esmeralda, epf represents Marietta, eph represents Hot Springs Mountain, epi represents Sand Mountain, epk represents Silver State Dunes, epm represents Mono Lake, epp represents Panaca, epq represents Sunnyside, epr represents Railroad Valley, eps represents Sand Pass, and ept represents Turtle Mountain.

	epe	epf	eph	epi	epk	epm	epp	epq	epr	eps	ept
epe	-	0.01	0.13	0.11	0.12	0.02	0.02	0.01	0.01	0.13	0.14
epf	0.01	-	0.14	0.11	0.12	0.02	0.02	0.01	0.01	0.14	0.14
eph	0.13	0.13	-	0.19	0.19	0.13	0.16	0.14	0.13	0.14	0.14
epi	0.10	0.10	0.18	-	0.15	0.12	0.13	0.12	0.10	0.16	0.17
epk	0.12	0.12	0.19	0.15	-	0.13	0.14	0.13	0.11	0.16	0.16
epm	0.02	0.02	0.12	0.12	0.12	-	0.03	0.02	0.02	0.13	0.14
epp	0.02	0.02	0.15	0.13	0.14	0.03	-	0.02	0.02	0.16	0.16
epq	0.01	0.01	0.13	0.11	0.12	0.02	0.02	-	0.01	0.14	0.15
epr	0.01	0.01	0.13	0.09	0.11	0.02	0.02	0.01	-	0.12	0.13
eps	0.13	0.14	0.13	0.16	0.16	0.13	0.15	0.14	0.12	-	0.01
ept	0.13	0.14	0.13	0.16	0.16	0.13	0.16	0.14	0.12	0.01	-

Table S3: Pairwise G_{st} values with lower (bottom triangle) and upper (top triangle) credible intervals for genus-level, common analysis at *k*=7 and all localities. Eab, ean and, eau represent *E. ancilla* populations; ebd represents the *E. battoides* population; eea and eej represent *E. enoptes* populations; egg and egl represent *E. glaucon* populations; eph, epi, epp, and epr represent *E. pallescens* populations.

	eab	ean	eau	ebd	eea	eej	egg	egl	eph	epi	epp	epr
eab	-	0.19	0.01	0.19	0.21	0.01	0.21	0.21	0.46	0.45	0.45	0.45
ean	0.18	-	0.19	0.02	0.03	0.19	0.02	0.02	0.44	0.44	0.44	0.44
eau	0.01	0.18	-	0.19	0.21	0.01	0.21	0.21	0.46	0.45	0.45	0.45
ebd	0.19	0.02	0.19	-	0.03	0.19	0.01	0.01	0.44	0.44	0.43	0.44
eea	0.20	0.02	0.20	0.03	-	0.21	0.03	0.04	0.45	0.44	0.44	0.44
eej	0.01	0.18	0.01	0.18	0.20	-	0.20	0.21	0.45	0.45	0.44	0.44
egg	0.20	0.02	0.20	0.01	0.03	0.20	-	0.01	0.45	0.45	0.44	0.44
egl	0.20	0.02	0.20	0.01	0.03	0.20	0.01	-	0.45	0.45	0.45	0.45
eph	0.45	0.43	0.45	0.43	0.44	0.44	0.44	0.44	-	0.04	0.04	0.03

Table S3.

Continued.

	eab	ean	eau	ebd	eea	eej	egg	egl	eph	epi	epp	epr
epi	0.44	0.43	0.44	0.43	0.43	0.44	0.44	0.44	0.04	-	0.04	0.03
epp	0.44	0.43	0.44	0.43	0.43	0.43	0.43	0.44	0.04	0.04	-	0.02
epr	0.44	0.43	0.44	0.43	0.43	0.43	0.43	0.44	0.03	0.03	0.02	-

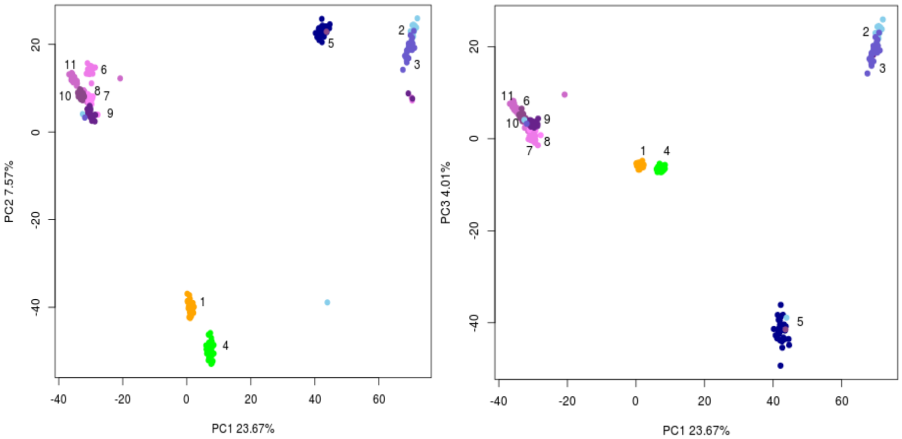


Figure S1: PCA plot for the analysis of common loci in the *E. pallescens*-only analysis.

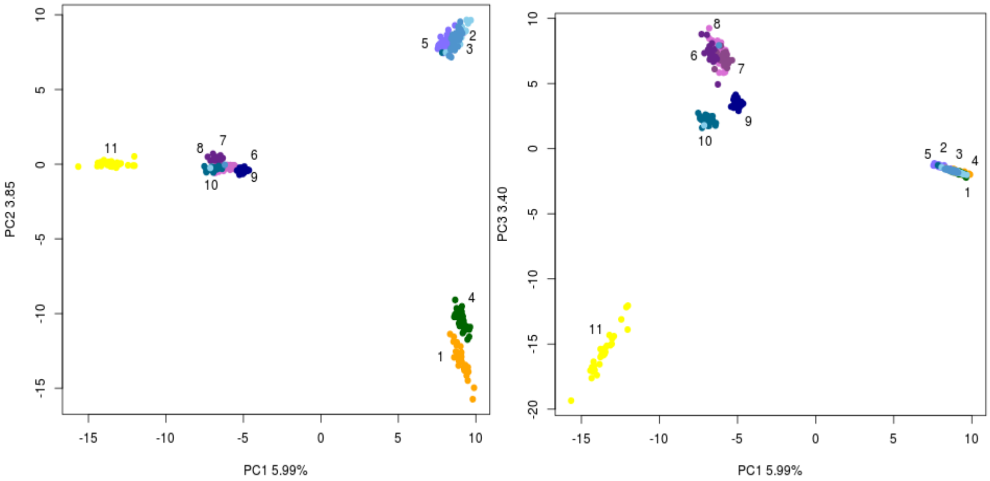


Figure S2: PCA plot for the analysis of rare loci in the *E. pallescens*-only analysis.

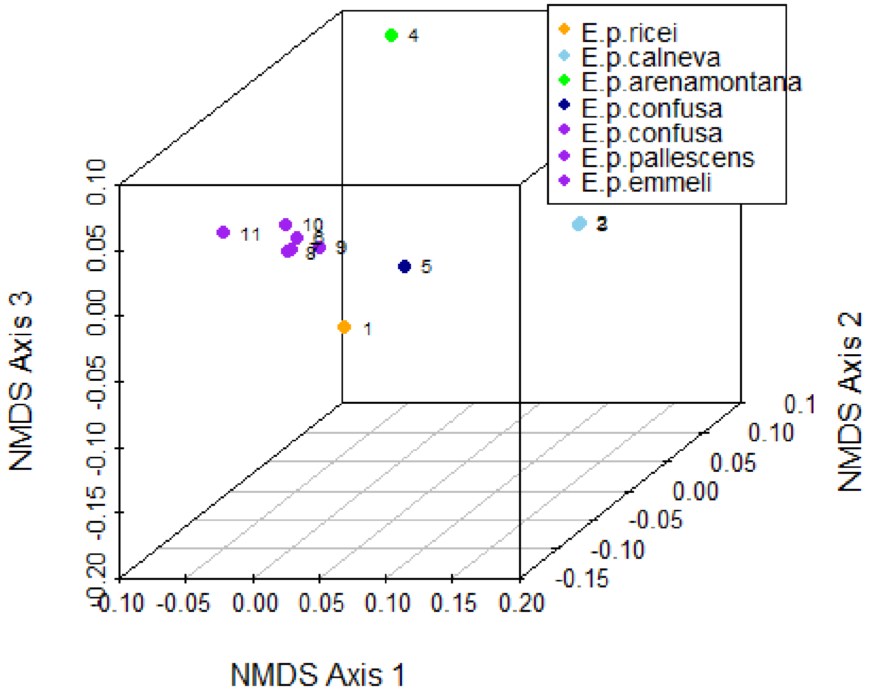


Figure S3: NMDS for the analysis of common loci in the *E. pallescens*-only analysis.

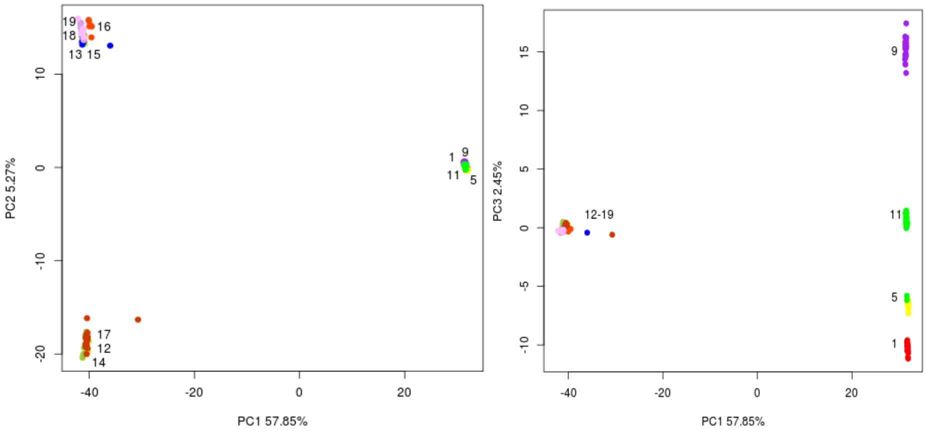


Figure S4: PCA plots for rare loci in the genus-level analysis.

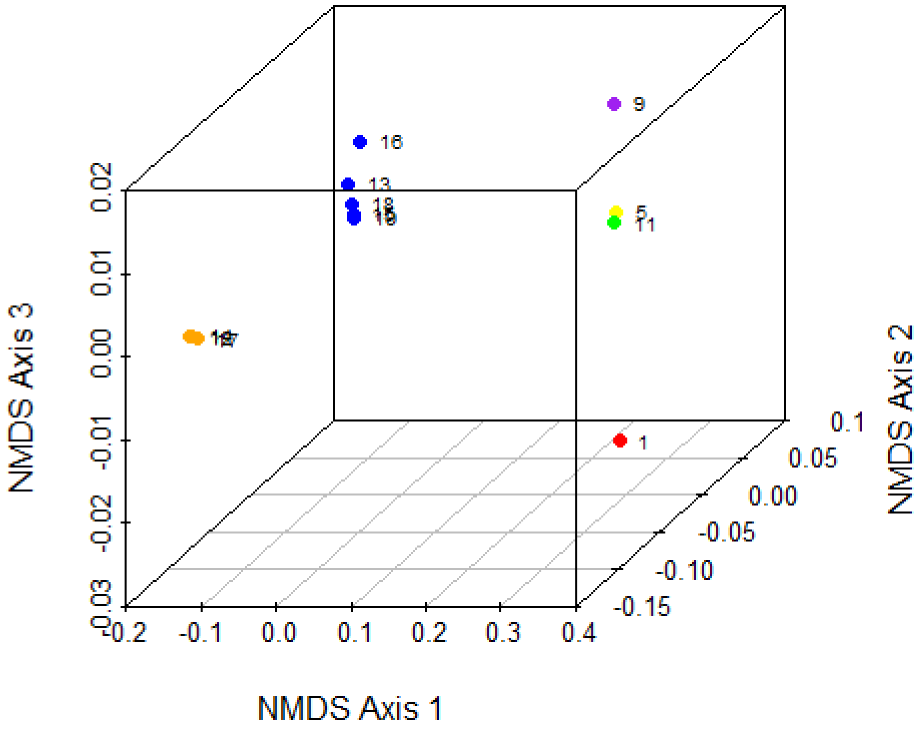


Figure S5: NMDS of the common loci for genus-level analysis.

REFERENCES

Arkle, R. and Pilliod, D. (2015). Persistence at distributional edges: Columbia spotted frog habitat in the arid Great Basin, USA. *Ecology and Evolution*, 5(17): 3704-3724.

Austin, G. and Leary, P. (2008). Larvel hostplants of butterflies in Nevada. *Holarctic Lepidoptera*. 12: 74-77.

Austin, G. T. (1998a). Checklist of Nevada butterflies. *Systematics of western North American butterflies* (ed. By T.C. Emmel), pp 837-844. Mariposa Press, Gainesville, FL.

Austin, G. T. (1998b) New subspecies of Lycaenidae (Lepidoptera) from Nevada and Arizona. *Systematics of western North American butterflies* (ed. By T.C. Emmel), pp. 539-572. Mariposa Press, Gainesville, FL.

Austin, G. T. (1998c). A new subspecies of *Euphilotes pallescens* (Lepidoptera: Lycaenidae) from the northern Great Basin of Nevada. *Systematics of western North American butterflies* (ed. By T.C. Emmel), pp. 815-818. Mariposa Press, Gainesville, FL.

Barton, N. & Slatkin, M. A Quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity, 56*: 409-415.

Benson, L., Currey, D., Dorn, R., Lajoie, K., Oviatt, C., Robinson, S., Smith, G., and Stine, S. (1990). Chronology of expansion and contraction of four Great Basin lake systems during the past 35,000 years. *Palaeogeography, Palaeoclimatology, Palaeoecology,* 78: 241-286.

Brock, J. and Kaufman, K. (2003). Field guide to butterflies of North America. Houghton Mifflin Company. New York, New York.

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

Ehrlich, P., Murphy, D., Sherwood, C., White, R., and Brown, I. (1980). Extinction, reduction, stability and increase: The responses of Checkerspot Butterfly (*Euphydryas*) populations to the California drought. *Oecologia*, 46: 101-105.

Emmel, J. F. and Emmel, T. C. (1998). A new *Euphilotes pallescens* subspecies (Lepidoptera: Lycaenidae) from the northern California-Nevada border region. *Systematics of western North American butterflies* (ed. By T.C. Emmel), pp. 277-282. Mariposa Press, Gainesville, FL.

Foote, A. & Morin, P. (2016). Genome-wide SNP data suggest complex ancestry of sympatric North Pacific killer whale ecotypes. *Heredity*, *117*(5): 316-325.

Forister, M., McCall, A., Sanders, N., Fordyce, J., Thome, J., O'Brien, J., Waetjen, D., and Gaggiotti, O. and Foll, M. (2010). Quantifying population structure using the *F*-model. *Molecular Ecology Resources*, 10: 821-830.

Galaska, M., Sands, C., Santos, S., Mahon, A., & Halanych, K. (2016). Geographic structure in the Southern Ocean circumpolar brittle star *Ophionotus victoriae* (Ophiuridae) revealed from mtDNA and single-nucleotide polymorphism data. *Ecology and Evolution*, *7*: 475-485.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sciences*, *7*(4): 457-511.

Gompert, Z., Lucas, L., Buerkle, A., Forister, M., Fordyce, J., and Nice, 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.

Hafner, J., Upham, N., Reddington, E., and Torres, C. (2008). Phylogeography of the pallid kangaroo mouse, *Microdipodops pallidus*: a sand-obligate endemic of the Great Basin, western North America. *Journal of Biogeography*, 35: 2102-2118.

Jahner, J., Gibson, D., Weitzman, C., Blomberg, E., Sedinger, J., & Parchman, T. (2016). Fine-scale genetic structure among greater sage-grouse leks in central Nevada. *BMC Evolutionary Biology*, *16*(127).

Kahle, D. & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1), 144-161.

Larson, W., Seeb, L., Everett, M., Waples, R., Templin, W., & Seeb, J. (2013). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinoook salmon (*Oncorrhynchus tshawytscha*). *Evolutionary Applications*, *7*: 355-369.

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

Mandeville, E., Parchman, T., McDonald, D., and Buerkle, A. (2015). Highly variable reproductive isolation among pairs of *Catostomus* species. *Molecular Ecology*. 24: 1856-1872.

Mayr, E. (1966). *Animal Species and Evolution*. The Belknap Press of Harvard University Press, Cambridge, MA.

McDonald, D., Parchman, T., Bower, M., Hubert, W., & Rahel, F. An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. *Proc Natl aced Sci U.S.A.*, *105*: 10837-10842.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10).

Munshi-South, J., Zolnik, C., & Harris, S. (2015). Population genomics of the Anthropocene: urbanization is negatively associated with genome-wide variation in white-footed mouse populations. *Evolutionary Applications*, *9*: 546-554.

Nei, M. (1987). Molecular evolutionary genetics. New York, NY. Columbia University Press.

Nevada Natural Heritage Program. (2015). *Nevada Natural Heritage Program at-risk plant and animal tracking list*. Available at: http://heritage.nv.gov/ sites/default/files/ library/track.pdf (last accessed 28 November 2015)

Novikova, P., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Shmid, K., Fedorenko, O., Holm, S., Säll, T., Schlötterer, C., Marhold, K., Widmer, A., sese, J., Shimizu, K., Weigel, D., Krämer, U. Koch, M., & Nordoborg, M. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, *48*(9).

Oksanen, J., Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Stevens, M., Szoecs, E., & Wagner, H. (2017). Vegan: Community ecology package. R package version 2.4-2. http://CRAN.R-project/package=vegan.

Onogi, A., Nurimoto, M., & Morita, M. (2011). Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, *12*(263).

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*: 7-11.

Pratt, G. F. and Emmel, J. F. (1998). A new subspecies of *Euphilotes pallescens* (Lycaenidae) from the Death Valley region of California. *Systematics of western North American butterflies* (ed. By T. C. Emmel), pp. 271-276. Mariposa Press, Gainesville, FL.

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

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/

Rahbek, C. (1995). The elevational gradient of species richness: a uniform pattern? *Ecography*, 18(2): 200-205.

Scott, J. (1992). The butterflies of North America: A natural history and field guide. Stanford, CA. Stanford University Press.

Shapiro, A. (2010). Compounded effects of climate change and habitat alteration shift patterns of butterfly diversity. *PNAS*, 107(5): 2088-2092.

Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution, 39*(1): 53-65.

Sparks, T. and Yates, T. (1997). The effect of spring temperature on the appearance dates of British butterflies 1883-1993. *Ecography*, 20(4): 368-374.

Tilden J. W. and Downey J.C. (1955). A new species of *Philotes* from Utah (Lepidoptera: Lycaenidae). *Bulletin of the Southern California Academy of Sciences*. pp. 25-29. The Academy. Los Angeles, CA.

Trimble, S. (1989). The Sagebrush Ocean: A natural history of the Great Basin. University of Nevada Press. Reno, NV.

U.S. Fish and Wildlife Service. (2007). Petition to list the Sand Mountain Blue Butterfly (*Euphilotes pallescens arenamontana*) as a threatened or endangered species under the U.S. Endangered Species Act. Available at: http://www.fws.gov/ nevada//nv_species/documents/smbb/smbb_petition.pdf (last accessed 28 November 2015).

Underwood, Z., Mandeville, E., & Walters, A. (2015). Population connectivity and genetic structure of burbot (*Lota lota*) populations in the Wind River Basin, Wyoming. *Hydrobiologia*, *759*.

Van Devender, T., Martin, P., Thompson, R., Cole, K., Jull, T., Long, A., Toolin, L., and Donahue, D. (1985). Fossil packrat middens and the tandem accelerator mass spectrometer. *Nature,* 317: 610-613.

Venables, W. & Ripley, B. (2002). Modern applied statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Whitney, K., Broman, K., Kane, N., Hovick, S., Randell, R., & Rieseberg, L. (2015). QTL mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. *Molecular Ecology*, *24*(9): 2194-3311.

Wilson, J. and Pitts, J. (2010). Phylogeographic analysis of the nocturnal velvet ant genus *Dilophotopsis* (Hymenoptera: Mutillidae) provides insights into diversification in the Nearctic deserts. *Biological Journal of the Linnean Society*. 101: 360-375.

Wilson, J.S., Sneck, M., Murphy, D.D., Nice, C.C., Fordyce, J.A., & Forister, M.,L. (2013). Complex evolutionary history of the pallid dotted-blue butterfly (Lycaenidae: *Euphilotes pallescens*) in the Great Basin of western North America. *Journal of Biogeography*. 40: 2059-2070.