POPULATION GENOMICS OF PECOS PUPFISH (*Cyprinodon pecosensis*)

Final report for New Mexico Department of Game and Fish
April 10, 2023

Prepared by: Amish, S.J - Conservation Genomics Consulting, Missoula, MT

POPULATION GENOMICS OF PECOS PUPFISH (*Cyprinodon pecosensis*)

Summary of DNA extractions, Sbf1 RAD library sequencing, genotyping, and assessment of Sheepshead minnow admixture

Methods & Results

DNA was extracted from 450 Pecos pupfish (*Cyprinodon pecosensis*) tissue samples that were collected at the Bitter Lake National Wildlife Refuge (BLNWR), Bottomless Lakes State Park (BLSP), and a Bureau of Land Management (BLM) property by the New Mexico Department of Game and Fish. Thirty additional samples, consisting of negative controls and duplicate samples, were included to bring the total number of samples sequenced to 480 (i.e., 5 plates of 96 samples; Table 1). Each of the 5 plates included at least 2 negative controls and 2 duplicate control samples. DNA was extracted from the tissue samples, *Sbf1* restriction site associated DNA libraries were prepared, and Illumina sequencing libraries were constructed. Laboratory methods generally followed the DNA extraction and "BestRAD" protocols outlined in Ali et al. 2016.

Table 1. *Summary table of the Pecos pupfish samples analyzed. Site name is used in Figs. 1, 2 and 3. Average DNA concentration for all samples from a waterbody is given in nanograms per microliter (ng/µl). Samples with low DNA had concentrations <15 ng/µl. Samples removed due to missing genotypes had >65% missing data.*

Site	Sample Year	Avg. DNA conc. (ng/ul)	# Samples extracted	# Duplicate samples	# Samples sequenced	# Samples low DNA	# Rem missing genotypes	# Samples genotyped	#Duplicates genotyped
BLN01	**2021**	**98.3**	**34**	**0**	**34**	**0**	**5**	**29**	**0**
BLN02	**2022**	**124.7**	**35**	**4**	**39**	**0**	**8**	**31**	**2**
BLN04	**All**	**109.0**	**37**	**5**	**42**	**0**	**8**	**31**	**4**
	2021	91.7	13	5	18	0	3	10	4
	2022	121.9	24	0	24	0	5	21	0
BLN20	**2021**	**104.5**	**35**	**4**	**39**	**0**	**2**	**37**	**4**
BLN07	**2021**	**93.9**	**32**	**0**	**32**	**0**	**1**	**31**	**0**
BLN09	**All**	**96.5**	**36**	**0**	**36**	**2**	**6**	**30**	**0**
	2021	167.6	1	0	1	0	1	0	0
	2022	94.4	35	0	35	2	5	30	0
BLN15	**2021**	**114.5**	**30**	**1**	**31**	**0**	**2**	**29**	**1**
BLN03	**2021**	**73.0**	**35**	**0**	**35**	**0**	**2**	**33**	**0**
BLN05	**2021**	**66.4**	**29**	**3**	**32**	**0**	**0**	**32**	**3**
BLM01	**2021**	**117.8**	**33**	**0**	**33**	**0**	**1**	**32**	**0**
BTLS01	**2021**	**99.2**	**30**	**4**	**34**	**0**	**2**	**32**	**4**
BTLS02	**2021**	**75.6**	**35**	**0**	**35**	**0**	**6**	**29**	**0**
BTLS03	**2021**	**76.5**	**30**	**0**	**30**	**0**	**1**	**29**	**0**
SHM	**2022**	**95.1**	**9**	**9**	**18**	**2**	**5**	**13**	**6**
	2022	**2.3**	**10**	**0**	**10**	**0**	**NA**	**0**	**0**
Grand Total		**94.8**	**450**	**30**	**480**	**4**	**49**	**418**	**24**

Sequence reads were generated for the Pecos pupfish RAD libraries on an Illumina HiSeq sequencer using paired-end reads 150 base-pairs long. The five plates were sequenced as a single library on 3 sequencing lanes. The resulting high-quality sequence data generated an average of approximately 16 million read pairs per sample. PCR duplicates represented an average of 18% of the read pairs and were removed using the clone_filter module in STACKs. Adapter sequences were removed and low-quality bases were trimmed from sequence reads using TRIMMOMATIC.

The remaining read pairs were then mapped to the recently published genome of a Caribbean pupfish species (*C. brontotheroides*; UCB_Cbro_1.0) that is closely related to the Pecos pupfish, following similar work on the Devils Hole pupfish (*C. diabolis*) by Tian et. al (2022). Read pairs were mapped using default short read parameters with the program BWA mem. GATK HaplotypeCaller was then used to take the mapped reads to generate individual g.vcf files, which were genotyped using GATK GenotypeGVCFs. The resulting raw genotypes were filtered in accordance with Broad Institute's "Best practices for non-model organisms" (Van der Auwera 2020). Genotypes were further filtered by requiring a minimum mapping quality of 30, minimum read depth of 7, heterozygote allele balance (AB) and minimum read depth (.25>AB<.75, 10), retaining only bi-allelic SNPs, normalizing to the reference, requiring a minor allele count of 3, removing loci missing > 50% individual genotypes, and removing individuals missing > 65% of loci genotypes. The resulting data set contained 418 samples and 973,036 SNP loci.

The presence of the sheepshead minnow (*C. variegatus*) in the region suggested the possible existence of *C. variegatus* X *C. pecosensis* hybrids in the data set, so 9 sheepshead minnow reference samples from Florida were also sequenced. To identify outlier samples (e.g., individuals with unique or discordant genetic variation) and to identify potential hybrids, the R package ADGENET was used to perform a principal component analysis (PCA) on the scaled allele frequencies. Figure 1 shows strong divergence between all Pecos pupfish samples and the 9 sheepshead minnow samples included in our analysis. Only the label for the uppermost site is visible in the cluster on the left because all Pecos pupfish samples were plotted so close together. Any hybridized samples would be expected to be intermediate to the Pecos pupfish and sheepshead minnow clusters on the horizontal axis (PC1). Using this test, we see no evidence of any *C. variegatus* X *C. pecosensis* hybrids.

After removing the sheepshead minnow samples, a second PCA was performed. Figures 2 & 3 show low levels of Pecos pupfish allele frequency differences among the waterbody clusters. In Figure 2, the horizontal axis (PC1) shows the strongest differences in allele frequencies between the BLSP Site BTLS02 and the rest of the waterbodies. The vertical axis (PC2) shows a similar level of allele frequency differentiation between BLM Site BLM01 and all other waterbodies. In Figure 3, the horizontal and vertical axis suggest further groupings, potentially by geography or eco-type. A map showing the locations from which the samples were collected would help refine our interpretation of the PCA results.

Figure 1. *PCA of the Pecos pupfish and sheepshead minnow samples. 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 33% of the variation in allele frequencies and differentiates the sheepshead minnow cluster (green, SHM) on the right from the cluster of all Pecos pupfish samples on the left (yellow, BLN05). Due to low variation among the pupfish samples, only the label for the BLN05 cluster is visible. The vertical axis (PC 2) explains 5% of the variation in allele frequencies and differentiates sheepshead minnow individuals (green dots). There is no evidence for* C. variegatus X C. pecosensis *hybrids, which would be intermediate to the two clusters along the horizontal axis.*

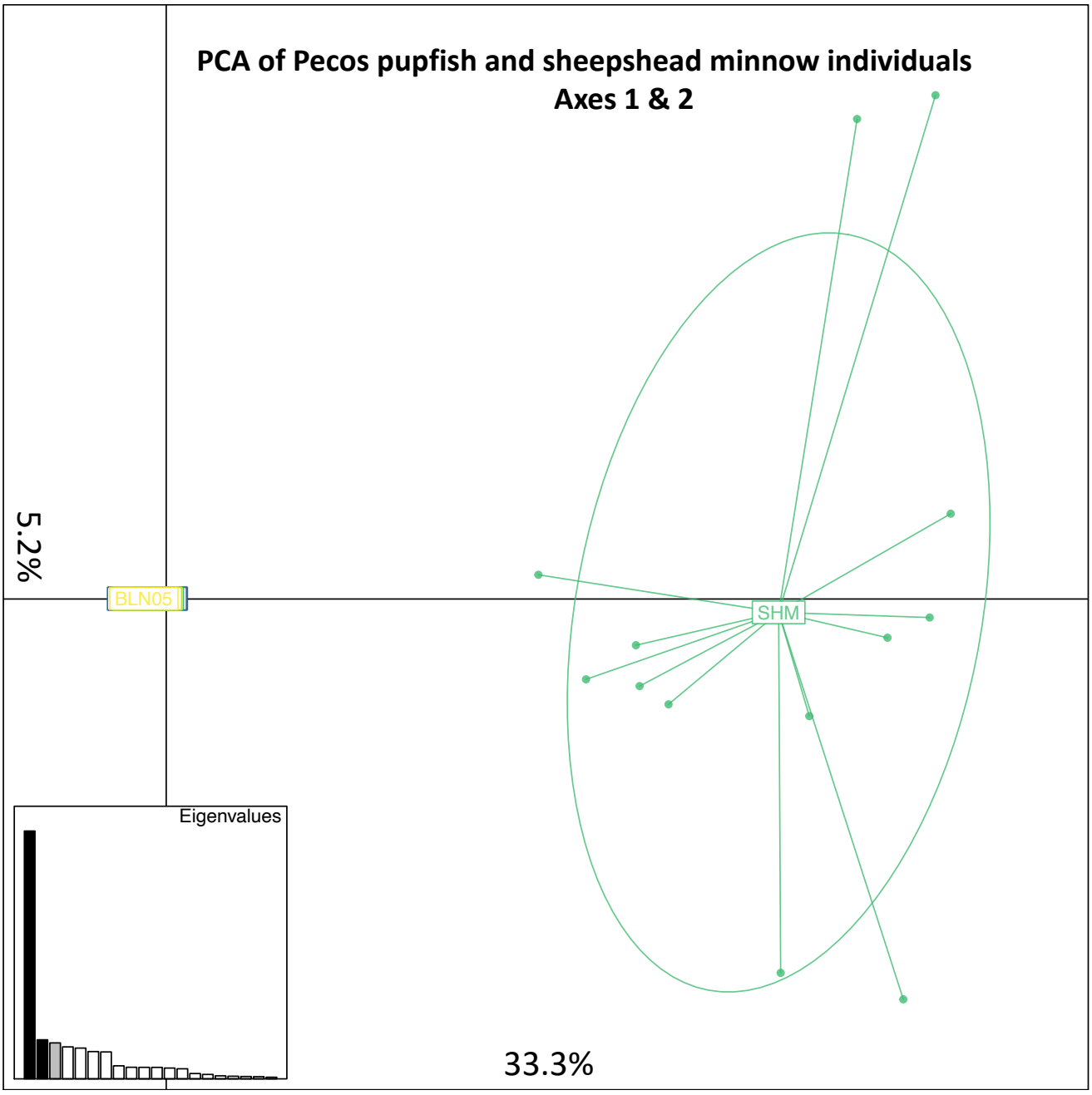


Figure 2. *PCA of the Pecos pupfish samples only (sheepshead minnow samples removed). Dots represent individuals, colored by sampling site, with the inertia ellipse representing the general shape of a group of individuals in the PC space. The horizontal axis (PC 1) explains 2.1% of the variation in the allele frequencies and differentiates the BLSP Site 02 cluster (purple, BTLS02) on the right from the clusters on the left, which include the two other sites at BLSP (BTLS01 and BTLS03; just right of center) and all of the BLNWR sites. The vertical axis (PC 2) explains 1.8% of the variation in allele frequencies and differentiates the BLM site (teal, BLM01) from all other sites.*

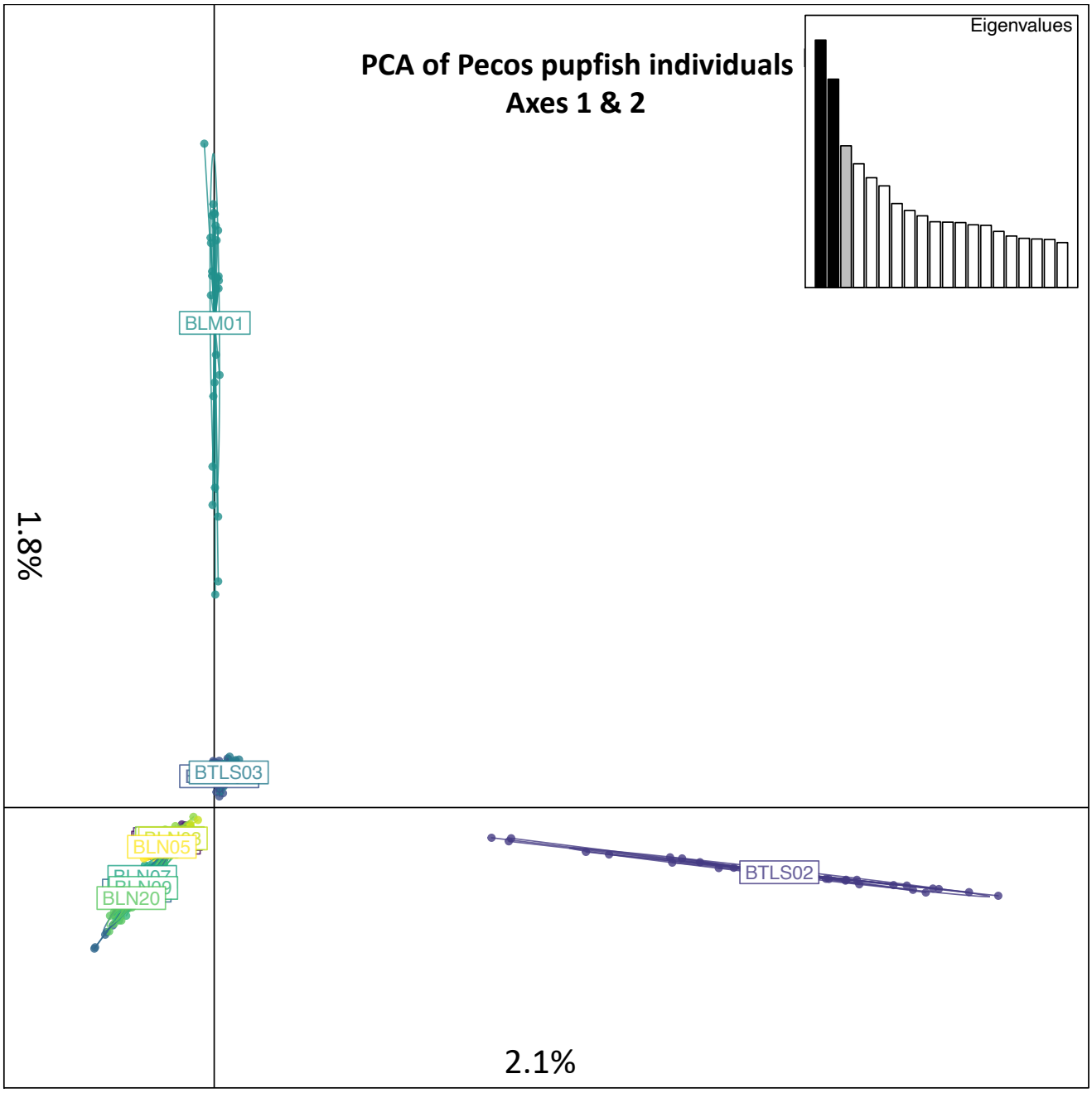
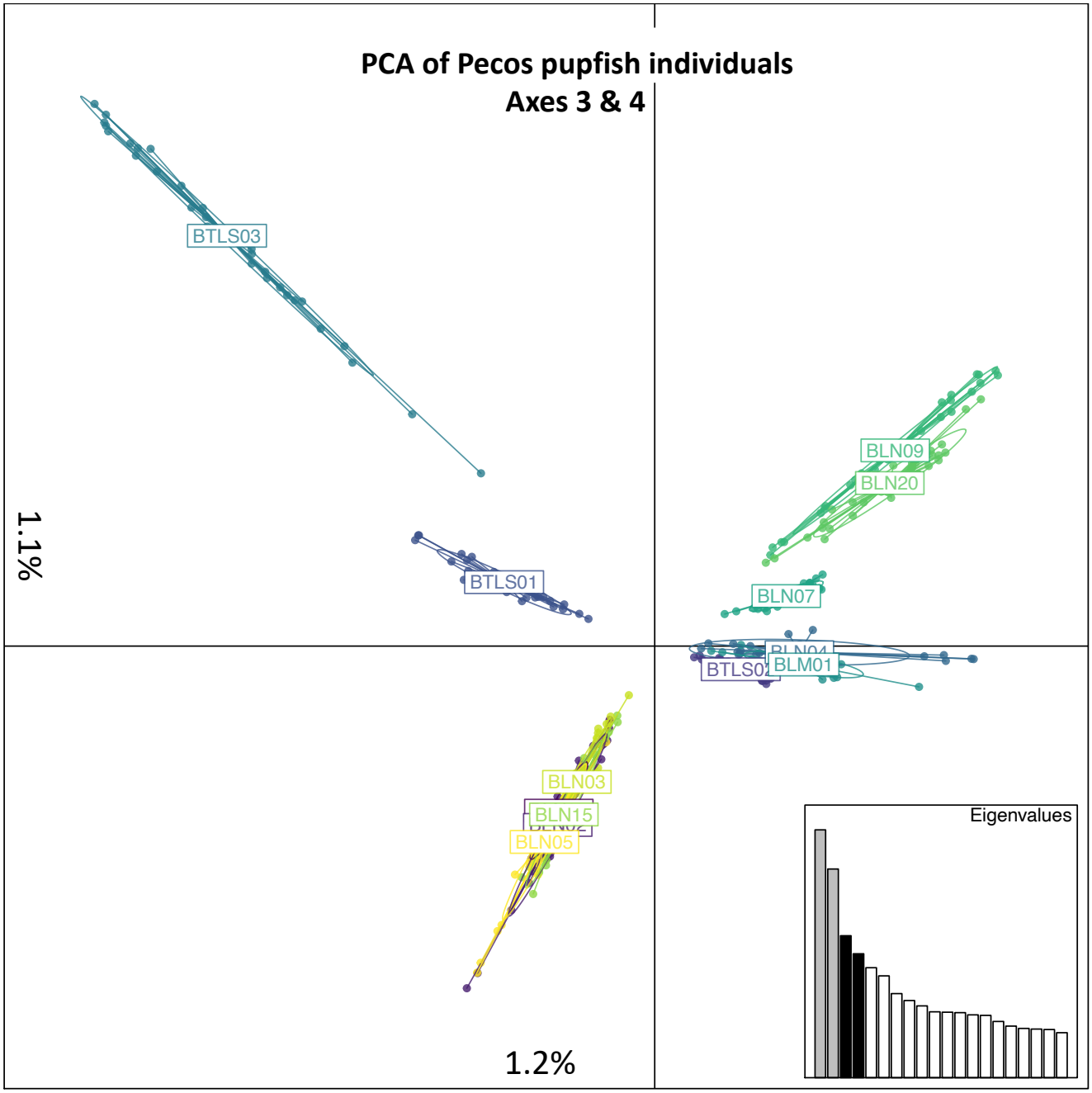


Figure 3. *PCA of the Pecos pupfish samples only. Dots represent individuals, colored by sampling site, with the inertia ellipse representing the general shape of a group of individuals in the PC space. The horizontal axis (PC 3) explains 1.2% of the variation in the allele frequencies and differentiates two BLNWR sinkhole sites (green, BLN09 and BLN20) on the right from BLSP Site 03 on the left (blue, BTLS03). The vertical axis (PC 4) explains 1.1% of the variation in allele frequencies and most strongly differentiates BLSP Site 03 (blue, BTLS03) from a subset of the BLNWR waterbodies (yellow, BLN05; light green, BLN03).*



References

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389-400.

Tian D., Patton A.H., Turner B.J., Martin C.H. (2022). Severe inbreeding, increased mutation load and gene loss-of- function in the critically endangered Devils Hole pupfish. *Proc. R. Soc. B* 289: 20221561. https://doi.org/10.1098/rspb.2022.1561.

Van der Auwera G.A. & O'Connor B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media.