Population Genomics of Pecos Pupfish (*Cyprinodon pecosensis*)

Final report to New Mexico Department of Game and Fish
July 17, 2023

Prepared by: Andrew R. Whiteley – Director of the University of Montana Conservation Genomics Lab

POPULATION GENOMICS OF PECOS PUPFISH (*Cyprinodon pecosensis*)

Introduction

The New Mexico Department of Game and Fish contracted with the University of Montana Conservation Genomics Lab to examine the population genetic structure (e.g., genetic differentiation among populations, genetic variation within populations, and patterns of inbreeding) of Pecos pupfish from 13 populations. We obtained genotypes following bioinformatic analysis of Restriction site-Associated (RAD)-sequencing (Seq) (generated with the *Pst*-1 6-base cutting restriction enzyme) data from Amish [1]. The following report provides population genomics analyses of these data.

Results

As reported by Amish [1], there was no evidence of admixture between sheepshead minnow (*Cyprinodon variegatus*) and Pecos pupfish. We excluded sheepshead minnow samples from subsequent analyses, except for the analysis of inbreeding (see below). We conducted several additional locus filtering steps not performed by Amish [1](see detailed methods below). Populations from which samples were analyzed are shown in Table 1.

Genetic differentiation among populations

We used principal components analysis (PCA) on the 117,319 locus data set to initially describe patterns of genetic structure among the 13 Pecos pupfish populations (Figures 1 and 2). BTLS02 was the most genetically divergent site, as indicated by its strong divergence along PC1 (Figure 1). BLM01 was the next most divergent site, as indicated by its strong divergence along PC2. BTLS01 and BTLS03 were somewhat divergent from the remainder of the sites. The remaining 9 sites were genetically similar along the first two principal components, as indicated by the large degree of overlap of individuals in PC space (Figure 1).

Examination of PC3 and PC4 revealed some of the more subtle genetic divergence that occurs among the examined sites (Figure 2). Genetic similarity between BTLS01 and BTLS03 was apparent because both occurred to the left of zero on PC axis 3 (Figure 2), but the distance between these two sites also revealed a degree of genetic differentiation between them. BLN09 and BLN20 were genetically similar. BLM01, BLN07, and BTLS02 clustered closely together (Figure 2), indicating very little allele frequency divergence among these three sites. BLN01, BLN02, BLN03, BLN04, BLN05, and BLN15 overlapped in PC axis 3 and 4 space (Figure 2), again indicating a lack of allele frequency divergency among these six sites. On the other hand, divergence apparent in either Figure 1 or Figure 2 reveals allele frequency divergence consistent with lack of gene flow.

Discriminant Analysis of Principal Components (DAPC) performed on the 117,319 locus data set provided a similar overall depiction of the patterns of genetic differentiation, with some slight differences. DAPC provides a test for groupings that minimizes variance within and maximizes variance among clusters for a given value of K (i.e., number of pre-defined genetic groups). The Bayesian Information Criterion (BIC) indicated that K = 5 or 6 were plausible (Figure 3). K = 5 had the lowest BIC (a model selection criterion where lower values are preferred) and BIC started to increase slightly at K = 6 (Figure 3). K = 5 is therefore the overall most plausible scenario, but we describe patterns present for K = 5 and 6 because their BIC values were very similar. BIC declined slightly again at K = 7 and K = 8, but these values are likely to be associated with over-splitting and over-interpretation.

At K = 5, cluster 3 (BLM01) and cluster 4 (BLN01, BLN02, BLN03, BLN05, and BLN15) were the most divergent along DA axis 1 (x-axis; Figure 4). Cluster 5 (BTLS02) was divergent from cluster 4 along both DA axis 1 and axis 2. Cluster 1 (BLN04) and cluster 2 (BTLS01, BTLS03, BLN07, BLN09, BLN20) were similar along DA axis 1, but divergent along DA axis 2. Cluster 1 (BTN04) and cluster 2 (BTLS01, BTLS03, BLN07, BLN09, and BLN20) were differentiated along DA axis 2 but not DA axis 1 (Figure 4).

At K = 6, cluster 1 (BLM01), cluster 5 (BTLS02), and cluster 6 (BTLS01 and BTLS03) occurred in the lower left quadrant. Cluster 2 (BLN04) and cluster 4 (BLN07, BLN09, and BLN20) occurred in the lower right quadrant. Cluster 3 (BLN01, BLN02, BLN03, BLN05, and BLN15) occurred in the upper left quadrant. The difference between K = 5 and K = 6 occurred with the splitting of BTLS01 and BTLS03 (cluster 6) from BLN07, BLN09, and BLN20 (cluster 4) revealing divergence among these sets of populations that was not detected at K = 5 (i.e., these populations appeared together in cluster 2 at K = 5). These results for K = 6, but not K = 5, reveal the presence of genetic differentiation between these BTLS and BLN sites that could be biologically meaningful but smaller than the differentiation observed among other populations in this data set.

To calculate population genomic summary statistics within populations, we randomly subsampled 10,000 SNPs without replacement. Examination of F_{ST} as a measure of genetic differentiation revealed patterns concordant with the PCA and DAPC results. Overall genetic differentiation was moderate and driven primarily by elevated genetic differentiation of two sites. Site BTLS02 was the most genetically divergent site (range of pairwise F_{ST}: 0.238 – 0.338; higher F_{ST} represents higher divergence Table 2). Site BLM01 was also genetically differentiated from other sites (range of pairwise F_{ST}: 0.126 – 0.338; Table 2). Sites that clustered closely together in PCA or DAPC space tended to also have relatively low pairwise F_{ST} values (Table 2).

Testing for locus-specific natural selection

We performed an initial analysis to test for the putative influence of natural selection on genetic divergence. There was strong genetic drift in several populations, which causes challenges when attempting to separate genetic drift and selection. We used the 117,319 locus data set for an OutFLANK analysis. The mean F_{ST} of neutral loci, as inferred with OutFLANK, was 0.126. Three-hundred and forty-eight loci (348/117,319 = 0.30%) were statistically significant outliers and were therefore consistent with divergent selection. Outlier loci occurred throughout the genome, with little evidence of many outliers occurring in nearby genome locations (Figure 6). Specific hypotheses should be formed if further testing for the effects of natural selection (and therefore possible local adaptation) is desired (e.g., divergence in populations among specific habitat types or ecotypes, if present).

Genetic variation within populations

We used the randomly subsampled 10,000 SNP data set to examine summary statistics of genetic variation within populations. Ten of the population samples had similar genetic variation, for example H_S was between 0.072 and 0.077 (Table 3; Figure 7), the proportion of polymorphic loci (*P*) was greater than 0.32, and allelic richness (*AR*) was greater than 1.17 (higher values reflect higher genetic variation for all three variables; Table 3). Three populations had lower genetic variation. BTLS02 had the lowest estimates of genetic variation (H_S = 0.0423, 95% CI 0.0418 – 0.0429; *AR* = 1.09; *P* = 0.134). BLM01 and BLN09 had slightly reduced (compared to the ten populations with more genetic variation) and similar (to each other) estimates of genetic variation (BLM01: H_S = 0.0664, 95% CI 0.0656 – 0.0670; *AR* = 1.15; *P* = 0.304; BLN09: H_S = 0.0686, 95% CI 0.0680 – 0.0692, *AR* = 1.16; *P* = 0.312; Table 3; Figure 7).

To ensure that the 10,000 locus subsamples accurately represented amounts of genetic variation within populations, we performed 5 replicate subsamples of 10,000 loci (Table 4). Estimates of H_S were similar, as indicated by the small mean standard deviation among the five replicates of 0.0011. Estimates of *P* were also similar, with slightly greater variation among replicates. The mean standard deviation among the five replicates for estimates of *P* was 0.0046 (Table 4). Further, a subsample of 50,000 loci yielded nearly identical estimates of H_S and *P* (data not shown). We conclude that each 10,000-locus replicate provided similar information, any one replicate could be used to perform population genomic analysis, and relative comparisons among populations would not change with analysis of more loci.

Inbreeding within populations

We used the less stringently filtered data set of 299,660 SNPs to examine inbreeding within populations to allow comparison to a separate analysis of desert pupfishes [2]. We limited inference regarding inbreeding, which is based on Runs of Homozygosity (ROH), to longer runs of consecutive homozygous loci within an individual. Longer runs correspond to inbreeding that has occurred within approximately the last 64 generations. More ancient inbreeding is associated with short runs of homozygosity that are more difficult to reliably detect. The distribution of F_{ROH} values with a cutoff of inbreeding within 64 generations in the past was strongly bimodal (Figure 8). All F_{ROH} (with a 64-generation cutoff) values greater than 0.10 belonged to individuals from BTLS02 (Table 5; Figure 8). BLM01 also had a slightly higher mean value of F_{ROH} (Table 5). To further visualize patterns of individual inbreeding, we randomly selected and plotted 10 individuals from the most inbred site (BTLS02), the second most inbred site (BLM01), the third most genetically depauperate site (but with a close to average across all 14 mean values for F_{ROH}; BLN09), a site with typical genetic variation and individual inbreeding (BLN01), and the sheepshead minnow (SHM; Figure 9). BTLS02 individuals clearly had greater evidence of inbreeding within the last 32 generations. There was a slight pattern for elevated, very recent (2 generations in the past) inbreeding in BLM01 when compared to the sites other than BTLS02 (Figure 9).

Contemporary N_e

Estimates of contemporary N_e in the two most genetically depauperate sites (BTLS02 and BLM01) were relatively large (i.e., >100; Table 5). Specifically, estimated contemporary N_e in BTLS02 was 321.8 (95% CI 55.3-inf) and in BLM01 was 665.3 (95% CI 80.6-inf). This suggests that current population sizes in these two sites are large enough to prevent pronounced effects of genetic drift and inbreeding. Precision was low and these estimates would likely change with larger sample sizes, but we are confident that new estimates would still be large (e.g., likely to be at least in the 100s).

Estimates of contemporary N_e were surprisingly small (<100) and had finite confidence intervals in three sites: 1) BLN04 had an N_e of 10.5 (95% CI 2.6 – 67.6); 2) BTLS01 had an N_e of 25.8 (95% CI 10.8 – 186.6); and 3) BLN20 had an N_e of 45.2 (95% CI 19.0 – 1758.5). Numbers for BLN04 reflect the highest confidence that contemporary N_e is small and of concern. There is a chance that larger sample sizes would yield somewhat larger estimates of N_e for BTLS01 and BLN20. We suspect, however, that more robust estimates would still be small for these sites.

The remaining sites had varying point estimates of contemporary N_e but infinite upper confidence limits. In sites where the point estimate was in the thousands and the upper confidence limit was infinity (e.g., BLN07; Table 5), we can be confident that contemporary N_e is in fact large. Sites where the point estimate of contemporary N_e was relatively small (e.g., < 50) but the upper confidence limit was infinity (e.g., BLN02 and BLN05; Table 5) have the greatest uncertainty. Larger samples might yield relatively small estimates of N_e, but it is more likely that estimates would be substantially larger.

Discussion

No admixture with SHM was detected for this set of 13 Pecos pupfish population samples [1]. It remains possible that admixture would be detected with analysis of SHM collected from the Pecos River. However, we consider this outcome highly unlikely. The SHM that were considered in this project were all highly divergent from Pecos pupfish [1]. SHM themselves were quite divergent from one another and could represent individuals from multiple populations in their native range [1]. It is beneficial to our analyses that wide genetic variation was captured in SHM and yet none of that variation was similar to that found in Pecos pupfish. Thus, we suspect that our admixture results are robust and would not change if future analyses include SHM from the Pecos River (i.e., within the native range of the Pecos pupfish).

We used two different approaches to examine spatial patterns of genetic structure. PCA is well-suited for examining raw patterns in allele frequency data among known sampling locations. DAPC adds a discriminant function step, asking whether there is support for clustering individuals in a way that maximizes variation among clusters but minimizes variation within clusters. Clusters are identified without inclusion of prior population information [3]. The clusters that emerge are thus driven solely by similarities in allele frequencies. Comparison of both approaches, along with analysis of $F_{ST,}$ can offer useful insights. From both PCA and DAPC, we can confidently infer that BLM01 and BTLS02 are highly genetically differentiated sites. These sites also had the highest pairwise F_{ST} values. DAPC revealed divergence of four more groups (at K = 6); (1) BLN04, (2) BTLS01 and BTLS03, (3) BLN01, BLN02, BLN03, BLN05, and

BLN15, and (4) BLN07, BLN09, and BLN20. The sites from these latter four groups clustered together along PC axes 1 and 2. PC axes 3 and 4 revealed that BLN04, BLN07, BLN09, and BLN20 were highly similar. In this instance, the discrepancy between DAPC and PCA lies with BLN04; we would conclude that it is more differentiated based on the DAPC than on the PCA results. DAPC and PCA were concordant in revealing that BLN01, BLN02, BLN03, BLN05, and BLN15 are all highly similar and, as a group, are somewhat divergent from other sites. PCA revealed greater divergence between BTLS01 and BTLS03 than did DAPC, but these two sites were still relatively close together in PC space. In combination, the PCA and DAPC analyses suggest that Pecos pupfish from the sites examined represent the following five, distinct genetic groups: (1) BTLS02, (2) BLM01, (3) BTLS01 and BTLS03, (4) BLN04, BLN07, BLN09, and BLN20 and (5) BLN01, BLN02, BLN03, BLN05, and BLN15. These results suggest that some of these populations are isolated (groups 1 and 2) and that gene flow is most likely within groups 3, 4, and 5 (mostly likely to commonly occur within group 5), but less likely (might occur, but more rarely) among them. Within these 5 groups, we observed some additional evidence that BLN04 was divergent within group 4 and that BTLS01 and BTLS03 exhibited some divergence. Therefore, there appears to be variation in rates of gene flow within these genetic groups.

The fifth group described above (containing BLN01, BLN02, BLN03, BLN05, and BLN15) appears to display geographical coherence among most sites in the group. Specifically, all of these sites are geographically close. The fourth group above (containing BLN04, BLN07, BLN09, and BLN20) also has some geographical coherence in that all sites except BLN20 are relatively close. These results allow predictions to be formed regarding possible subterranean gene flow among these sinkholes. However, there is also a risk of over-interpreting these results. Analyses with this many SNPs can reveal patterns that are statistically significant but not biologically meaningful. That said, pairwise F_{ST} values tend to be substantially below 0.05 within each of these two groups (groups 4 and 5) and tend to be greater than 0.05 for comparisons between the two groups. For example, an F_{ST} of 0.007 (BLN01 *vs.* BLN02) compared to 0.059 (BLN01 *vs.* BLN20) could be biologically meaningful, with the expectation that migrants are exchanged more commonly between the former two populations than between the latter two.

Genetic variation provides an indication of past demography and future adaptive potential [4]. Three sites have lower genetic variation (BTLS02, BLM01, and BLN09). Our estimates of genetic variation (H_O, H_S, AR, and *P*) were derived from a random representation of loci from throughout the genome. BTLS02 in particular appears to have either gone through a genetic bottleneck, either a short duration but severe bottleneck or it could have occurred at chronically low effective population size. This site, and to a lesser extent BLM01 and BLN09, are likely to have lower potential to adapt to future, changing environmental conditions based on our estimates of lower standing genetic variation in these populations. The remaining ten populations had similar status in terms of measures of genetic variation.

The most evidence of recent inbreeding (within the last 64 generations) occurred in BTLS02. This site had the highest mean inbreeding coefficient and there was strong evidence for a bout of elevated inbreeding that occurred approximately 32 generations ago. BLM01 had elevated, very recent inbreeding (within 2 generations), but lacked the signal of inbreeding for earlier generations that was observed for BTLS02. There was no evidence of recent inbreeding for other populations. The models we used indicated a high level of inbreeding a very long time ago (approximately 500 generations ago) for all of the populations. However, this inference is based on the portion of the data with the lowest signal-to-noise ratio because it is based on very short runs of homozygosity. We suggest that additional examinations of different ROH model types and sensitivities would be required to have greater confidence in inferences regarding inbreeding taking place more than 64 generations ago. This work could use publicly available data from other pupfish species, which yielded similar estimates of F_{ROH}-based inbreeding coefficients to those observed here for more recent generations [2].

Estimates of contemporary N_e were generally informative, despite low precision for many estimates due to a small sample size. If sample size is small relative to the true (unknown) N_e, point estimates will lack accuracy and confidence intervals will be wide [5]. However, large values (approximately greater than 500) tell us that N_e is large enough for genetic drift to have minimal effect on allele frequencies. For the 6 populations with N_e larger than 500 in this study, estimates obtained with larger samples would be more accurate and precise. However, we can be confident that estimates of N_e re-estimated with larger sample sizes would still be large [6]. In general, to increase accuracy and precision, larger sample sizes would be required, likely on the order of at least 100 individuals per site based on the data presented here. Fewer loci than the complete set generated here, on the order of 1000's of SNPs, would be sufficient to obtain accurate estimates. However, it is our opinion that the current estimates can still inform specific management actions (see specific recommendations immediately below).

Recommendations for specific populations:

BTLS02: This site had the lowest observed genetic variation, the most evidence for inbreeding, and was highly genetically differentiated from other sites. All of these observations are likely due to this population having historically undergone a population bottleneck (i.e., the population went through a period of time at a small enough size to experience pronounced genetic drift and inbreeding). However, contemporary N_e was relatively large. This suggests that current abundance is relatively high (probably in the thousands), many adults currently contribute to population-level reproduction , there is no high reproductive skew, and there have not been dramatic population size fluctuations in recent history. Thus, we might infer that strong genetic drift and inbreeding in the past might not have had large negative fitness consequences. For example, largely deleterious alleles may have been purged (removed via natural selection) from the population. Inevitably, some deleterious alleles must have gone to fixation to create the higher fixed genetic load observed in this population. However, this fixed genetic load appears not to have large fitness consequences under current environmental (abiotic and biotic) conditions. This population might have limited adaptive potential in the future due to its low genetic variation, and inbreeding depression could become apparent if the environment becomes more stressful. If correct, this combination of past inbreeding but relatively large contemporary N_e suggests genetic rescue is not currently needed. Management focus should instead be on maintaining large abundance and N_e. Future monitoring could be used to determine if conditions change and evidence accrues that fitness has been impaired (e.g., if abundance, genetic variation, or N_e decline), whereupon this population might become a good candidate for genetic rescue. This set of inferences rests on the assumption that the contemporary N_e is accurate and reflects conditions over the last several generations. Genetic examination of a larger sample or consideration of demographic data would lend this conclusion more certainty.

BLM01: This site had the second lowest genetic variation and the second highest mean individual inbreeding coefficient. This site was also highly genetically differentiated from other populations and also had a relatively large contemporary estimate of N_e. Genetic concerns for this population are similar to those outlined for BTLS02. However, it also appears that BLM01 did not go through as severe a population bottleneck as BTLS02. The modest signal of very recent inbreeding could warrant further examination of contributing factors such as reduction in habitat quantity or quality in the last several generations. Thus, based on population genetic theory, we would expect fitness issues associated with genetic load of deleterious mutations to be less severe in BLM01 than BTLS02. It should be noted that standing deleterious variation and the effects of genetic drift are highly stochastic, so it remains possible that fitness consequences of inbreeding in these two populations could be reversed.

BLN09: This site had the third lowest genetic variation but was otherwise very genetically similar to other populations, including having a large point estimate of contemporary N_e (> 1000). These results do not warrant treating this population differently from those at the other sites (discussed below). The low estimate of genetic variation might be a sampling artifact, possibly due to sampling related individuals. However, given the similarity of BLN09 to other populations based on the results of the PCA and DAPC analyses, this explanation is only plausible if the same sample effects did not influence analyses of genetic differentiation.

BLN04 (and to a lesser extent BTLS01 and BLN20): These three sites had relatively high genetic variation and low genetic differentiation, suggesting a lack of strong population bottlenecks historically. However, for BLN04 in particular, the point estimate of contemporary N_e was very small and confidence intervals were highly constrained. This suggests that some combination of a small current population size, reproductive contribution by only a few adults, highly skewed reproductive among a larger set of adults, or population fluctuations might be a cause for concern regarding this population's persistence. These data warrant investigation into whether demographic or even habitat factors, such as those that might limit the quantity or quality of spawning habitat, might be negatively affecting population persistence at these sites.

Remaining sites: The remaining sites had similar amounts of genetic variation, were moderately genetically differentiated from one another, and had relatively high (or had among the most uncertain) estimates of contemporary N_e. Several clusters were detected, which appear to routinely exchange gene flow within each cluster and might represent different, functioning metapopulations (e.g., BLN01, BLN02, BLN03, BLN05, and BLN15 vs. BLN04, BLN07, BLN09, and BLN20). Depending on the ecological setting, management efforts could focus on ensuring that gene flow is maintained among these populations. The patterns detected here could guide any such efforts. It could be valuable to re-estimate contemporary N_e for sites with relatively small point estimates but very large confidence intervals, especially if other deterministic stressors are present.

Methods

Genotype filtering

As reported by Amish [1], there was no evidence of admixture between sheepshead minnow and Pecos pupfish. We excluded sheepshead minnow samples from subsequent analyses.

A total of 405 Pecos pupfish from 13 collection sites passed the filtering criteria from Steve Amish's report [1] on the same data set. VCF files created by Steve Amish were used as a starting point for subsequent analyses presented here. Initial analyses performed here revealed a subset of Single Nucleotide Polymorphisms (SNPs) that appeared to be paralogs (i.e., multiple loci with highly similar sequences - possibly duplicate genes - that mapped to the same genome location and were considered to be the same locus). These loci exhibited a pattern where all, or nearly all, individuals were heterozygous. We excluded this set of loci by testing all loci using program HDPlot [7] and removing loci with $|D| > 25$ (D is a specialized metric that measures the deviation of the number of sequencing reads from an expectation and is highly informative for detecting paralogous loci [2]); and observed heterozygosity > 0.60. 948,049 loci passed the above filtering criteria and also had a minimum of 50% locus missingness (i.e., at least 50% of individuals were successfully genotyped at a given locus). 409,307 loci passed the filtering criteria and had 80% locus missingness (i.e., at least 80% of individuals were successfully genotyped at a given locus). Of the 948,049 SNPs at 50% missingness, 299,660 were polymorphic within the 13 Pecos pupfish populations. Of the 409,307 SNPs at 80% missingness, 117,319 were polymorphic within the 13 Pecos pupfish populations. We used two missingness criteria because we wanted to exclude loci with more missing data (i.e., use the 80% missingness criteria) for most analyses. However, we used the more relaxed locus missingness criterion (50%) for analysis of Runs Of Homozygosity (ROH; see below), so that we could replicate the filtering procedures used by Tian et al. [2] for other desert pupfishes. Additionally, the analysis of ROH should be robust to missing data but also benefits from using the largest possible number of loci.

Genetic differentiation among populations

We initially examined patterns of genetic differentiation using Principal Components Analysis (PCA) with the 117,319 locus data set. We used the function *dudi.pca* from R package *adegenet* [8]. Genotypes were scaled with the function *gen.scale* and missing genotypes , once allele frequencies were calculated, were replaced with mean allele frequency values. PCA results for Pecos pupfish were not sensitive to data scaling and centering as revealed by highly similar groupings across PC axes when PCA was performed with and without scaling and centering (data not shown). Further, we performed the same analysis with and without the set of putative paralogs and did not observe changes to the patterns revealed by PCA. Results are highly similar to those shown by Amish [1], who did not remove paralogs.

We used DAPC to further examine patterns of genetic structure. DAPC performs a discriminant analysis after performing PCA. DAPC provides a test for a pre-determined range of numbers of genetic groups (K). We performed DAPC for a range of K values and used a Bayesian Information Criteria (BIC) analysis to determine the range of K values that were most consistent with the data. DAPC differs from PCA in that the discriminant analysis (DA) provides a statistical test of the groupings revealed by PCA. Individuals are permutated among groups and the groups that minimize variance within and maximize variance among clusters for a given value of K emerges as the optimal solution for that K. DAPC does not have an underlying evolutionary model and therefore is not prone to defining discrete genetic groups when the underling genetic structure is continuous in nature (i.e., Isolation By Distance [IBD]; Jombart [3]).

Testing for locus-specific natural selection

We used the R package OutFLANK [9] to test for outlier loci. Outlier loci are loci that exhibit extreme genetic divergence and therefore are possibly influenced by natural selection instead of neutral evolutionary processes (genetic drift and gene flow). OutFLANK performs well in comparison to other approaches for testing for signs of natural selection[9]. We used a heterozygosity cutoff of 0.1 at a locus because loci with low heterozygosity are more likely to be either false positives for detecting outliers or the occurrence of low heterozygosity is a an indicator of loci with sequencing error. We consider outliers as putatively under selection, with the recognition that it is difficult, if not impossible, to separate genetic drift from natural selection with this type of correlational analysis.

Genetic variation within populations

Estimates of within population genetic variation were obtained with the hierfstat R package [10]. We estimated observed (H_O) and expected (H_E) heterozygosity. We summarized mean within-population expected heterozygosity as H_S, following standard population genomics nomenclature. We estimated the proportion of polymorphic loci within populations (*P*) with in-house R code using the package *dplyr*. We used hierfstat to estimate mean allelic richness (*AR*), an estimate of allelic diversity that uses a rarefaction approach to standardize estimates from population samples with different sample sizes. . *AR* tends to provide less information than H_S and *P* about genetic variation with SNPs because there is a maximum of two alleles per locus. We calculated confidence intervals for estimates of H_S using in-house code that bootstraps across individuals (i.e., draws individuals within populations with replacement) to generate 100 replicate data sets.

Inbreeding within populations

Occurrence of ROH are an indication of inbreeding. More precisely, an ROH is a string of adjacent loci that are homozygous, which indicates identity by descent at that set of adjacent loci. We used the R package RZooRoH to examine ROH [11]. This package examines homozygosity by descent (HBD) and has been shown to perform the best among available methods for this type of analysis of RAD-Seq data [11, 12]. We chose a fixed set of HBD bins in order to compare individuals among the three populations that were the most genetically differentiated and depauperate in genetic diversity (BTLS02, BLM01, BLN09). We compared HBD in these three sites to one representative of the more interconnected set of Pecos pupfish populations (BLN01 was arbitrarily chosen) and sheepshead minnows (SHM).

Contemporary N_e

We estimated contemporary (within the last several generations) N_e for the 13 Pecos pupfish population samples to provide information about current effective population size. We used NeEstimator V2 [13] with a random mating model and a minimum allele frequency cutoff of 0.05. We used a 10,000 locus random data subset, which is well-suited for estimating N_e (in terms of number of loci)[14]. We tested for consistency among the 10,000 locus data subsets by generating estimates with a second 10,000 locus random data subset, which yielded highly similar results (data not shown).

Literature Cited

1. Amish, S.J., *Population genomics of pecos pupfish (Cyprinodon pecosensis)*. 2023, University of Montana, Final report for New Mexico Department of Game and Fish.
2. Tian, D., et al., *Severe inbreeding, increased mutation load and gene loss-of-function in the critically endangered Devils Hole pupfish.* Proceedings of the Royal Society B: Biological Sciences, 2022. **289**(1986): p. 20221561.
3. Jombart, T., S. Devillard, and F. Balloux, *Discriminant analysis of principal components: a new method for the analysis of genetically structured populations.* BMC Genetics, 2010. **11**(1): p. 1-15.
4. Kardos, M., et al., *The crucial role of genome-wide genetic variation in conservation.* Proceedings of the National Academy of Sciences, 2021. **118**(48): p. e2104642118.
5. Waples, R.S. and C. Do, *Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution.* Evolutionary Applications, 2010. **3**(3): p. 244-262.
6. Waples, R.S., *What is Ne, anyway?* Journal of Heredity, 2022. **113**(4): p. 371-379.
7. McKinney, G.J., et al., *Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations.* Molecular Ecology Resources, 2017. **17**(4): p. 656-669.
8. Jombart, T., *adegenet: a R package for the multivariate analysis of genetic markers.* Bioinformatics, 2008. **24**(11): p. 1403-1405.
9. Whitlock, M.C. and K.E. Lotterhos, *Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST.* The American Naturalist, 2015. **186**(S1): p. S24-S36.
10. Goudet, J., *Hierfstat, a package for R to compute and test hierarchical F-statistics.* Molecular Ecology Notes, 2005. **5**(1): p. 184-186.
11. Bertrand, A.R., et al., *RZooRoH: An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments.* Methods in Ecology and Evolution, 2019. **10**(6): p. 860-866.
12. Lavanchy, E. and J. Goudet, *Effect of reduced genomic representation on using runs of homozygosity for inbreeding characterization.* Molecular Ecology Resources, 2023. **23**(4): p. 787-802.
13. Do, C., et al., *NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data.* Molecular Ecology Resources, 2014. **14**(1): p. 209-214.
14. Waples, R.K., W.A. Larson, and R.S. Waples, *Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci.* Heredity, 2016. **117**(4): p. 233-240.
15. Weir, B.S. and C.C. Cockerham, *Estimating F-statistics for the analysis of population structure.* Evolution, 1984: p. 1358-1370.

Table 1. Thirteen Pecos pupfish population samples examined in this report. Population labels are generic to prevent identification of true location names. Location represents a broad regional location identity.

Population	Location
BLN01	Bitter Lake National Wildlife Refuge Site 1
BLN02	Bitter Lake National Wildlife Refuge Site 2
BLN04	Bitter Lake National Wildlife Refuge Site 4
BLN20	Bitter Lake National Wildlife Refuge Site 20
BLN07	Bitter Lake National Wildlife Refuge Site 7
BLN09	Bitter Lake National Wildlife Refuge Site 9
BLN15	Bitter Lake National Wildlife Refuge Site 15
BLN03	Bitter Lake National Wildlife Refuge Site 3
BLN05	Bitter Lake National Wildlife Refuge Site 5
BLM01	Bureau of Land Management Overflow
BTLS01	Bottomless Lakes State Park Site 1
BTLS02	Bottomless Lakes State Park Site 2
BTLS03	Bottomless Lakes State Park Site 3

Table 2. Pairwise F_{ST} values (based on Weir and Cockerham [15]) among 13 Pecos pupfish population samples.

	BLN01	BLN02	BTLS02	BTLS01	BLN04	BTLS03	BLM01	BLN07	BLN09	BLN20	BLN15	BLN03	BLN05
BLN01	--												
BLN02	0.007	--											
BTLS02	0.252	0.254	--										
BTLS01	0.026	0.027	0.238	--									
BLN04	0.05	0.053	0.287	0.064	--								
BTLS03	0.068	0.07	0.277	0.048	0.104	--							
BLM01	0.137	0.139	0.338	0.126	0.174	0.165	--						
BLN07	0.051	0.053	0.281	0.057	0.061	0.096	0.167	--					
BLN09	0.091	0.096	0.32	0.096	0.092	0.129	0.204	0.074	--				
BLN20	0.059	0.064	0.286	0.07	0.066	0.108	0.18	0.074	0.079	--			
BLN15	0.008	0.002	0.257	0.029	0.054	0.07	0.14	0.052	0.094	0.064	--		
BLN03	0.001	0.008	0.25	0.026	0.048	0.069	0.136	0.048	0.091	0.06	0.007	--	
BLN05	0.003	0.008	0.249	0.028	0.049	0.067	0.136	0.052	0.092	0.06	0.007	0.004	--

Table 3. Summary of genetic variation in 13 Pecos pupfish population samples. Population label corresponds to Table 1. Sample size is the number of individuals examined from each site. H_O is observed heterozygosity. H_S is mean expected heterozygosity, shown with 95% confidence intervals in parentheses. 95% CIs were generated based on bootstrapping across individuals. An extra significant digit was used for the confidence intervals around H_S compared to other variables. F_{IS} is the degree of departure of observed and expected heterozygosity. *AR* is mean allelic richness. *P* is the proportion of polymorphic loci. All summaries are based on a randomly selected 10,000 SNP subset.

Population	**Sample Size (N)**	H_O	H_S **(95% CI)**	F_{IS}	***AR***	***P***
BLN01	29	0.075	0.0756 (0.0748-0.0762)	0.008	1.18	0.492
BLN02	31	0.075	0.0760 (0.0756-0.0765)	0.011	1.19	0.492
BLN04	31	0.073	0.0731 (0.0719-0.07398)	0.002	1.17	0.374
BLN20	37	0.070	0.0716 (0.0711-0.0721)	0.011	1.17	0.379
BLN07	31	0.073	0.0724 (0.0720-0.0730)	0.003	1.17	0.413
BLN09	30	0.067	0.0686 (0.0680-0.0692)	0.018	1.16	0.312
BLN15	29	0.075	0.0766 (0.0760-0.0772)	0.014	1.19	0.488
BLN03	33	0.074	0.0745 (0.0740-0.0749)	0.010	1.18	0.506
BLN05	32	0.076	0.0765 (0.0760-0.0771)	0.003	1.19	0.519
BLM01	32	0.065	0.0664 (0.0656-0.0670)	0.013	1.15	0.304
BTLS01	32	0.074	0.0756 (0.0747-0.0763)	0.011	1.18	0.475
BTLS02	29	0.043	0.0423 (0.0418-0.0429)	-0.011	1.09	0.134
BTLS03	29	0.073	0.0735 (0.0731-0.0739)	0.011	1.17	0.372

Table 4. Summary statistics (mean expected heterozygosity, H_S and proportion of polymorphic loci [*P*]) for five independent, randomly subsampled, 10,000 locus data sets. The top table shows five replicate subsamples for H_S. Each subsample is designated as 10k1-5, which stands for 10,000 locus ("10k"), replicates 1 through 5. The mean and standard deviation (SD) across the five replicates is shown. The bottom table shows the same information for *P*. Sample size (N) per population sample is shown.

Population	**Sample Size (N)**	**H_S_10k5**	**H_S_10k4**	**H_S_10k3**	**H_S_10k2**	**H_S_10k1**	**Mean**	**SD**
BLN01	29	0.076	0.074	0.076	0.074	0.075	0.075	0.0010
BLN02	31	0.076	0.075	0.077	0.074	0.076	0.076	0.0011
BLN04	31	0.073	0.072	0.073	0.072	0.073	0.073	0.0006
BLN20	37	0.072	0.071	0.073	0.070	0.073	0.072	0.0013
BLN07	31	0.072	0.073	0.075	0.071	0.074	0.073	0.0016
BLN09	30	0.069	0.069	0.069	0.067	0.072	0.069	0.0018
BLN15	29	0.077	0.075	0.077	0.074	0.076	0.076	0.0013
BLN03	33	0.074	0.073	0.076	0.074	0.075	0.074	0.0011
BLN05	32	0.076	0.075	0.077	0.074	0.076	0.076	0.0011
BLM01	32	0.066	0.066	0.066	0.065	0.065	0.066	0.0006
BTLS01	32	0.076	0.077	0.077	0.075	0.076	0.076	0.0008
BTLS02	29	0.042	0.044	0.046	0.044	0.044	0.044	0.0014
BTLS03	29	0.074	0.075	0.075	0.073	0.075	0.074	0.0009

Population	**Sample Size (N)**	***P*_10k5**	***P*_10k4**	***P*_10k3**	***P*_10k2**	***P*_10k1**	**Mean**	**SD**
BLN01	29	0.492	0.494	0.498	0.495	0.497	0.495	0.0021
BLN02	31	0.492	0.487	0.490	0.481	0.491	0.488	0.0045
BLN04	31	0.374	0.378	0.368	0.374	0.375	0.374	0.0036
BLN20	37	0.379	0.376	0.378	0.379	0.387	0.380	0.0042
BLN07	31	0.413	0.423	0.421	0.408	0.419	0.417	0.0059
BLN09	30	0.312	0.311	0.314	0.305	0.313	0.311	0.0036
BLN15	29	0.488	0.475	0.484	0.475	0.480	0.481	0.0055
BLN03	33	0.506	0.500	0.504	0.505	0.511	0.505	0.0040
BLN05	32	0.519	0.507	0.513	0.516	0.511	0.513	0.0045
BLM01	32	0.304	0.314	0.306	0.302	0.304	0.306	0.0046
BTLS01	32	0.475	0.482	0.480	0.475	0.487	0.480	0.0054
BTLS02	29	0.134	0.142	0.141	0.141	0.138	0.139	0.0032
BTLS03	29	0.372	0.391	0.389	0.376	0.378	0.381	0.0082

Table 5. Mean of individual inbreeding coefficients (F_{ROH}) based on inbreeding estimated to have occurred within the most recent 64 generations (T64) for 13 Pecos pupfish populations and the sheepshead minnow (SHM). Contemporary effective population size was estimated based on the linkage disequilibrium-based approach implemented in NeEstimator V2. 95% confidence intervals were generated by jackknifing across individuals, 'inf' stands for infinity. Low estimator precision is due to small sample size relative to the underlying (unknown but large) true N_e at many of these sites. F_{ROH}(T64) estimates were generated with 299,660 loci. N_e estimates were generated with 10,000 randomly subsampled loci, due to the need to minimize linkage for this estimator (the linkage disequilibrium (LD) signal needs to be due to genetic drift, not physical linkage [14]). We did not estimate contemporary N_e for SHM because the precise population of origin is uncertain for this set of individuals collected in the eastern US.

Population	**F_{ROH}(T64)**	**N_e (95% CI)**
BLN01	0.020	3894 (127.2-inf)
BLN02	0.018	46.8 (15.1-inf)
BLN04	0.027	10.5 (2.6-67.6)
BLN20	0.031	45.2 (19-1758.5)
BLN07	0.020	1363.3 (118.4-inf)
BLN09	0.023	1123.5 (52.5-inf)
BLN15	0.016	110.7 (26.4-inf)
BLN03	0.018	2376.3 (89.6-inf)
BLN05	0.020	26.2 (13.5-inf)
BLM01	0.038	665.3 (80.6-inf)
BTLS01	0.018	25.8 (10.8-186.6)
BTLS02	0.142	321.8 (55.3-inf)
BTLS03	0.021	1296.3 (84.2-inf)
SHM	0.013	--

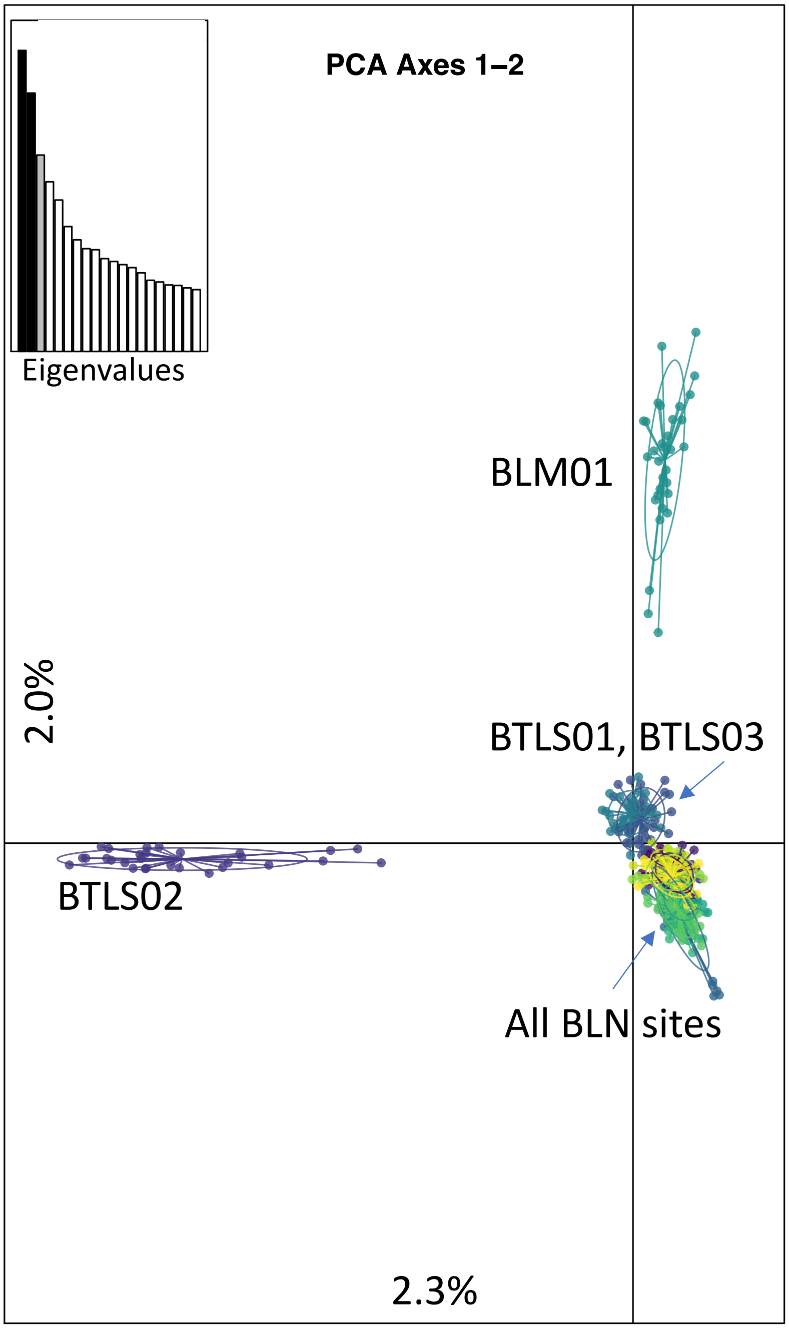

Figure 1. Principal Components Analysis (PCA) of samples from 13 Pecos pupfish populations (sheepshead minnow samples excluded). Dots represent individuals, colored by sampling site, with the inertia ellipse representing the general shape of a group of individuals in the PC space. The horizontal axis (PC 1) explains 2.3% of the variation in the allele frequencies and differentiates the Bottomless Lakes State Park (BTL) site 02 cluster (purple, BTLS02) on the left from all other clusters of Pecos pupfish samples on the right. The other two BTL sites (BTLS01 and BTLS03) and the BLM site (BLM01) were just right of center (note that BTLS01 is obscured by BTLS03), and all Bitter Lake National Wildlife Refuge (BLN) sites were right of center on PC axis 1. The vertical axis (PC 2) explains 2.0% of the variation in allele frequencies and primarily differentiated the BLM01 site (teal) from all other sites. Inset shows a histogram of eigenvalues, which roughly correspond to the proportion of overall variance explained by a PC axis. The first and second PC axis eigenvalues are shown in black because those axes are shown in this figure.

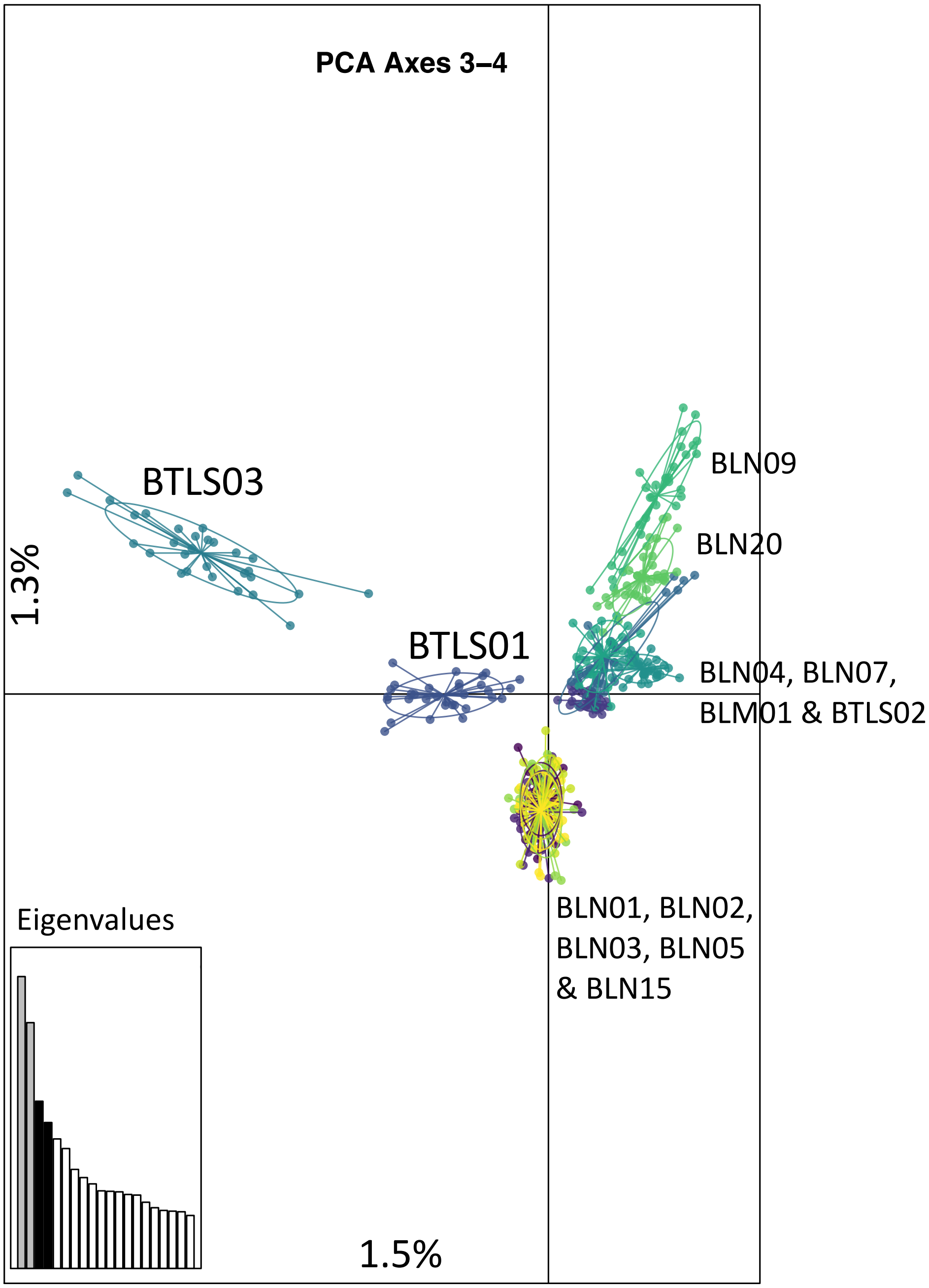

Figure 2. Principal Components Analysis (PCA) of samples from 13 Pecos pupfish populations. Dots represent individuals, colored by sample site, with the inertia ellipse representing the general shape of a group of individuals in the PC space. The horizontal axis (PC 3) explains 1.5% of the variation in the allele frequencies and primarily differentiates two Bitter Lake National Wildlife Refuge (BLN) sinkhole sites (green, BLN09 and BLN20) on the right from the Bottomless Lake site 03 on the left (blue, BTLS03). The vertical axis (PC 4) explains 1.3% of the variation in allele frequencies and most strongly differentiates BTLS03 from a subset of the BLN waterbodies. Note that the cluster with the lowest PC4 values contains overlapping sites BLN01, BLN02, BLN03, BLN05, and BLN15. Inset shows a histogram of eigenvalues, which roughly correspond to the proportion of overall variance explained by a PC axis. The third and fourth PC axis eigenvalues are shown in black because those axes are shown in this figure. The first and second PC axis eigenvalues are shown in grey (corresponding to Figure 1), the remainder in white.

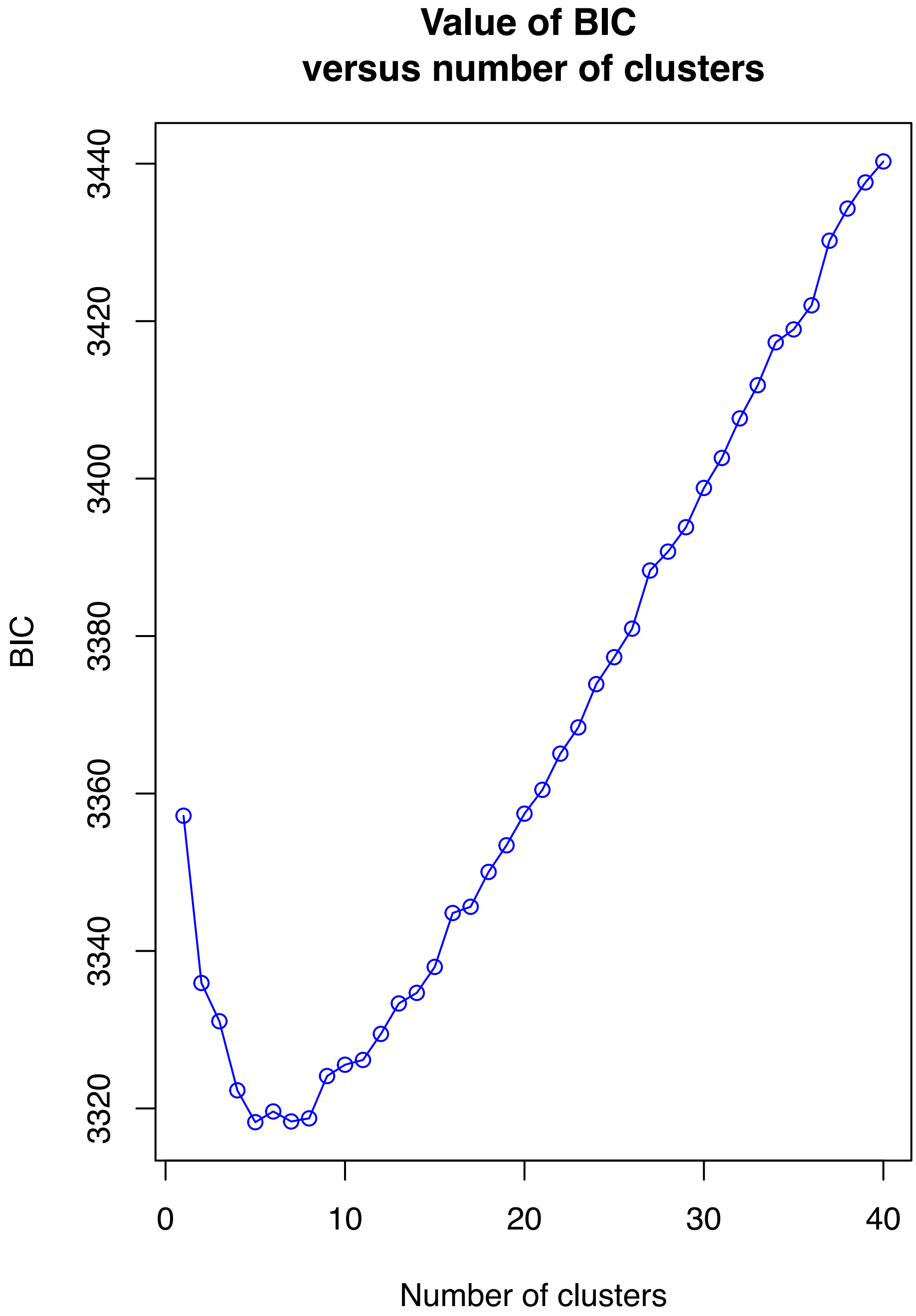

Figure 3. Bayesian Information Criteria (BIC) plot for Discriminant Analysis of Principal Components (DAPC) analysis of 117,309 Single Nucleotide Polymorphisms (SNPs) in 13 Pecos pupfish populations. Lower BIC values indicate a number of pre-assigned genetic clusters (K) that provides a more plausible explanation of genetic variation in this data set from 13 Pecos pupfish populations than other K values. General guidelines suggest interpreting the smallest value of K associated with the 'trough' in BIC values to represent the most likely number of genetically distinct groups . Here, K = 5 is the lowest BIC and BIC starts to increase at K = 6. We therefore based interpretation on these two K values. BIC declines slightly again at K = 7 and K = 8, but these values are likely to cause over-splitting and over-interpretation.

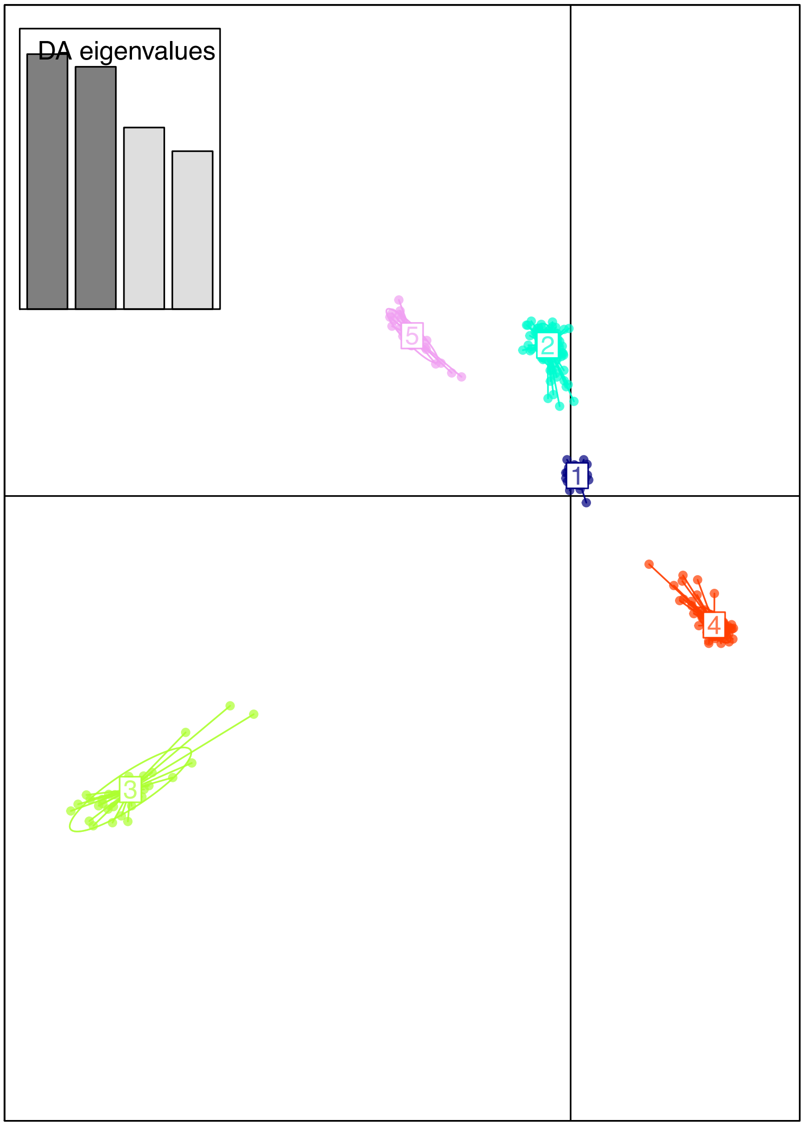

Figure 4. Discriminant Analysis of Principal Components (DAPC) results for 13 Pecos pupfish samples for K = 5 showing discriminant axes 1 (x-axis) and 2 (y-axis). Cluster numerical assignment is arbitrary. Each circle represents one individual. Inertia ellipses represent the general shape of a group of individuals in the DA space. Cluster 1 (dark blue) contains individuals from Bitter Lake National Wildlife Refuge (BLN) site 04 (BLN04). Cluster 2 (teal) contains individuals from sites at Bottomless Lakes State Park (BTL) (i.e., BTLS01 and BTLS03 and from BLN (i.e., BLN07, BLN09, and BLN20). Cluster 3 (lime green) contains individuals from the Bureau of Land Management site (i.e., BLM01). Cluster 4 (red) contains individuals from BLN01, BLN02, BLN03, BLN05, and BLN15. Cluster 5 (pink) contains individuals from BTLS02. Insert shows eigenvalues from the DA (see Figures 1 and 2 for an explanation of eigenvalues).

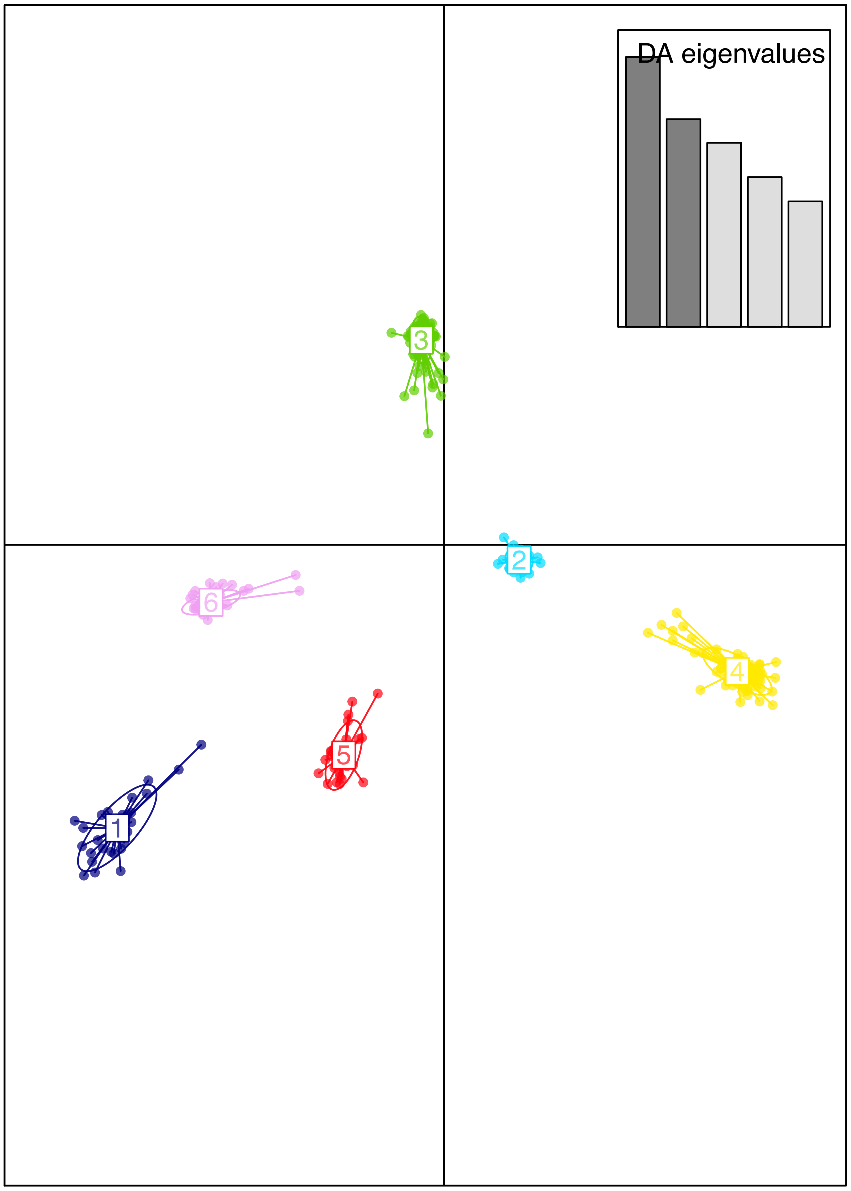

Figure 5. Discriminant Analysis of Principal Components (DAPC) results for 13 Pecos pupfish samples for K = 6 showing discriminant axes 1 (x-axis) and 2 (y-axis). Cluster numerical assignment is arbitrary. Each circle represents one individual. Inertia ellipses represent the general shape of a group of individuals in the DA space. Cluster numerical assignment is arbitrary. Cluster 1 (dark blue) contains individuals from the Bureau of Land Management site (BLM01). Cluster 2 (teal) contains individuals from Bitter Lake National Wildlife Refuge (BLN) site 04 (BLN04). Cluster 3 (lime green) contains individuals from sites BLN01, BLN02, BLN03, BLN05, and BLN15. Cluster 4 (yellow) contains individuals from sites BLN07, BLN09, and BLN20. Cluster 5 (red) contains individuals from Bottomless Lakes State Park (BTL) site 02 (BTLS02). Cluster 6 (pink) contains individuals from sites BTLS01 and BTLS03.

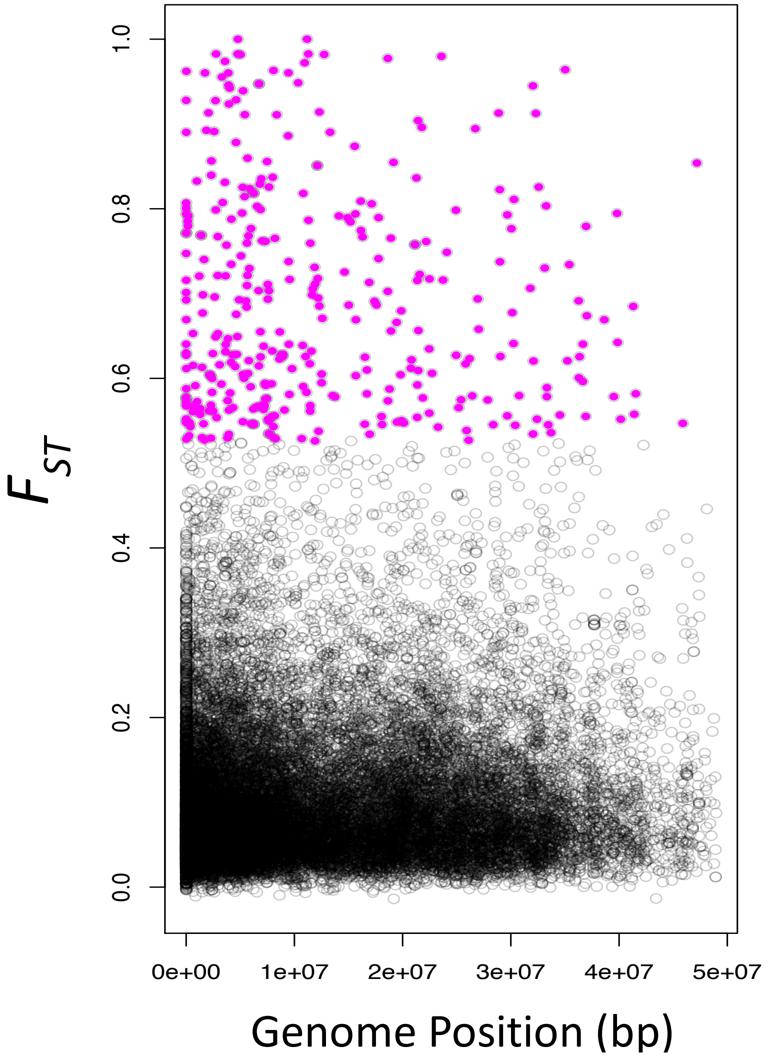

Figure 6. Results for outlier locus test performed using OutFLANK. Each data point represents a single locus. Genome position in base pairs is on the x-axis. F_{ST}, a measure of genetic differentiation, is on the y-axis. Loci represented by black circles did not exceed the F_{ST} cutoff value and are therefore more likely to be neutral (not influenced by natural selection). Loci represented by pink circles were statistically significantly different in terms of their F_{ST} values and are putative outliers. That is, allele frequencies of these pink loci are more divergent. potentially due to divergent natural selection. This analysis was based on 117,319 loci. Loci given a genome position value of zero could not be mapped to the genome.

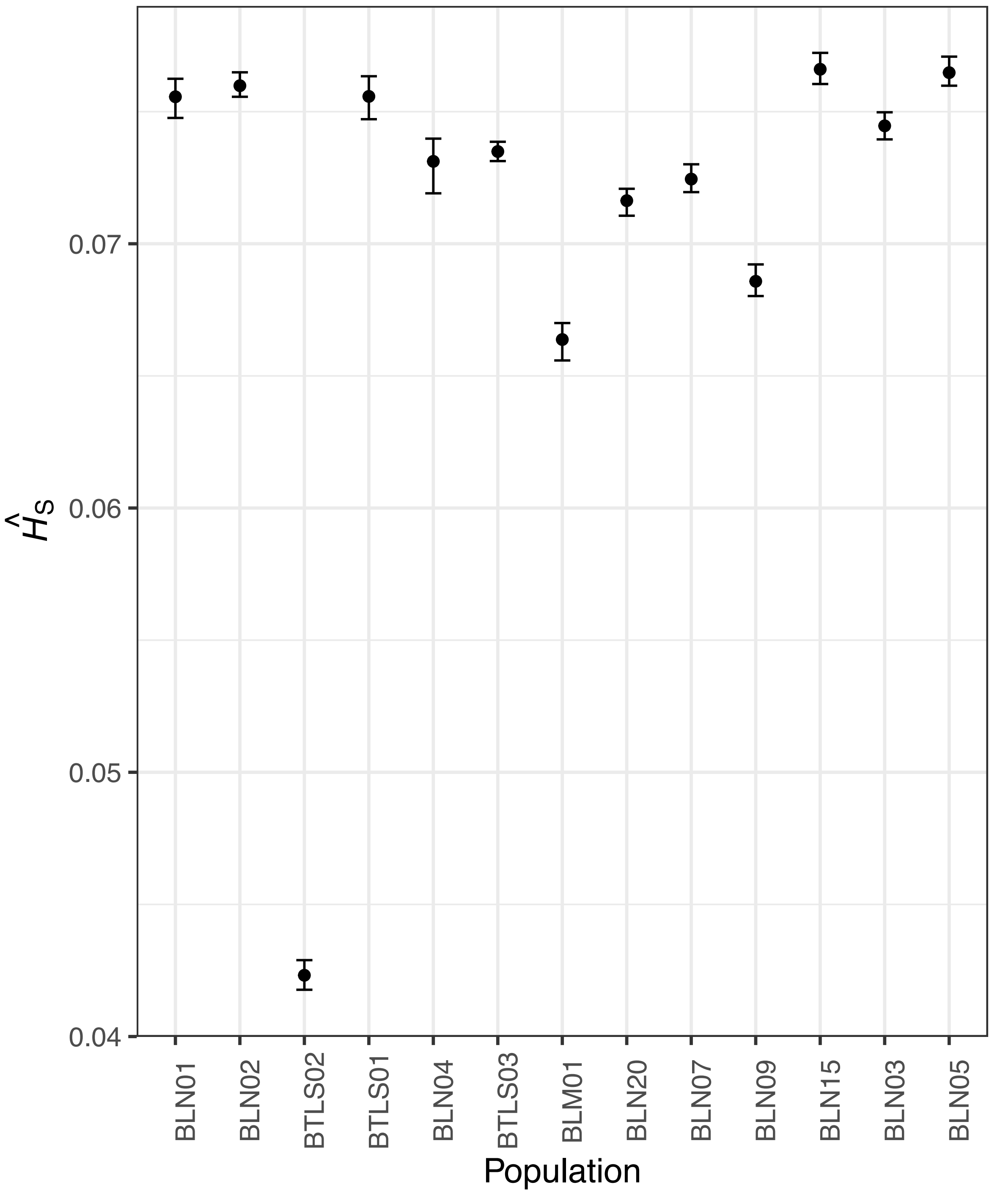

Figure 7. Estimated mean within-population expected heterozygosity (H_S) for samples from 13 Pecos pupfish populations. Error bars represent confidence intervals generated by bootstrapping across individuals. A 10,000 locus subset of the data was used to generate estimates.

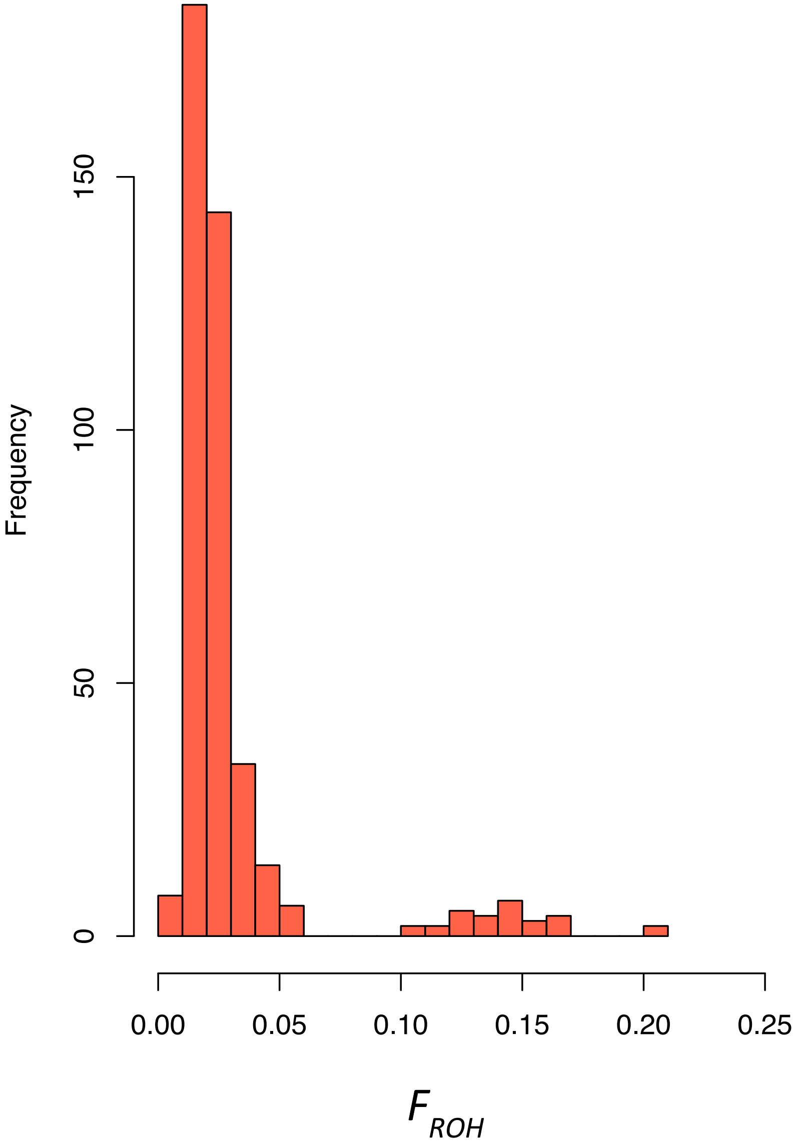

Figure 8. Distribution of individual inbreeding coefficients (F_{ROH}) based on a cutoff of T = 64 generations in the past. Long runs of homozygosity are indicative of more recent inbreeding. Evaluation of the likelihood of inbreeding was estimated only for the most recent 64 generations. All values greater than 0.10 (i.e., more indicative of inbreeding) belong to the Bottomless Lakes State Park 02 (i.e., BTLS02) population sample.

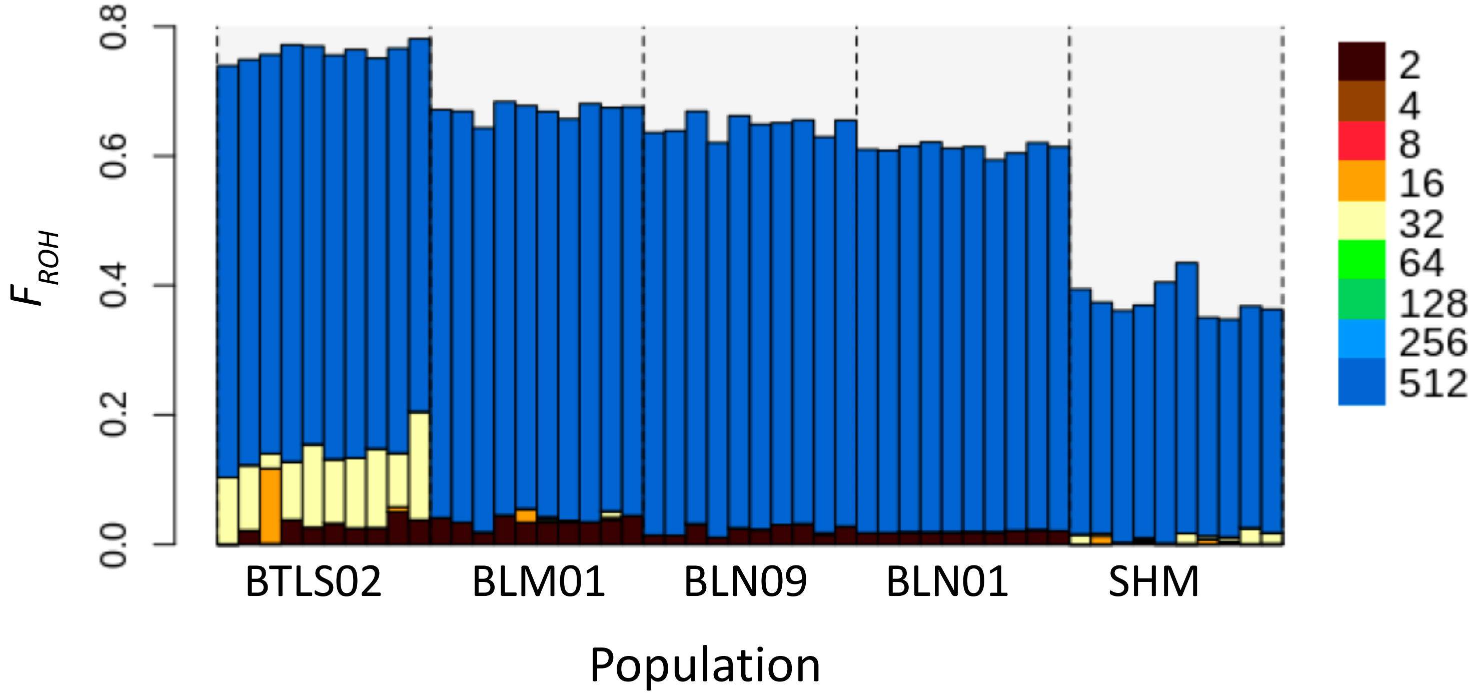

Figure 9. RZooRoH results for five Pecos pupfish populations. Population labels are on the x-axis. F_{ROH}, or individual inbreeding coefficients, are on the y-axis. Each column represents a single individual, and 10 individuals were randomly selected from each site due to the computational intensity of the model. $\mathrm{F}_{\mathrm{ROH}}$ are shown in nine color-coded categories, corresponding to the approximate generations in the past during which mating among common ancestors is estimated to have occurred (2 through 512 generations). The blue bars, representing estimated inbreeding approximately 512 generations in the past, are based on the weakest signal in the data set (short runs of homozygosity) and therefore provide the least robust inference. Bars corresponding to more recent inbreeding (less than or equal to approximately 32 generations in the past) are based on longer runs of homozygosity and are expected to be more reliable. Individuals from Bottomless Lakes State Park (BTL) site 02 (i.e., BTLS02) had a clear signal of elevated recent inbreeding compared to the other individuals and populations shown. BLM = Bureau of Land Management; BLN = Bitter Lake National Wildlife Refuge; SHM = Sheepshead Minnow.