Title

Applying RAD Sequencing to Conserve the Genetic diversity of California Freshwater Fish

Author

su, yingxin

Publication Date

2023

Peer reviewed|Thesis/dissertation

Applying RAD Sequencing to Conserve the Genetic diversity of California Freshwater Fish

By

YINGXIN SU

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Animal Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Andrea Schreier, Chair

Andrew Rypel

Peter B. Moyle

Committee in Charge

2023

Acknowledgement

Firstly, I extend heartfelt gratitude to my parents, Caiwen Yu and Li Su, and my boyfriend, Shunyang Wang, for their unwavering love and support. My adorable parrots - Lemon & Lime, CoCo, Bai & Hui, Ginger, and Bean have been my warm companions, adding cheerful moments to my journey.

Secondly, I am profoundly thankful to the faculty at the Genomic Variation Laboratory for their invaluable assistance and financial support which were instrumental in the completion of my PhD. I appreciate Amanda J. Finger and Matthew A. Campbell for our engaging weekly discussions which propelled my projects forward and provided a platform for brainstorming solutions to technical challenges in RAD Sequencing. Our conversations about travels and parrots added a dash of fun to the academic rigor. Special thanks to Andrea Schreier, the Principal Investigator of Genomic Variation Laboratory, whose guidance was indispensable whenever I encountered hurdles in my PhD journey. I always felt better after talking with her. I am grateful to Melanie LaCava for meticulously reviewing and editing my dissertation, her expertise in writing was my lifesaver amidst a sea of grammatical errors.

Thirdly, I acknowledge the support from the faculty of the Animal Science Department and the Center for Watershed Science. My gratitude extends to my dissertation committee members: Andrea Schreier, Peter Moyle, and Andrew Rypel, as well as the qualifying exam committee: Andrea Schreier, Andrew Rypel, Rachel Bay, Jackson A. Gross, and Hao Cheng. Collaborating with Peter Moyle on Chapter One during the COVID-19 pandemic was a memorable experience, and I am thankful for his accessibility amid the systemic disruptions.

Moreover, my PhD journey was brightened by the camaraderie of many amazing individuals. A shoutout to Ensieh Habibi, my delightful neighbor in the lab office room, and Xinyu Tang, a cherished friend with whom tea/milk tea sessions, lunches, and snack breaks were always rejuvenating. I enjoyed the wonderful times at TAFS 2023 with Alana Luzzio, Anderson Tate, Emily Funk, and Sean Canfield.

The completion of my PhD is a testament to the collective support and wonderful interactions with all mentioned and many more. Thanks for having all of you, and I am looking forward to hearing your big accomplishment in your work or life. Let's be a killer in science.

Dissertation Abstract

Freshwater fishes in California are challenged by anthropogenic activities, especially habitat fragmentation and introduction of invasive species. Restriction Site Associated (RAD) Sequencing, a commonly applied Next Generation Sequencing technique in non-model organisms, can provide essential and actionable genomic information for managers to assess population statuses and for developing species conservation plans. In this dissertation, I explored varied application of RAD sequencing to two native California species, Speckled Dace (*Rhinichthys osculus*) and Paiute cutthroat trout (*Oncorhynchus clarkii seleniris*). The Speckled Dace are widespread in western North America, and to this point, have been considered a single species. In Chapter One, the California Speckled dace is divided into three distinct evolutionary lineages based on the genetic differentiation detected by thousands of single nucleotide polymorphisms extracted using RAD sequencing. Paiute cutthroat trout are a subspecies of cutthroat trout threatened by competition and hybridization with introduced non-native trout. To protect and recover the subspecies, populations were previously translocated by conservation biologists to nine refuge populations thought to approximate historical habitats. Yet long-term genetic monitoring is required to monitor genetic diversity over space and time to assess the efficacy of these management actions. In Chapter Two, we selected 1,114 SNPs that generate comparable results in genetic diversity and genetic population structure to 6,187 SNPs in RAD sequencing after resolving technical questions related to the RAD sequencing approach. Results ultimately demonstrate the SNP panel is useful for genetic monitoring of population structure and heterozygosity, and thus helpful for conservation management. In Chapter Three, we applied 1,114 SNPs to estimate genetic diversity, genetic population structure for all nine refuge populations. We also applied these SNPs to estimate parent-offspring relationships and the

change of genetic diversity after the translocation between two refuge populations. We also applied these SNPs to estimate parent-offspring relationships and the change of genetic diversity after the translocation between two refuge populations.

Table of Contents

Chapter 1: Population genomic analysis of the Speckled Dace species complex (*Rhinichthys osculus*) identifies three distinct lineages in California.

Abstract

Speckled Dace *Rhinichthys osculus* is small cyprinoid fish that is widespread in western North America. In California and elsewhere it is currently treated as a single species with multiple subspecies, many undescribed. However, these subspecies may represent evolutionary lineages that are cryptic species because they cannot be distinguished using standard morphometric techniques. To determine evolutionary lineages within California populations of Speckled Dace, we collected samples from 38 locations in the western USA, with a focus on California. We used RAD sequencing to extract thousands of SNPs across the genome to identify genetic differences among all the samples. We performed principal component analysis, admixture analysis, estimated pairwise F_{ST}, and constructed molecular phylogenies to characterize population genetic and phylogenetic relationships among sampled Speckled Dace populations. Our analyses detected three major lineages of Speckled Dace in California that align with geography: 1) Sacramento River, Central California Coast, Klamath River and Warner Basin; 2) Death Valley and Lahontan Basin; 3) Santa Ana River basin, in southern California. These lineages fit well with the geologic history of California, which has promoted long isolation of populations of Speckled Dace and other fishes. The presence of distinct evolutionary lineages indicates that Speckled Dace in California should be managed with distinct population segments to preserve

within-species diversity. This study highlights the importance of genetic analyses for conservation and management of freshwater fishes.

1. Introduction

The Speckled Dace *Rhinichthys osculus* is a small (usually <10 cm total length) cyprinoid (Cypriniformes, Leuciscidae) fish that is widely distributed across western North America. It is found in northern Mexico and southern, central, and northern California, the Great Basin, and the Pacific Northwest to southwestern Canada (Moyle 2002; Smith et al. 2017). Despite its wide distribution, the Speckled Dace is considered to be one highly variable species, albeit with numerous subspecies, many undescribed and of uncertain taxonomic status (Moyle 2002; Smith et al. 2017). We refer to the species therefore as the Speckled Dace complex (SDC). The SDC diverged from the Longnose Dace *Rhinichthys cataractae* of eastern and northwestern North America over 6 million years ago (mya) (Spencer et al. 2008). The common ancestor of the SDC was presumably initially isolated from Longnose Dace in the ancestral Columbia River and then spread throughout the western USA, parts of Mexico, and British Columbia as the result of geologic events that connected and disconnected watersheds (Smith et al. 2017). Speckled Dace Complex populations are found in a wide array of habitats, from desert springs to large rivers and lakes to (most typically) small to medium-sized streams. Their morphology is variable but generally reflects the habitat in which a particular population lives. For example, narrow caudal peduncles and large pectoral fins characterize swift-water populations while more robust bodies, thicker caudal peduncles and smaller pectoral fins characterize slow-water populations (Sada et al. 1995; Page and Burr 2011; Smith et al. 2017). Consequently, morphological and meristic differences do not reflect phylogenetic relationships among populations of the SDC, resulting in a long and complex taxonomic history (Smith et al. 2017).

The most comprehensive study of the systematics of the SDC is that of Smith et al. (2017) who compared dace populations from throughout western North America, using mitochondrial DNA (mtDNA), morphology, fossils, and the geologic record of the entire region. While their analyses indicated multiple lineages, they concluded that there was considerable, if sporadic, gene flow among populations, reflecting complex geologic events that promoted both connectivity and isolation. Smith et al. (2017) suggest that gene flow has prevented the formation of morphologically distinct populations that might be defined as species through the process of reticulate evolution.

Overall, genetic studies have produced mixed results as to whether or not any lineages in the SDC are distinct enough to be designated as species or subspecies. The default position is to follow Spencer (2008) and Smith et al. (2017) that the SDC is a single species throughout its range because the various populations lack diagnostic characteristics that would allow them to be described as distinct phylogenetic entities. This default position is particularly problematic for California, a region rich in endemic fish species, many of which are threatened with extinction (Leidy and Moyle 2021; Moyle 2002). Notably, California SDC populations are among those most distant from the region of origin in the Columbia River and also among the most southern populations of the taxon. California SDC populations thus reflect their remarkable record of colonizing new regions during the wetter periods of the Pleistocene and then adapting to new conditions as waters they colonized became smaller and more isolated (Smith et al. 2017).

In this paper, we analyze Speckled Dace relationships using restriction-site associated DNA sequencing data (RAD-seq). This approach is well suited for analyzing the SDC because it uses thousands of loci distributed across the genome from each individual rather than only a

single locus or handful of loci compared with earlier methods. For further discussion of this approach to resolving issues with identifying cryptic fish species, see Baumsteiger et al. (2017).

We investigated the following question using RAD sequencing to look at relationships among populations of Speckled Dace collected over much of their wide range: Based on genetic distinctness, should members of the SDC in California be treated as a single lineage or multiple lineages for conservation and management?

1.1 Previous Genetic Studies of Speckled Dace

Historically, many populations of Speckled Dace were described as separate species. For example, Jordan and Evermann (1896) list nine species, which had mostly been described based partially on their isolation from other populations and partially on morphological and meristic characteristics even though these characters overlapped among populations (Jordan and Evermann 1896). However, the presence of many isolated populations of Speckled Dace with similar adaptations to local environments and hence convergent morphologies suggests that cryptic species exist within the SDC and that some of the recognized subspecies (listed in Smith et al. 2017) could be recognized as species (Evermann and Meek 1896).

The advent of molecular genetic techniques resulted in renewed efforts to examine diversity within the SDC. Resulting genetic information was used to test hypotheses of evolutionary relationships among populations and generate biogeographic scenarios relating Speckled Dace to the development of the western aquatic landscape (Smith et al. 2017). Thus far, mitochondrial DNA has been the primary genetic approach used to investigate the systematics of Speckled Dace (Oakey et al. 2004; Smith et al. 2017). Smith et al. (2017) compared dace populations from throughout western North America and concluded that while geographically-based lineages existed, as shown by Oakey et al. (2004), there was no basis for declaring them

separate species. Oakey et al. (2004) used restriction sites in the mitochondrial genome of dace distributed across the western USA to construct a molecular phylogeny. They found a close match between mtDNA patterns and the geologic history and isolation of drainage basins, concluding that the SDC consisted of three main evolutionary lineages: (1) Colorado River Basin and Los Angeles Basin, (2) Great Basin (Snake River, Bonneville, Death Valley and Lahontan basins) and (3) Columbia and Klamath-Pit Rivers (Oakey et al. 2004). Pfrender et al (2004) showed that mtDNA patterns reflected long isolation of populations in five river basins in Oregon and suggested that some of the lineages were distinct enough to be considered species. In contrast, Billman et al (2010) did not find species-level differences in mtDNA among SDC members found in Great Basin waterways (Snake, Bonneville, Lahontan).

More narrowly, Ardren et al. (2010) applied mtDNA analysis to dace from throughout the Warner Basin and concluded that some lineages could qualify as a species. Hoekzema and Sidlauskas (2014) also examined SDC fish from the Warner Basin, along with dace from five other isolated Great Basin populations in Oregon. They used mtDNA and nuclear DNA (nuclear s7 intron) and found that dace in the Warner Basin were different, potentially at the species level, from dace in the other five basins (Hoekzema and Sidlauskas 2014).

Recognizing the limitations of mtDNA for determining evolutionary lineages, Mussmann et al. (2020) compared isolated populations of Speckled Dace from throughout the Death Valley region, in the Owens and Amargosa river basins, using double-digest RAD. They found that the region has four distinct evolutionary lineages of *R. osculus* with each of the lineages being recognizable as a Distinct Population Segment (DPS) for management purposes.

2. Methods

2.1 Sampling and DNA sequencing

We obtained samples from 38 locations across several major zoogeographic regions throughout the range of Speckled Dace (Figure 1-1 and Supplemental Table S.1-1). Samples from Butte Lake in Lassen Volcanic National Park were included to determine if the population is native or introduced. Fin clips were taken from live adults or from the whole fish stored in ethanol, and the fin clips were dried on Whatman qualitative filter paper and stored at room temperature. DNA was extracted from fin clips with a magnetic bead–based protocol (Ali et al. 2016) quantified using Quant-iT PicoGreen dsDNA Reagent (Thermo Fisher Scientific) with an FLx800 Fluorescence Reader (BioTek Instruments). Genomic DNA was used to generate *SbfI* RAD libraries (Ali et al. 2016) and sequenced with paired-end 100-bp reads on an Illumina HiSeq 2500. Demultiplexing was performed requiring an exact match with well and plate barcodes (Ali et al. 2016). Sequencing coverage was assessed at the 50 bp position of each *de novo* RAD contig (see below) across all individuals using the depth function in SAMtools (Li et al. 2009).

2.2 RAD De Novo Assembly and Alignments

To generate a reference sequence for Speckled Dace, we performed RAD *de novo* assembly on eight individuals from the Walker River (Supplemental Material S.1). Specific details of the *de novo* assembly methods may be found in Baumsteiger et al. (2017), but briefly, a bioinformatic pipeline including a genome assembler was used to construct a partial reference for Speckled Dace. After *de novo* assembly, the mem algorithm in the Burrows–Wheeler aligner (BWA) was used to align each sample to the reference under the default parameters. SAMtools

was used to convert SAM files to BAM files, calculate the percentage of aligned reads, remove PCR duplicates, filter for the proper pairs, and combine the alignments if needed (Li et al. 2009). After this process, we removed low-coverage individuals with less than 70,000 mapped reads.

2.3 Genetic population structure

2.3.1 PCA

To begin investigating population structure, we used Analysis of Next Generation Sequencing Data (ANGSD 1.9) to identify SNPs (-SNP_eval 1e-12), calculate genotype likelihoods using SAMtools model (-GL 1), infer major and minor alleles directly from the genotype likelihoods (-doMajorMinor 1), and estimate allele frequencies assuming a fixed major allele with unknown minor allele (-doMaf 2) (Korneliussen et al. 2014). Only reads with a mapping quality score above 20 (-minMapQ 20) and only bases with a quality score above 20 (-minQ 20) were used in this process. Furthermore, only SNPs with a minor allele frequency of at least 0.01 (-minMaf) and that were represented in at least 50% of the included samples (-minInd 88) were included. These SNPs were then used to calculate a covariance matrix (-doCov 1), which was used to generate eigenvalues and eigenvectors for Principal Component Analysis (PCA). The percentage of total genetic variation explained by each PC was calculated, and PCs explaining a relatively large proportion of genetic variation were plotted with ggplot2 (Wickham 2009). To view the substructure within groups from the initial PCA, subsequent PCAs were performed on samples from each group using the same methods described above.

2.3.2 Admixture

To further assess population structure in Speckled Dace, we used the same parameters as above to generate a beagle file with ANGSD 1.9 for admixture analysis. The beagle output file was then used as the input file for NGSadmix (Skotte et al. 2013). The parameter K (the number

of clusters NGSAdmix partitions samples into), was tested from 2 to 9, and each run had a minor allele frequency filter of 0.01. After population structure was initially characterized, we repeated the procedure as described above on subsets of samples in order to determine substructure within each group.

2.3.3 Pairwise F_{ST}

To quantify the genetic divergence among populations, we calculated genome-wide pairwise F_{ST} values for population units identified by the analysis above and/or the sample collection locations. The folded site allele frequencies (SAF) were estimated for each group with RealSFS in ANGSD 1.9. The SAF files for the pairwise locations were the input to estimate two-dimensional site frequency spectrum (SFS). SFS files were then indexed by the SAF files to generate F_{ST} files. Then we estimated the weighted genome-wide F_{ST} values from the F_{ST} files with the stats function set in RealSFS.

2.4 Molecular Phylogeny

To further investigate the relationships among different genetic lineages, a range-wide phylogenetic tree was generated using SVDQuartets (Chifman and Kubatko 2014; Chifman and Kubatko 2015) and IQ-TREE 1.6.12 (Schrempf et al. 2019). Relict Dace (*Relictus solitarius*) and Tui Chub (*Siphatales bicolor*) were used as outgroups to root molecular phylogenies (Supplemental Material S.2). For SVDQuartets, samples were pooled by the locations where they were collected (Sub-region 1 of Supplemental Table S.1-1). If a significant genetic difference was shown between locations within a geographic region in PC or admixture analyses, the region was separated into two tips (Sub-region 2 of Supplemental Table S.1-1) accordingly.

We used ANGSD 1.9 to perform genotype calling, and we used the same parameters as mentioned above except we generated a VCF file (-dovcf 1). BCFtools 1.13 was used to prune

the SNPs with r^2 greater than 0.9 within each RAD contig (https://samtools.github.io/bcftools/bcftools.html). The pruned VCF file was transformed into NEXUS format by vcf2phylip (https://github.com/edgardomortiz/vcf2phylip). The pruned NEXUS file was analyzed by SVDQuartets within PAUP* 4.0 (Swofford 2002). We used a multispecies coalescent model to construct the phylogeny with 1,000,000 random quartets and 100 bootstraps.

We used the same SNP alignment with IQTREE as we did with SVDQuartets. The optimal substitution model was determined with ModelFinder under the Bayesian information criterion (-m MFP selected TVM+F+R4) (Kalyaanamoorthy et al. 2017) and 1,000 bootstraps were performed with ultrafast bootstrap approximation (-bb 1000) (Hoang et al. 2017). The resulting consensus tree was visualized with ggtree (Yu et al. 2017). Individuals with a substantial degree of missing data (<10,000 contigs with mapped reads) were pruned from the consensus tree for presentation.

3. Results

3.1 Sequencing, de novo RAD assembly, and alignment

Demultiplexed sequence data is available on from the NCBI Sequence Read Archive under BioProject PRJNA851170and code for analyses is available at *https://github.com/yingxins/speckled-dace/tree/Script*. The final assembly contained 17,639 contigs, with a mean contig length of 456.20, a maximum length of 788, and a minimum length of 89 (Supplemental Material S.3). After filtering individuals with sequencing and mapping quality, there were 175 individuals and 421,929 SNPs for range-wide analysis. The mean individual coverage (i.e., the average coverage across all the contigs in one individual) was 7.69, with a maximum 24.88, a minimum of 2.50, and a standard deviation of 3.92 (Supplemental

Figure S.1-1). For further analyses, there were 76 individuals and 108,746 SNPs for Group One, 67 individuals and 196,412 SNPs for Group Two, and 32 individuals and 142,578 SNPs for Group Three.

3.2 Genetic structure

3.2.1 Range-Wide

Across all Speckled Dace samples range-wide, the first two PCs explained 16.7% of the total variance and divided our samples into three clusters (Figure 1-2A). Because the percentage of genetic variation explained by PCA is highly influenced by the minor allele frequency (MAF) cutoff (we selected 0.01 as a cutoff) and our samples cover several distinct lineages, 16.7% genetic variation is a considerable amount of genetic variation to be explained. Group One (upper right) consists of populations of Speckled Dace from Sacramento River, Central California Coast, Klamath River, Warner Basin and Butte Lake. Group Two (upper left) is made up of the Speckled Dace from Amargosa River, Long Valley, Owens River Basin, and Lahontan Basin. Group Three (lower middle) is composed of populations from the Santa Ana River as well as locations outside California such as the Bonneville Basin, Washington Coast, Columbia River, and Lower Colorado River. Speckled Dace from the four regions outside California showed that the California populations we sampled are distinct from populations in the rest of the range of Speckled Dace.

We next used an admixture analysis of range-wide samples to complement our PCA. The admixture analysis was run with K = 2-9 (Supplemental Figure S.1-3). At K=3, members of each group in admixture analysis comprised Group One, Group Two and Group Three as indicated by PCA (Figure 1-2B). Furthermore, pairwise F_{ST} calculated between different populations varied from 0.16 (Speckled Dace from Owens River and Amargosa River) to 0.68 (Speckled Dace from

Amargosa River and Santa Ana River) (Supplemental Table S.1-2). Taken together, these results revealed that the Speckled Dace groups in California have highly variable levels of genetic divergence, and taxonomic revision may be warranted.

Our SVDQuartets range-wide phylogenetic analyses indicated that the Speckled Dace in California are mainly distributed into two monophyletic groups with the exception of Santa Ana Speckled Dace, which is a distinct evolutionary lineage (Figure 1-2C). Similar to the results of PCA and admixture analyses, Speckled Dace from the Sacramento River, Central California Coast, Klamath River, Warner Basin, and Butte Lake belong to the same monophyletic group (Group One, bootstrap support = 100%). Lahontan, Long Valley, Amargosa, and Owens Speckled Dace are another monophyletic group (Group Two, bootstrap support = 100%), which is the sister group of Group One. Speckled Dace collected from the Santa Ana River are the sister lineage of Speckled Dace from the lower Colorado River drainages (bootstrap support = 100%). Speckled Dace from the Santa Ana River and lower Colorado River are the earliest-branching lineage in the species-tree, followed by subsequent branching of (a) Speckled Dace from the Washington Coast and the Columbia River, (b) Speckled Dace from the Snake River and Bonneville, and (c) all other California Speckled Dace.

The phylogeny generated by IQTREE indicates three main lineages of California Speckled Dace (Supplemental Figure 1-S.4): (1) Group One as previously defined, and, (2) Group Two as previously defined are distinct lineages that are sister to each other; (3) Speckled Dace from the Santa Ana River are sister to Speckled Dace from the lower Colorado River drainage (placement and monophyly bootstrap support = 100%), and this combined lineage is sister to Group One and Group Two (bootstrap support = 100%). Speckled Dace from the Columbia River and Washington Coast combined are the earliest-branching lineage in this

phylogeny (placement and monophyly bootstrap support = 100%), and sister to all other Speckled Dace. Individuals sequenced from Snake River (n=4) and Bonneville (n=2) were filtered out due to low number of contigs with aligned reads. Overall, the genetic structure/divergence of Speckled Dace in California is hierarchical with multiple levels of genetically distinct lineages (Figure 1-2).

3.2.2 Group One: Klamath River, Central California Coast, Sacramento River, and Warner Basin

Group One Speckled Dace include Speckled Dace collected from Klamath and Sacramento rivers and Central California Coast plus Speckled Dace collected from Butte Lake and Warner Basin. After our range-wide data showed that Group One is distinct, we performed additional PC and admixture analyses using only Group One samples. For this PCA, the first three PCs explain the largest proportion of the genetic variation (Supplemental Figure S.1-2B). PC1 splits the Warner Basin population from populations in the other regions. PC2 separates the Klamath River population from Sacramento River populations (Figure 1-3A). PC3 separates the Central California Coast populations from populations in the Sacramento Basin (Pit River, Goose Lake, and Sacramento River) (Figure 1-3B). The Butte Lake population clusters with Speckled Dace from Sacramento River in all the PCs, indicating genetic similarity. Admixture analysis supports the PCA (Figure 1-3C): all the locations are separated in different K values. More specifically, admixture analysis splits out the Warner Basin and Central California Coast population when K = 2; the Klamath River and Sacramento River populations are distinct when K = 3. At K=4, the Butte Lake population is separated from the Sacramento River population. Pairwise F_{ST} analysis also supports the results from PC and admixture analyses; the highest F_{ST} values are found between Warner Basin and the other locations (mean: 0.32) and between the

Central California Coast and the other locations (mean: 0.28) (Supplemental Table S.1-2), whereas the F_{ST} value between the Klamath and Sacramento River populations is 0.18. The F_{ST} value between Sacramento River and Butte Lake populations is only 0.084. The Central California Coast population has relatively low pairwise F_{ST} values with Sacramento River and Butte Lake, which are 0.19 and 0.23 respectively.

The range-wide SVDQuartets analysis (Figure 1-2C) and the IQTREE phylogeny (Supplemental Figure S.1-4) are concordant with the above, placing Speckled Dace from the Klamath River, Central California Coast, Sacramento River, and Warner Basin into one clade (bootstrap support =100%). The Warner Basin population diverges first within the Group One clade in both the SVDQuartets species-tree and the IQTREE phylogeny. Subsequent to the branching of the Speckled Dace, both with the SVDQuartets species-tree and IQTREE phylogeny, Speckled Dace from the Klamath population separate. A well-supported clade of Speckled Dace representing the Klamath population is present in the IQTREE phylogeny. The remaining sampling locations - Butte Lake, Central California Coast, and Sacramento River - are not monophyletic with regard to sampling location when analyzed in a concatenated phylogenetic framework. Instead, these four geographic regions are mixed together into a single well-supported Group One subclade, which may be caused by lower genetic differentiation (Supplemental Figure S.1-4).

3.2.3 Group Two: Death Valley and Lahontan Basin

Speckled Dace in Group Two includes samples from three locations in the Death Valley region (Amargosa River, Owens River, and Long Valley) and four in the Lahontan Basin. To investigate the genetic structure within Group Two, we performed PC and admixture analyses on these samples. The first two PCs explain the largest proportion of the genetic variation

(Supplemental Figure S.1-2C): Amargosa River and Owens River populations are very close to each other in both PCs; both PC1 and PC2 split Lahontan Basin and Long Valley populations from Owens River and Amargosa River populations (Figure 1-4A). Admixture analysis supports the results of the PC analysis. The Lahontan Basin population is split from all the other Speckled Dace when K =2, and the Long Valley population is split from populations from the Owens River and Amargosa River populations when K = 3. At K=4 and K=5, we observed the local substructure in Amargosa River population, which is not discussed in this paper (Figure 1-4B) (but see Mussmann et al. 2020). Although not as obvious as in Group One, F_{ST} results support the PC and admixture analyses. The F_{ST} value between Speckled Dace collected from the Owens and Amargosa rivers is 0.16, which is the lowest of all Group 2 pairwise F_{ST} values. This is consistent with their close distance in PC analysis and differentiation at higher K values using admixture analyses. The F_{ST} values between Long Valley-Owens River and between Long Valley-Amargosa River are 0.38 and 0.30, respectively, which is concordant with their separation in the PC analyses and early split in admixture analyses (Supplemental Table S.1-2). However, though the Lahontan Basin population is the first lineage to separate in the admixture analysis, it does not have the highest pairwise F_{ST} values; the F_{ST} value between Amargosa River and Lahontan Basin is 0.33, which is higher than F_{ST} values for Lahontan-Owens (0.25) and Lahontan-Long Valley (0.26). These contradictory findings may be the result of multiple evolutionary events (e.g. hybridization between taxa or genetic drift in a dynamic landscape).

The range-wide SVDQuartets analysis is concordant with PC and admixture analyses in Group Two. The Lahontan Basin population, which split at K = 2, is the sister group of all the Death Valley populations. The result of IQTREE is also similar to the result of SVDQuartets, but exhibits structuring between Martis Creek, Humboldt River and Walker River sampling locations

from the Lahontan Basin (Supplemental Figure S.1-4). The Owens and Amargosa rivers, which show little genetic divergence in the PC and admixture analyses, are sister lineages in the SVDQuartets species-tree and form a monophyletic clade in the IQTREE phylogeny. The position of the Speckled Dace from Long Valley in the SVDQuartets species-tree and IQTREE phylogeny as sister to Owens River + Amargosa samples is also supported by admixture analysis, where Long Valley splits after Lahontan Basin but before Amargosa and Owens rivers (bootstrap support = 100%). Although PC analyses and pairwise F_{ST} indicate that the Long Valley population is a separate lineage from the Amargosa River and Owens River populations, this incongruence could be caused by overestimation of genetic divergence caused by genetic drift in a small population under long isolation.

3.2.4 Group Three: Santa Ana River

The only California Speckled Dace lineage in Group Three is that of Speckled Dace from the Santa Ana River, which clusters with non-California Speckled Dace (Supplemental Table S.1-1). To investigate the distinctiveness of the Santa Ana River population, we performed PC analysis, admixture analysis, and estimated pairwise F_{ST} for sample collections in Group Three. PC and admixture analyses show that Santa Ana River population is strikingly genetically different from non-California Speckled Dace in the Group Three. In the PC analysis for Group Three, the largest proportion of the genetic variation is explained by PC1 and PC2 (Supplemental Figure S.1-2D). Both PC1 and PC2 split the Santa Ana River population from all other Speckled Dace lineages. Admixture analysis for Group Three was run from K = 2 to K = 5, and the Santa Ana River population split from the other locations (Lower Colorado Basin, Bonneville, Columbia Basin, and Washington) coast from K = 3 to K = 5 (Figure 1-5B). In addition, the

Santa Ana River population has high pairwise F_{ST} values with all other California Speckled Dace and non-California Speckled Dace (Supplemental Table S.1-2B).

The range-wide SVDQuartets analysis places the Santa Ana population sister to samples collected from lower Colorado Basin (bootstrap support = 100%). IQTREE also places the Santa Ana population and lower Colorado Basin as sister lineages (bootstrap support = 100%), and indicates that each of these two populations are monophyletic as well (bootstrap support = 100%). The results of admixture analysis and PCA support the genetic affinity between Santa Ana Speckled Dace and those from the Colorado Basin: the lower Colorado population clusters with Santa Ana River population when K = 2 and the lower Colorado population is the closest lineage to the Santa Ana River population in the PCA. However, due to the limited number of samples of non-California Speckled Dace, we did not further explore the relationship between non-California Speckled Dace and Speckled Dace from Santa Ana River.

4. Discussion

4.1 The Speckled Dace Has Multiple Lineages.

Our genomic data analyses show that the Speckled Dace has hierarchical, genetically distinct lineages. In other words, they are genetically divergent at different levels as opposed to having relatively uniform relatedness as might be expected for a single widespread population. Our genetic analysis of California populations divides them into three lineages with sub-lineages. These lineages and sub-lineages coincide with zoogeographic regions that are largely isolated from one another and that contain other endemic fishes, suggesting long isolation (Moyle 2002). If allopatry sustains the genetic divergence for a sufficient duration, presumably phenotypic and genotypic differences will accumulate. It is thus unlikely that the split lineages can merge into a single lineage again and unlikely that genetic differences will be lost to hybridization upon

secondary contact. Thus, we assume that hybrid individuals from two distinct lineages would be poorly adapted to whatever ecological system in which they occur and have reduced fitness (Coyne and Orr 2004). We therefore find it appropriate to label geographically isolated lineages with large genomic differences as distinct lineages and geographically isolated lineages with less genomic differentiation as sub-lineages (Freudenstein et al. 2016). These lineages and sub-lineages will be a precondition of the formal species delimitation of taxa within SDC. A more comprehensive definition of species in the SDC, will be presented in a separate paper that formally describes species and subspecies in California.

4.1.1 Speckled Dace from Klamath River, Sacramento River, Central California Coast, and Warner Basin are a Single Genetically Distinct Lineage.

In all the analyses, Speckled Dace from the Klamath River, Sacramento River, Central California Coast and Warner Basin are a monophyletic lineage (Group One). These dace have relatively low F_{ST} values within the lineages compared to F_{ST} values between them and other populations. For example, F_{ST} values between Sacramento River and Klamath River populations and the Warner Basin population are 0.17 and 0.29, but the F_{ST} values between the Sacramento River population and Speckled Dace from Death Valley (Amargosa River, Owens River, Long Valley) are 0.56, 0.51, 0.44 respectively. The Klamath River and Sacramento River populations have less genetic divergence from each other than either does from the Warner Basin population, but the geographical basins in which each occurs are all well-defined and contain other endemic fishes. This means the isolation of these three basins has been long enough for diversification,

although the Sacramento and Klamath river lineages have closer ties to each other than either does to the Warner Basin. The Klamath River has not drained south into the ancestral Sacramento River system since the end of the Pliocene (*c.a.* 3 mya); however, extensive deformation (e.g. downfaulting of the Klamath Graben) and volcanism (both from the Cascade-arc and Medicine Lake volcanic field) have occurred almost continuously in the northern Sacramento and southern Klamath Basins (Colman et al. 2004). It is probable that this volcanism led to repeated drainage captures of headwater streams (habitat for dace and other native fishes with a common genetic heritage between the two basins) and allowed for intermittent gene flow. This intermittent inter-basin connectivity was presumably less frequent between the Klamath/Sacramento streams and those in the Warner Basin. The onset of Great Basin faulting and extension of the transnational Walker Lane belt into Surprise Valley resulted in uplift of the Warner Range (*c.a.* 3 mya) which likely resulted in a permanent topographic division of the Basins sometime in the late-Pleistocene (1.0-0.1 mya) (Egger and Miller 2011).

The large geographic extent of the Klamath and Sacramento drainages has resulted in some geographic population structure in each basin, creating genetically distinct population segments that need further investigation for determination of taxonomic status (Oakey et al. 2004). For example, Speckled Dace from the Central California Coast (San Luis Obispo Creek, Santa Maria River, Monterey Bay drainages) show enough genetic differentiation that this group of populations could be recognized as a genetically distinct sub-lineage within dace populations from the Sacramento drainage.

4.1.2 Speckled Dace from Butte Lake is an Introduced Population.

Butte Lake is located in Lassen Volcanic National Park and historically drained into the Lahontan basin, so Speckled Dace from Butte Lake were assumed to be genetically related to the

Speckled Dace in Lahontan Basin. However, Speckled Dace from Butte Lake have much greater similarity to Speckled Dace from the and Central California Coast and the Sacramento River than to Lahontan Speckled Dace in all the analyses. Therefore, we classify Speckled Dace from Butte Lake as part of the Central California Coast or Sacramento River population and hypothesize that the population most likely represents a bait-bucket introduction. Butte Lake drains northward from Mount Lassen through Butte Creek (which also has Speckled Dace) and may have been connected at one time to the Eagle Lake watershed in the Lahontan Basin, although frequent lava flows have obscured drainage patterns. The three other fishes present in Butte Lake, Tahoe Sucker (*Catostomus tahoensis)*, Lahontan Redside (*Richardsonius egregius)*, and Tui Chub (*Siphatales bicolor*) are Lahontan basin fishes, lending credence to the bait bucket hypothesis.

4.1.3 Speckled Dace from Death Valley and the Lahontan Basin are a single lineage.

In range-wide PCA, admixture analysis, and the phylogenetic analyses, Speckled Dace from Death Valley (Owens River, Amargosa River, Long Valley) and the Lahontan Basin are most closely related to each other (Group Two). The two geographic regions – Death Valley and Lahontan Basin – are isolated geological basins, and Speckled Dace reflect this division in analyses of Group Two (e.g. Figure 1-4B). Similar to Speckled Dace from the Sacramento and Klamath rivers, we also consider Speckled Dace from Death Valley and Lahontan Basin a single lineage with two distinct sub-lineages with a common ancestor.

Within the Death Valley lineages, the Amargosa River and Owens River populations only show small genetic differences, a finding consistent with Mussmann et al. (2020). Smith et al. (2017) found that Speckled Dace from the Amargosa River shared haplotypes with dace from

the Owens Valley. Dace from Oasis Valley, Nevada, the headwaters of the Amargosa River, and from Ash Meadows (Bradford Spring), are sister lineages of Speckled Dace from Owens River. Unlike the situation for Speckled Dace from Klamath and Sacramento rivers, the Owens River and Amargosa River watersheds are internal drainages that were connected via a chain of large lakes during extended wet periods in the late Pleistocene. Given the results of our analyses and their recent geographic separation and isolation, we place Speckled Dace from the Amargosa and Owens rivers as one sub-lineage. Although Speckled Dace from Long Valley are in the same watershed as the Owens Valley, dace from Long Valley are genetically distinct from dace in the Owens and Amargosa rivers. Mussmann et al. (2020) observed a similar pattern in the phylogeny and F_{ST} estimates. This is probably the result of genetic drift due to isolation of small dace populations in small streams flowing into the Long Valley caldera. Climatic shifts and volcanism in the southern Mono Lake Basin subsequently isolated the Owens Basin from the Lahontan Basin. Here we treat the Long Valley population as a sub-lineage under Lahontan Basin Speckled Dace.

Speckled Dace from the Walker River, Humboldt River, Eastern Sierra Nevada streams, and Death Valley system streams are one lineage: the Lahontan Speckled Dace *R. o. robustus*, which is a widely recognized taxon (Deacon and Williams 1984; Rutter 1902; Moyle 2002). Although Lahontan Basin Speckled Dace split at K=2 in the admixture analysis for Group Two, F_{ST} values between Lahontan-Owens and Lahontan-Long Valley Speckled Dace are somewhat small: 0.25 and 0.26, respectively, and even lower than the F_{ST} between Long Valley-Owens Speckled Dace.

Long Valley Speckled Dace originally occurred in a series of Hot Spring marshes in the Long Valley Caldera, located in the headwaters of Owens Valley watershed. Upstream

movement of Speckled Dace is prevented by a series of large drops in the Owens River Gorge. The Long Valley caldera drains into the Owens River before it enters this canyon as it flows from Long Valley to the Owens Valley. Downfaults of the Owens Valley occurred approximately 1.5 mya (Hildreth and Fierstein 2016). Low F_{ST} values suggest that 1) either a hybridization event took place between Speckled Dace from the Lahontan Basin and Owens River, creating Long Valley Speckled Dace, or 2) Lahontan and Long Valley share an early evolutionary history and were later separated by geologic change. The presence of Owens Tui Chub (*Siphatales bicolor snyderi*) and Owens Sucker (*Catostomus fumeiventris*) in Long Valley support the first hypothesis because the closest relatives of both taxa are in the Lahontan basin (Moyle 2002). Therefore, we consider Lahontan Speckled Dace to have two genetic distinct lineages in the Death Valley region: Speckled Dace in Death Valley that includes Amargosa and Owens River systems and Speckled Dace in Long Valley. However, the presence of other *R. osculus* subspecies, some described, in the Lahontan Basin indicates that additional subspecies will likely eventually be added to the list (Deacon and Williams 1984).

4.1.4 Speckled Dace from the Santa Ana River are a Distinct Lineage

Our range-wide analyses revealed that Speckled Dace from the Santa Ana River Basin are strikingly different from all other populations of Speckled Dace in California (Group Three). Speckled Dace in the Santa Ana River share more genetic similarities with Speckled Dace from the lower Colorado Basin, Bonneville, Washington Coast, and the Columbia River than with other dace lineages in California. Due to the small number of samples, the genetic diversity within non-California basins is not discussed in this paper. The evolutionary history of Speckled Dace from the Santa Ana River can be linked most closely with Speckled Dace from the lower Colorado Basin because they did not split from each other in the admixture analysis with all the

samples from K = 3 to K = 8 (Supplemental Figure S.1-3). According to Smith et al. 2017, Speckled Dace collected from Colorado Basin and Speckled Dace collected from Santa Ana Basin are sister lineages in the Colorado Group with relatively weak bootstrapping support in the mtDNA phylogeny. Regarding pairwise F_{ST} values, we find that Santa Ana Basin Speckled Dace have high genetic divergence from both California and non-California Speckled Dace. The lower bootstrapping in Smith et al. (2017) is likely caused by high genetic divergence and relatively low diversity in mtDNA.

In our study, we clarify the genetic distinctness of Speckled Dace from the Santa Ana Basin. All analyses show that these dace have remarkably large genetic differences from other populations (Supplemental Table S.1-2). Due to their unique genetic structure, Santa Ana Speckled Dace are clearly a distinct lineage. This same basic conclusion was reached by Cornelius (1969) who conducted a detailed study of the morphometrics and meristics of Santa Ana Basin Speckled Dace, as well as of dace from neighboring streams (Sacramento Basin), the Virgin River (Lower Colorado basin), and Lake Tahoe (Lahontan basin). His study was the first to link the origins of Speckled Dace from the Santa Ana Basin to the lower Colorado River Basin. Details of how this connection occurred can be found in Axen and Fletcher (2010), McClay and Bonora (2001), and Dorsey and Langenheim (2015).

5. Conclusion and Conservation Implications

If we view the SDC as a single lineage, it is a species that does not merit special consideration for conservation because of its wide distribution and large population size. However, our genetic analyses show as that Speckled Dace in California have a hierarchical order of divergence and that the different levels of genetic distinctness divide California Speckled Dace into multiple lineages. More specifically, our genetic analyses place all California

populations into three major genetic lineages: 1) Speckled Dace from the Sacramento River, Central California Coast, Klamath River, and Warner Basin; 2) Speckled Dace from Death Valley, Long Valley, and Lahontan Basin, 3) Speckled Dace from the Santa Ana River. Each distinct population within the three lineages can represent a sub-lineage under the main lineage. Each of these lineages needs further study to locate populations within them that need special protection. Indeed, the naming a species can improve its protection.

In this study, the populations across geographical regions are genetically divergent at different levels, depending on time and degree of isolation from other populations of Speckled Dace. However, populations in different geographical regions face different environmental threats. For example, the Amargosa region in Death Valley is one of the hottest and driest places in North America, where fish depend on springs that draw ancient water from underground aquifers; pumping water from these aquifers threatens to deplete this small flow (Robbins 2017; Belcher et al. 2019). In the Death Valley region, the Ash Meadows Speckled Dace was listed as Endangered in 1984 which is probably the main reason it still exists. The Center for Biological Diversity CBD has filed a petition (2020) to have that status apply to all dace populations in the Death Valley region, based on Mussmann (2020).

In the Los Angeles region, the Santa Ana Speckled Dace persists through urbanization, which has eliminated much of their habitat throughout the Santa Ana, San Gabriel and Los Angeles River systems and altered much of what is left. The dace presently inhabits small-isolated streams and creeks in the Santa Ana and San Gabriel watersheds where they are vulnerable to the effects of fire, droughts and floods (Nunziata et al. 2013; SAWPA 2004). A petition to list the Santa Ana Speckled Dace as threatened under the federal Endangered Species Act in 1996 was rejected largely because of lack of a formal species description (USFWS 1996).

The genetic and evolutionary distinctiveness of Santa Ana Speckled should no longer be an issue, so listing is justified.

These taxonomic issues will be further explored in a paper devoted solely to the taxonomy of the SDC in California. We can now combine our knowledge of genetic divergence with that of ecosystem status and characteristics to design distinct conservation management and policy strategies for different populations of Speckled Dace.

The focus of this paper is Speckled Dace in California, so how our findings relate to Speckled Dace outside of California is not discussed. However, it seems likely that there are non-California lineages that can also be designated (or redesignated) as species or subspecies when genomic methods are applied with careful sampling. Further genomic research in other zoogeographic areas will undoubtedly reveal heretofore unrecognized genetic structure.

6. Acknowledgements

The authors thank Amber Manfree for making the maps (Figure 1-1). The following people provided fish samples for us: Scot Lucas (Klamath River), Mollie Ogaz (Sacramento River system), Nick Buckmaster and Steve Parmenter (Owens Valley, Lahontan Basin), Kevin Guadalupe and Gregory Munson (Virgin River, Oasis Valley, Amargosa River), Hal Fairfield (Death Valley region), Jennifer Pareti (Santa Ana River), and Claire Ingel (South California).

7. Reference

Ali, O. A., and coauthors. 2016. RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. Genetics 202(2):389-400.

Ardren, W., J. Baumsteiger, and C. Allen. 2010. Genetic analysis and uncertain taxonomic status of threatened Foskett Spring speckled dace. Conservation Genetics 11:1299-1315.

Axen, G., and J. Fletcher. 2010. Late Miocene-Pleistocene Extensional Faulting, Northern Gulf of California, Mexico and Salton Trough, California. International Geology Review 40:217-244.

Baumsteiger, J., P. B. Moyle, A. Aguilar, S. M. O'Rourke, and M. R. Miller. 2017. Genomics clarifies taxonomic boundaries in a difficult species complex. PloS one 12(12):e0189417-e0189417.

Belcher, W. R., D. S. Sweetkind, C. B. Hopkins, and M. E. Poff. 2019. Hydrogeology of Lower Amargosa Valley and groundwater discharge to the Amargosa Wild and Scenic River, Inyo and San Bernardino Counties, California, and adjacent areas in Nye and Clark Counties, Nevada, 2018-5151, Reston, VA.

Billman, E. J., and coauthors. 2010. Phylogenetic Divergence in a Desert Fish: Differentiation of Speckled Dace within the Bonneville, Lahontan, and Upper Snake River Basins. Western North American Naturalist 70(1):39-47.

Colman, S. M., J. Platt Bradbury, and J. G. Rosenbaum. 2004. Paleolimnology and paleoclimate studies in Upper Klamath Lake, Oregon. Journal of Paleolimnology 31(2):129-138.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA.

Dorsey, R. J., and V. E. Langenheim. 2015. Crustal-scale tilting of the central Salton block, southern California. Geosphere 11(5):1365-1383.

Egger, A. E., and E. L. Miller. 2011. Evolution of the northwestern margin of the Basin and Range: The geology and extensional history of the Warner Range and environs, northeastern California. Geosphere 7(3):756-773.

Evermann, B. W., and S. E. Meek. 1896. A report upon salmon investigations in the Columbia river basin and elsewhere on the Pacific coast in. Retrieved from the Library of Congress, Washington.

Freudenstein, J. V., M. B. Broe, R. A. Folk, and B. T. Sinn. 2016. Biodiversity and the Species Concept—Lineages are not Enough. Systematic Biology 66(4):644-656.

Hildreth, W., and J. Fierstein. 2016. Long Valley Caldera Lake and reincision of Owens River Gorge, 2016-5120, Reston, VA.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35(2):518-522.

Hoekzema, K., and B. L. Sidlauskas. 2014. Molecular phylogenetics and microsatellite analysis reveal cryptic species of speckled dace (Cyprinidae: Rhinichthys osculus) in Oregon's Great Basin. Mol Phylogenet Evol 77:238-50.

Jordan, D. S., and B. W. Evermann. 1896. The Fishes of North and Middle America, volume 47. US National Museum Bulletin

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14(6):587-589.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15(1):356.

Leidy, R., and P. Moyle. 2021. Keeping up with the status of freshwater fishes: A California (USA) perspective. Conservation Science and Practice.

Li, H., and coauthors. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

McClay, K., and M. Bonora. 2001. Analog Models of Restraining Stepovers in Strike-Slip Fault Systems. AAPG Bulletin 85(2):233-260.

Moyle, P. B. 2002. Inland fishes of California. University of California Press., Berkeley.

Mussmann, S. M., M. R. Douglas, D. D. Oakey, and M. E. Douglas. 2020. Defining relictual biodiversity: Conservation units in speckled dace (Leuciscidae: Rhinichthys osculus) of the Greater Death Valley ecosystem. Ecology and evolution 10(19):10798-10817.

Oakey, D. D., M. E. Douglas, and M. R. Douglas. 2004. Small Fish in a Large Landscape: Diversification of Rhinichthys osculus (Cyprinidae) in Western North America. Copeia 2004(2):207-221.

Robbins, J. 2017. The Amargosa River defies the desert (p. 1). New York Times, New York, NY.

SAWPA, S. A. R. W. P. A. 2004. Old, grand prix, and padua fires (October, 2003): burn impacts to water systems and resources. Santa Ana River Watershed Area. San Bernardino National Forest, California. Santa Ana Watershed Project Authority, Riverside, California.

Schrempf, D., B. Q. Minh, A. von Haeseler, and C. Kosiol. 2019. Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity. Molecular Biology and Evolution 36(6):1294-1301.

Skotte, L., T. S. Korneliussen, and A. Albrechtsen. 2013. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics 195(3):693-702.

Smith, G., J. Chow, P. Unmack, D. Markle, and T. Dowling. 2017. Evolution of the Rhinichthys osculus complex (Telostei: Cyprinidae) in Western North America. Miscellaneous Publications Museum of Zoology, University of Michigan 204:45.

Spencer, J. E., G. R. Smith, and T. E. Dowling. 2008. Middle to late Cenozoic geology, hydrography, and fish evolution in the American Southwest. Pages 0 *in* M. C. Reheis, R. Hershler, and D. M. Miller, editors. Late Cenozoic Drainage History of the Southwestern Great Basin and Lower Colorado River Region: Geologic and Biotic Perspectives, volume 439. Geological Society of America.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10, volume Version 4.0.

USFWS. 1996. Endangered and threatened wildlife and plants: 90-day finding on a petition to list the Santa Ana Speckled Dace, Santa Ana Sucker, and the Shay Creek threespine stickleback as endangered.

Wickham, H. 2009. Ggplot2: Elegant Graphics for Data Analysis. New York. Springer, New York:.

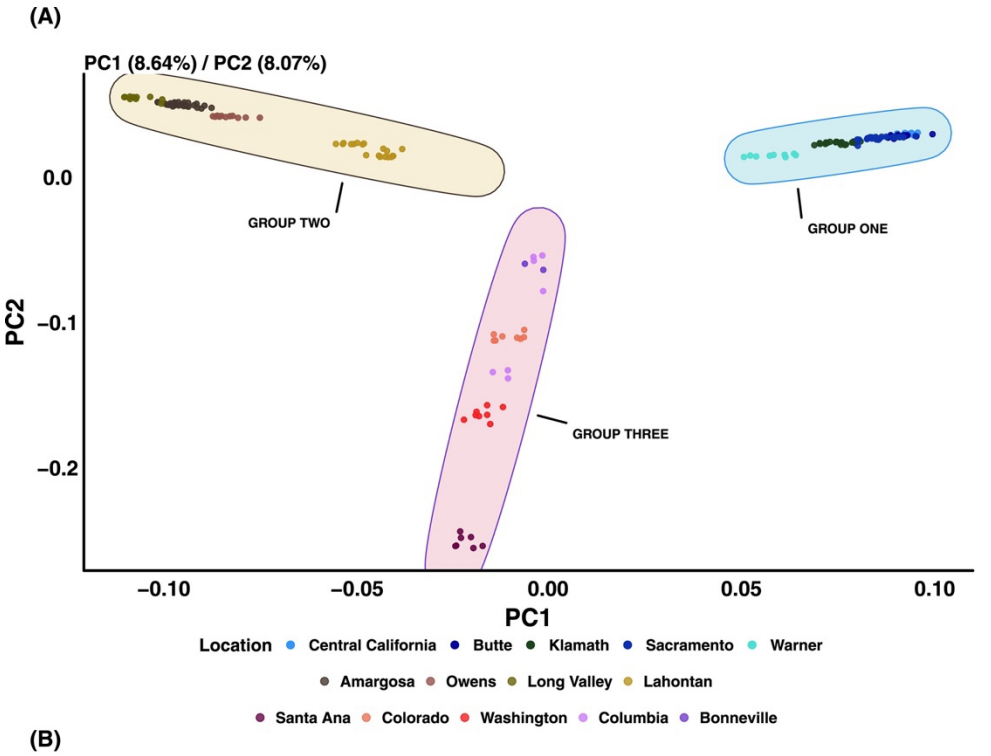
Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8(1):28-36.

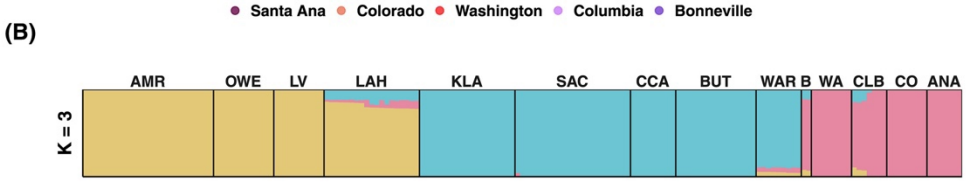
8. Figures and Tables

FIGURE 1-1. Sampling map. Map of sampling sites in which Speckled Dace were collected for this study. The location represented by each number and the number of individuals sampled are detailed in Supplemental Table S.1-1.


State
Drainage basin
Waterway
Pacific Ocean
Washington
Columbia R.
Willamette R.
Columbia
Salmon R.
Montana
Oregon
Idaho
Rogue R.
Snake R.
Wyoming
Klamath
Eel R.
Humboldt R.
Sacramento
Lahontan
Bonneville
San Joaquin
Nevada
Utah
Green R.
Colorado
California
San Juan R.
Colorado R.
Death Valley
Arizona
New Mexico
Gila R.
0 100 200 300 km
0 100 200 mi

FIGURE 1-2. Range-wide Speckled Dace population structure. (A) Principal Component Analysis (PC) of all samples. Color represents locations from which the Speckled Dace were collected. 16.7% genetic variation is explained in total (PC1 explains 8.6% variation while PC2 explains 8.1% variation). Three groups are distinguishable. Group One includes Speckled Dace from Sacramento River (SAC), Central California Coast (CCA), Klamath River (KLA), Butte Lake (BUT) and Warner Basin (WAR). Group Two includes Speckled Dace from Amargosa River (AMR), Owens River (OWE), Long Valley (LV), and Lahontan Basin (LAH). Group Three includes Speckled Dace from the Santa Ana River (ANA) and from Washington Coast (WA), Columbia River (CLB), Bonneville Basin (B), and Colorado River Basin (CO). (B) Admixture analysis of all samples when K = 3, which means we assumed the current populations were admixed by three populations in the past. (C) SVDQuartets results of all samples. Group One and Group Two are monophyletic and are the sister groups of each other, while Santa Ana Speckled Dace were clustered with Speckled Dace from Lower Colorado Basin and are the sister group of all other Speckled Dace included in this study.


(A)
PC1 (8.64%) / PC2 (8.07%)
GROUP TWO
GROUP ONE
GROUP THREE
PC2
0.0
−0.1
−0.2
−0.10
−0.05
0.00
0.05
0.10
PC1
Location Central California Butte Klamath Sacramento Warner
Amargosa Owens Long Valley Lahontan
Santa Ana Colorado Washington Columbia Bonneville


(B)
AMR OWE LV LAH KLA SAC CCA BUT WAR B WA CLB CO ANA
K = 3

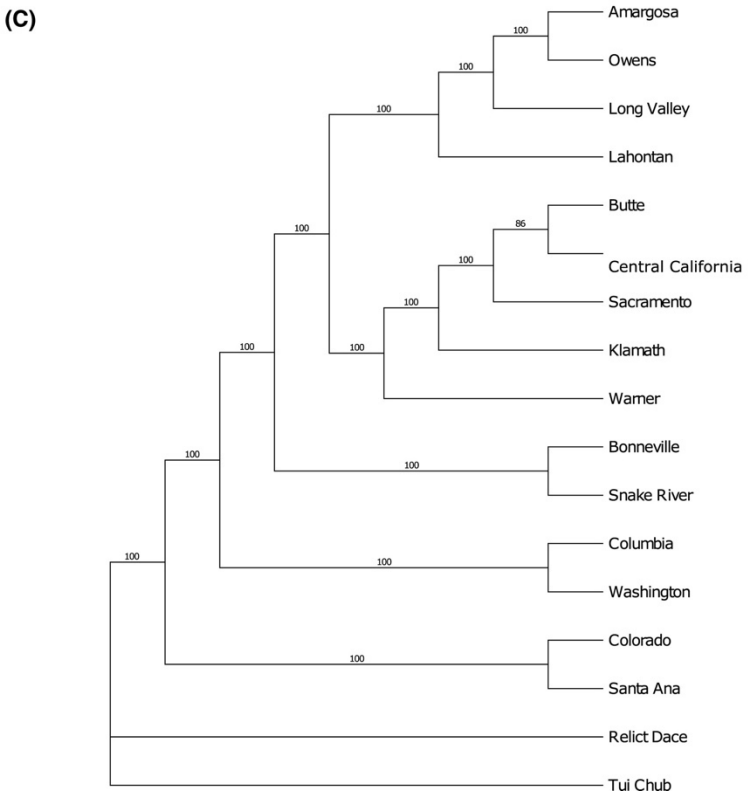

(C)
Amargosa
Owens
Long Valley
Lahontan
Butte
Central California
Sacramento
Klamath
Warner
Bonneville
Snake River
Columbia
Washington
Colorado
Santa Ana
Relict Dace
Tui Chub

FIGURE 1-3. Population structure of Speckled Dace from the Sacramento River, Central California Coast, Klamath River, and Warner Basin. (A) Principal Component Analysis of samples in Group One; color represents locations of the Speckled Dace. 12.3% of the genetic variation is explained by PC1(7.6%) and PC2 (4.7). (B) PC analysis of samples in Group One when genetic variation is explained by PC1 (7.6%) and PC3 (3.7%). (C) Admixture analysis of samples in Group One when K = 2, 3, 4, 5. The abbreviation labels represent locations as in Figure 1-2.

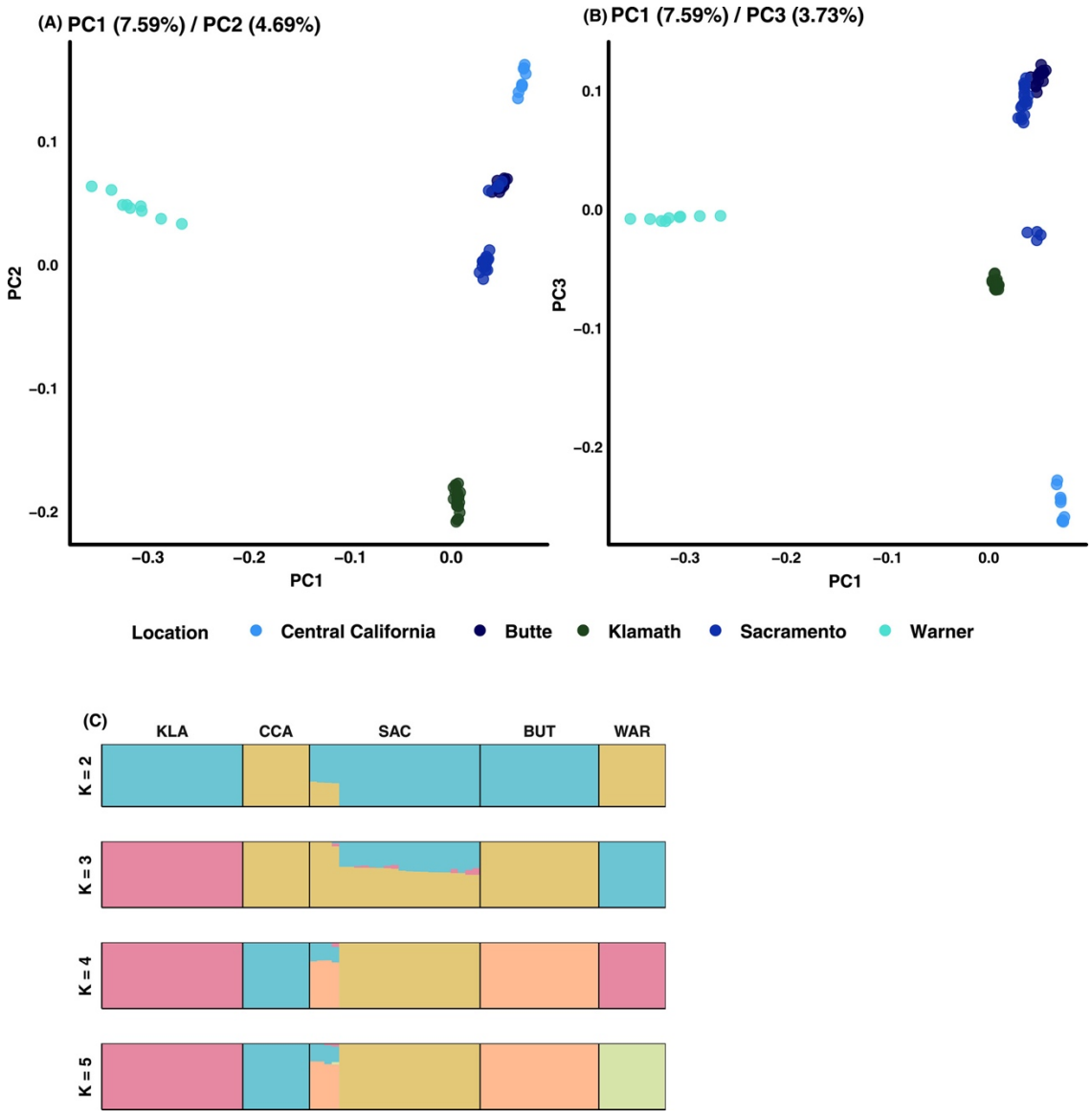


FIGURE 1-4. Death Valley Speckled Dace and Lahontan Speckled Dace population structure. (A) Principal Component Analysis of samples in Group Two; color represents the locations where Speckled Dace were collected. 20.13% of the genetic variation is explained by PC1(10.3%) and PC2 (9.8%). (B) Admixture analysis of samples in Group Two when K = 2, 3, 4, 5. The abbreviation labels represent the locations as in Figure 1-2.

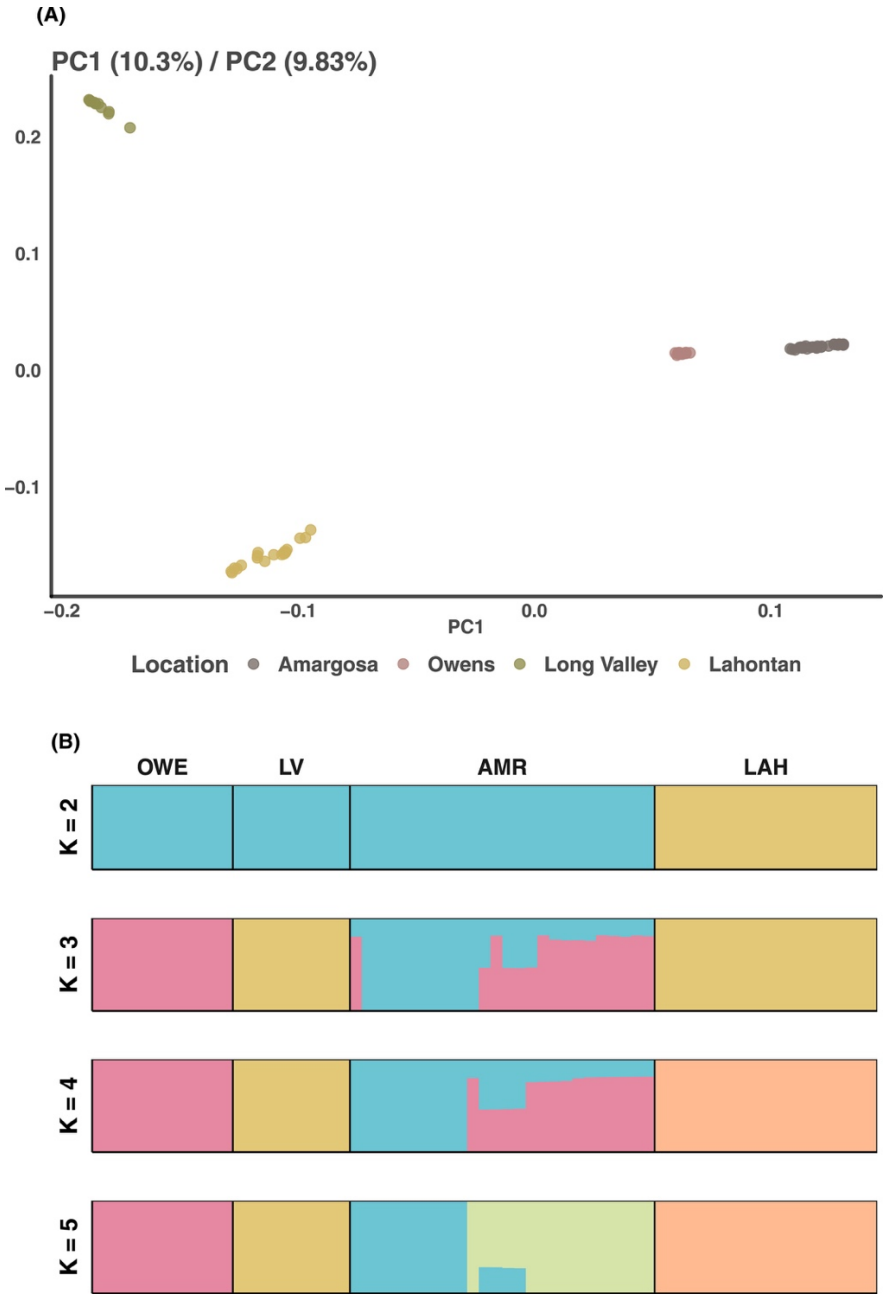
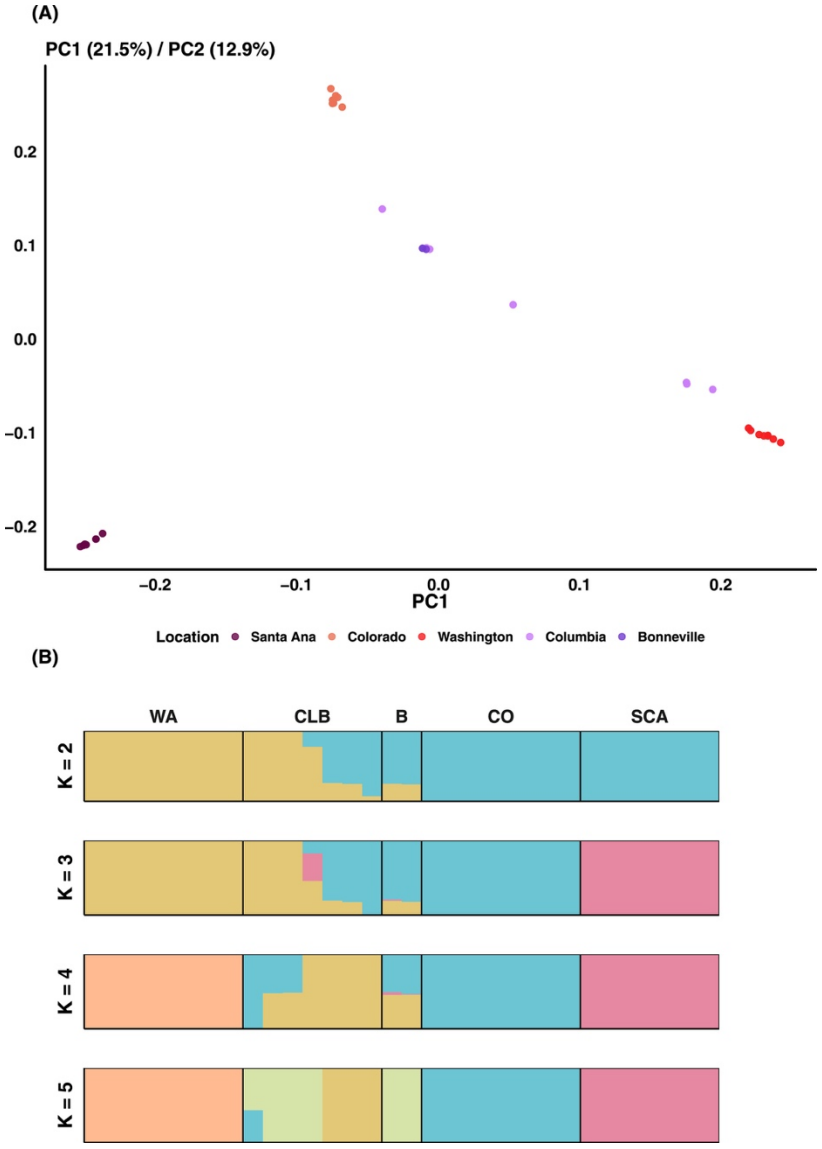
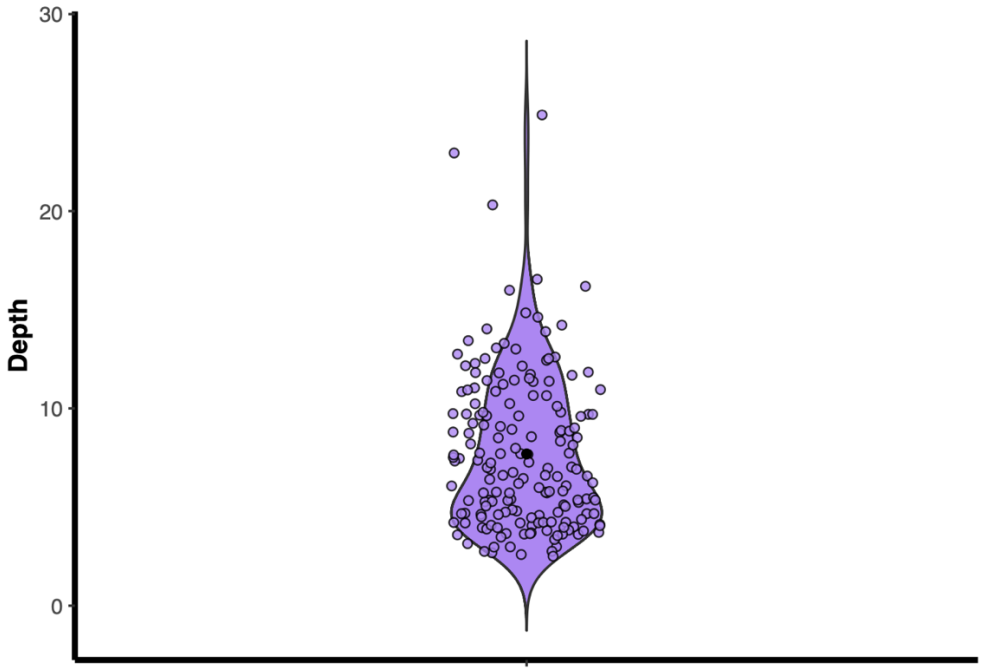


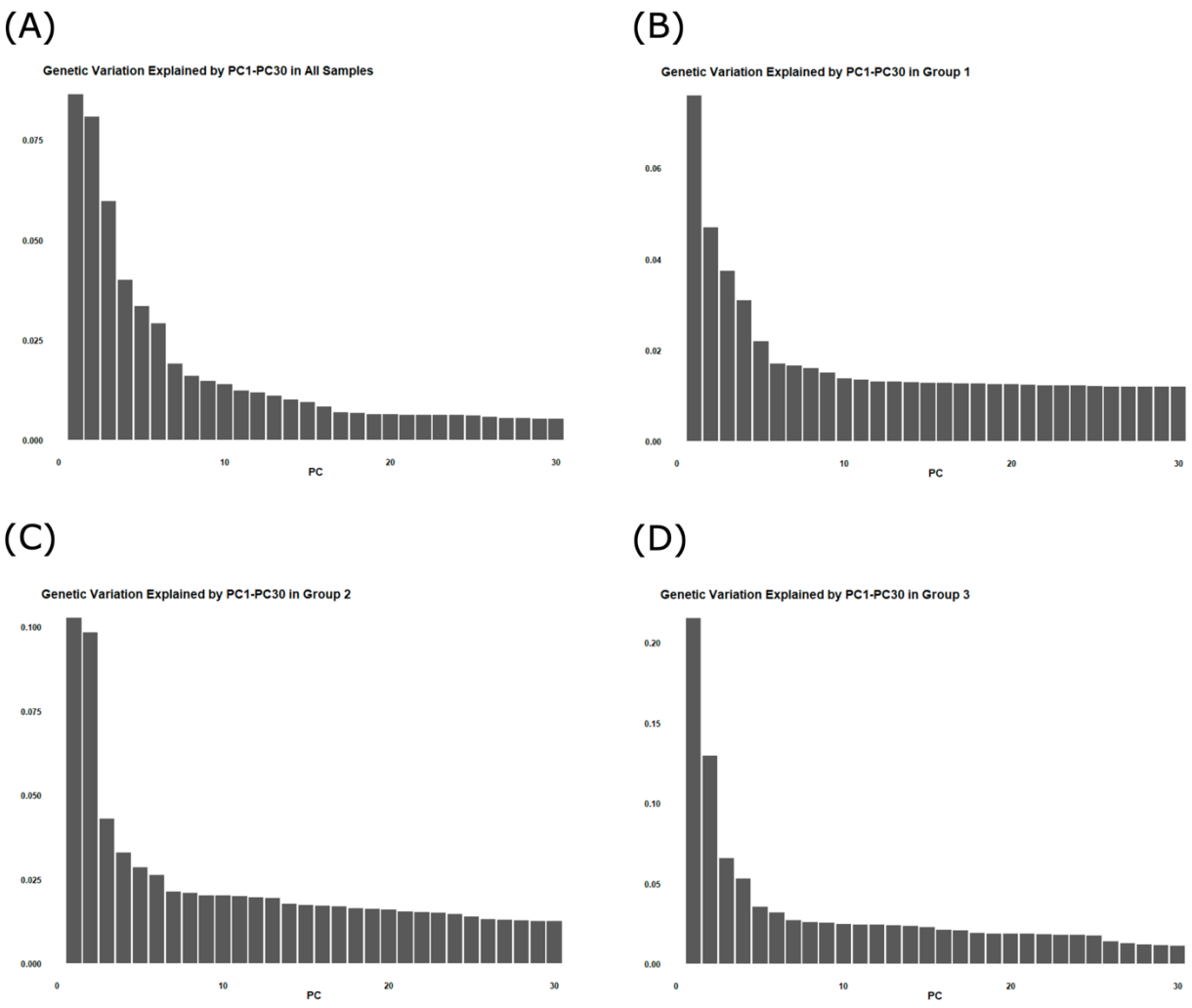
FIGURE 1-5. Non-California and Santa Ana Basin Speckled Dace population structure. (A) Principal Component Analysis of samples in Group Three. Color represents locations where Speckled Dace were collected. 34.4% genetic variation is explained in total (PC1 explains 21.5% variation while PC2 explains 12.9% variation). (B) Admixture analysis of samples in group three when K =2, 3, 4, 5. The abbreviations correspond to locations in Figure 1-2. Both PC and admixture analyses support Speckled Dace from the Santa Ana River as distinct from non-California Speckled Dace but have a distant relationship to Speckled Dace from the Lower Colorado Basin.



SUPPLEMENTAL FIGURE S.1-1. Contig depth. The distribution of mean depth of all the contigs at 50 bp in each individual.

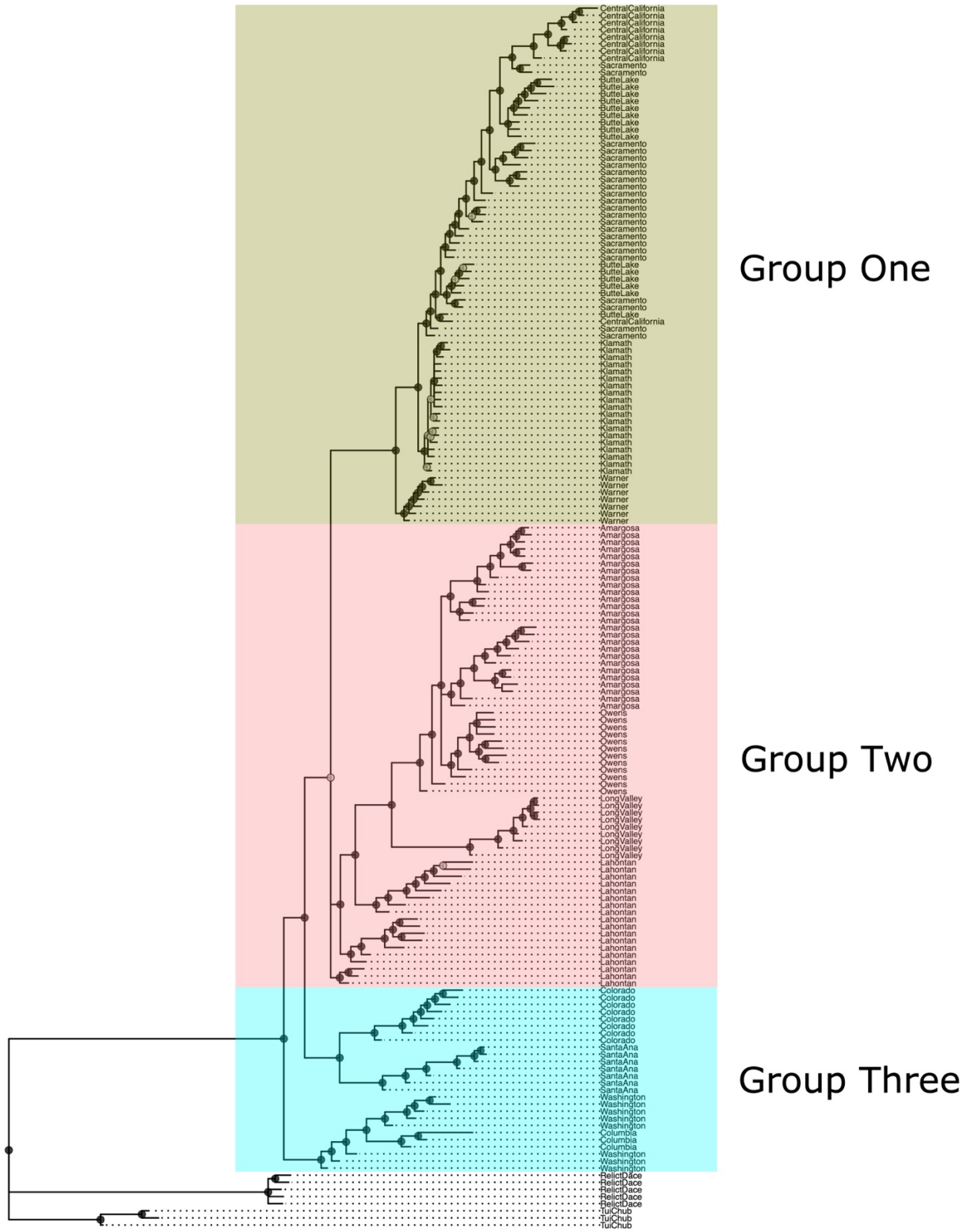


SUPPLEMENTAL FIGURE S.1-2. Genetic variation explained by each Principal Component. (A) The percentage of genetic variation explained by the first 30 PCs for range wide PCA. (B) The percentage of genetic variation explained by the first 30 PCs for PCA for California Speckled Dace, Warner Speckled Dace and Speckled Dace in Butte Lake (Group One). (C) The percentage of genetic variation explained by the first 30 PCs for PCA for Death Valley Speckled Dace and Lahontan Speckled Dace (Group Two). (D) The percentage of genetic variation explained by the first 30 PCs for PCA for Non-California and Santa Ana Speckled Dace (Group Three).



SUPPLEMENTAL FIGURE S.1-3. Admixture analysis of all the samples from K = 2-9. K refers to the number of the ancestral populations that the current populations are admixed from. Each color represents one of the ancestral populations. The abbreviations correspond to locations in Figure 1-2.

SUPPLEMENTAL FIGURE S.1-4. Maximum-Likelihood phylogeny generated by IQTREE. Tips are labeled according to Sub-region 1 following Supplemental Table S.1-1, except Butte Lake and Santa Ana, which are labeled by Sub-region 2, with the exception of the outgroup that is labeled according to species. The circle at each node indicates bootstrap support with black > 90% and gray > 75% but < 90%. Nodes with < 75% bootstrap support are presented as polytomies. The tree is rooted with Relict Dace and Tui Chub and groups of Speckled Dace as found by Principal Component Analysis (Figure 1-2) are indicated.


CentralCalifornia
CentralCalifornia
CentralCalifornia
CentralCalifornia
CentralCalifornia
CentralCalifornia
CentralCalifornia
CentralCalifornia
Sacramento
Sacramento
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
Sacramento
ButteLake
ButteLake
ButteLake
ButteLake
ButteLake
Sacramento
Sacramento
ButteLake
CentralCalifornia
Sacramento
Sacramento
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Klamath
Warner
Warner
Warner
Warner
Warner
Warner
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Amargosa
Owens
Owens
Owens
Owens
Owens
Owens
Owens
Owens
Owens
Owens
Owens
Owens
LongValley
LongValley
LongValley
LongValley
LongValley
LongValley
LongValley
LongValley
LongValley
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Lahontan
Colorado
Colorado
Colorado
Colorado
Colorado
Colorado
Colorado
Colorado
SantaAna
SantaAna
SantaAna
SantaAna
SantaAna
SantaAna
SantaAna
Washington
Washington
Washington
Washington
Washington
Columbia
Columbia
Columbia
Washington
Washington
Washington
RelictDace
RelictDace
RelictDace
RelictDace
RelictDace
TuiChub
TuiChub
TuiChub
Group One
Group Two
Group Three

SUPPLEMENTAL TABLE S.1-1. Collection information for samples analyzed in this study. Sample sizes for geographic regions are indicated and further subdivided at Sub-region 1 and Sub-region 2 levels. Specific sampling locations are presented with GPS coordinates and corresponding location number to Figure 1 with the number of samples included in analyses after quality control as detailed in the text.

Geographic Region	Sub-region 1	Sub-region 2	Location	GPS Coordinates (Latitude, Longitude)	filtered sample number
Pacific Northwest (n=15)	Columbia (n=7)	Lower Columbia	Cowlitz River (3)	(46.28972, -122.88000)	3
		Snake River	Pingree (10)	(43.12720, -112.51317)	2
			Tincup Creek (11)	(41.97013, -110.96920)	2
	Washington Coast (n=8)	Chehalis River (2)		(46.61944, -122.93722)	4
		Deschutes River (1)		(47.00175, -122.89638)	4
Great Basin (n=46)	Bonneville (n=2)	Bear River (12)		(41.61901, -112.11223)	2
	Lahontan (n=19)	Walker River	Poore Creek (16)	(38.34278, -119.52881)	8
		Humboldt River	Susie Creek (13)	(40.81223, -116.04667)	8
		Truckee River	Martis Creek (15)	(39.34555, -120.61777)	3
		Honey Lake	Willow Creek	(40.34611, -120.26361)	0
	Warner (n=9)	Hart Lake	Honey Creek (7)	(42.42111, -120.11000)	3
			Clover Creek (6)	(42.47500, -120.18000)	2
		Warner Valley	Twenty Mile Creek (8)	(42.11384, -119.93161)	1
		Coleman Lake	Foskett Spring (9)	(42.06944, -119.83764)	3
	Eagle Lake (n=16)	Butte Lake	Butte Creek (14)	(40.56472, -121.29138)	16
California (Pacific) (n=58)	Klamath (n=19)	Mid-Klamath	Lone Pine Bar (5)	(41.54883, -123.52476)	4
		Lower Klamath	Klamath River near Happy Camp (4)	(41.767682, -123.403174)	15
	Sacramento (n=23)	Pit River	Ash Creek (21)	(41.16111, -120.83000)	4
			Bear Creek (20)	(41.11433, -121.55361)	4
		Goose Lake	Thomas Creek (19)	(42.30416, -120.53250)	4
			Howard Creek (18)	(42.30138, -120.73160)	4
			Quartz Creek (17)	(42.29222, -120.74527)	3
		Sacramento River	Dye Creek (22)	(40.07075, -122.11222)	4
	Southern CA Coast (n=7)	Santa Ana River	Cajon Creek (36)	(34.27475, -117.45261)	3
			Lytle Creek (37)	(34.23899, -117.49892)	4
	Central CA Coast (n=9)	San Luis Obispo Creek	Brizziolari Creek (24)	(35.28000, -120.66889)	4
		Santa Maria River	Sisquoc River (25)	(34.83086, -120.16056)	1
		Monterey Bay	San Lorenzo River (23)	(36.99305, -122.03277)	4
Colorado River (n=8)	Lower Colorado (n=8)	Virgin River	Mesquite Div (38)	(36.81289, -108.00407)	8
Death Valley (n=48)	Amargosa (n=26)	Amargosa River	Beatty Oasis (31)	(36.91398, -116.75551)	8
			Valley-CoffR-Spring (30)	(37.04113, -116.71248)	4
			Amargosa Canyon (35)	(35.84956, -116.23078)	4
		Ash Meadows	Bradford Spring 1 (33)	(36.40214, -116.30236)	4
			Bradford Spring 2 (34)	(36.40124, -116.30304)	4
			Roger's Spring (32)	(36.47915, -116.32646)	2
	Owens (n=22)	Owens River	Rock Creek Ditch (27)	(37.45932, -118.59280)	4
			Matlick Ditch (29)	(37.37602, -118.42450)	4
			Fish Slough (28)	(37.47551, -118.40095)	4
		Long Valley	Whitmore Hot Springs (26)	(37.62977, -118.81017)	10

SUPPLEMENTAL TABLE S.1-2. Pairwise F_{ST} for Speckled Dace. Lower pairwise F_{ST} numbers indicate lower genetic differentiation between the two populations. (A) Pairwise F_{ST} of California samples and samples that share the same node with California samples. (B) Pairwise F_{ST} between non-California samples (Washington, Columbia, Colorado, Washington) and Speckled Dace from the Santa Ana River.

(A)

	Santa Ana	Amargosa	Owens	Long Valley	Lahontan	Sacramento	Central California	Klamath	Butte Lake
Amargosa	0.683038								
Owens	0.655648	0.163662							
Long Valley	0.495095	0.382165	0.300142						
Lahontan	0.549713	0.32826	0.246113	0.259542					
Sacramento	0.645057	0.555592	0.511999	0.443290	0.412116				
Central California	0.753472	0.620699	0.586420	0.488401	0.463870	0.194572			
Klamath	0.549713	0.571066	0.535016	0.456907	0.428349	0.165698	0.326771		
Butte Lake	0.679071	0.576948	0.533901	0.454898	0.424014	0.083905	0.228633	0.217210	
Warner	0.679669	0.559563	0.507790	0.418762	0.388788	0.288543	0.429950	0.319681	0.32908

(B)

	Bonneville	Columbia	Colorado	Washington
Santa Ana	0.597602	0.553871	0.544243	0.697827

SUPPLEMENTAL MATERIAL S.1-1. Individuals selected for reference genome. Names of the filtered BAM files of eight Speckled Dace collected from the Walker River, which are selected to generate a reference genome.

SUPPLEMENTAL MATERIAL S.1-2. Outgroup sequences. Sequences of Tui Chub and Relict Dace used as outgroups for phylogenetic analysis.

SUPPLEMENTAL MATERIAL S.1-3. The list of contigs in the reference genome. Contigs included in the reference genome with the average, minimum, and maximum lengths.

Chapter 2: Creation of a baseline for genetic monitoring of Paiute Cutthroat Trout

Abstract

Refuge populations are frequently created by managers as safeguards against species or lineage extinctions. Yet refuge populations are often small and geographically fragmented, thereby risking loss of genetic diversity and promoting potential rapid genetic divergences. Quantifying genetic baselines is essential for monitoring genetic change in refuge populations, and for informing management actions, such as future translocations. One useful approach for monitoring genetic structure and diversity in refuge populations is to develop a panel of single nucleotide polymorphism (SNPs). The SNP panel should contain non-paralogous, unlinked, neutral markers that can be reliably genotyped to accurately reflect genetic diversity. Paiute cutthroat trout (*Oncorhynchus clarkii seleniris*, PCT) is a subspecies of cutthroat trout historically distributed only in Lower Silver King Creek. However, the subspecies experienced massive population declines because non-native trout were introduced into their native habitat, competing and hybridizing with them. PCT have been restored in nine small refuge populations over the past ~100 years, but the genetic structure and diversity of these populations remains unknown. In this study, we developed a SNP panel to monitor genetic structure and diversity of the nine PCT refuge populations. We applied restriction-site associated (RAD) sequencing on 854 PCT samples collected from 1996 to 2021 from all nine refuge populations to select 1,114 SNPs for our panel to detect the genetic population structure and calculate the population heterozygosity. We compared results between our larger RAD sequencing dataset containing 6,187 SNPs to our 1,114 panel SNPs and found that the panel SNPs generate comparable results

to RAD sequencing SNPs in population heterozygosity. They also produce more resolved estimates of genetic population structure, demonstrating the usefulness of our SNP panel for current and future monitoring and conservation efforts.

1. Introduction

Many species at risk of extinction are restricted to fragmented populations that no longer experience connectivity. These fragmented populations are often numerically small, which simultaneously puts them at risk of losing genetic diversity while promoting divergence over time. One particular kind of fragmented population an at-risk species may have is a refuge population. While definitions of refuge population vary, the one that is applied here is a population created by managers or remains *in situ* under protection as a safeguard against species or lineage extinction (Meretsky et al. 2006). Refuge populations may be especially at risk because upon creation they undergo founder effects, where only a small subset of genetic diversity from the source population(s) is transferred to the new population. This creates an initial reduction in diversity corresponding to the number of individuals used for initial establishment (Finger et al. 2012; Fraser 2008; Templeton et al. 2001). Refuge populations also commonly have small population sizes which further leads to genetic drift and differentiation over time.

To maintain the health of refuge populations, active management may be undertaken. Translocating individuals between populations is one management tool that promotes gene flow among refuge populations. Throughout the western North America, translocations remain one of few tools available to assist with the protection and recovery of native trout species (Stead et al. 2022; Larig and Fausch 2002). Translocations may replace historic gene flow among now isolated populations and mitigate both the loss of the genetic diversity and the increase in genetic

differentiation through drift or local adaptation (Aitken and Whitlock 2013; Hoban et al. 2021; Waters et al. 2015). It is especially critical that translocation efforts be paired with robust genetic monitoring so that success can ultimately be measured; translocations to restore the gene flow among refuge populations are only successful if translocated individuals successfully breed with the resident population, infusing new genetic material. If translocated individuals do not integrate into the recipient population, significant resources may have been wasted, which can setback a conservation program or species recovery.

It is helpful to quantify a genetic baseline before translocation and refuge efforts are initiated at scale. For example, an accurate baseline can be used to inform genetic monitoring and evaluate success over space (across multiple populations), and over time within single populations of interest. A genetic baseline typically includes information on genetic diversity, effective population sizes, and genetic population structure. Additional applications of the baseline include prioritization of populations for additional management actions, genetic stock identification, and evaluations geared towards effectiveness of other habitat, ecosystem and population conservation activities (Clemento et al. 2014).

Genetic markers must be carefully selected for monitoring, particularly when chosen from a large reduced representation sequencing data set (e.g., RAD sequencing). Although large genome wide-datasets provide significant power and can be used without a reference genome, they often take significant time to produce (sequencing is often outsourced), require substantial experience to analyze, and can produce inconsistent results making combining data sets difficult (Deagle et al. 2015; Flanagan and Jones 2018; Leigh et al. 2018; Meek and Larson 2019). These characteristics suggest methods such as RAD sequencing are ideal for generating genetic baselines, with thousands of individuals available for marker discovery. Yet this tool may also

become inefficient for tasks requiring a high throughput of many individuals over prolonged periods of time such as genetic monitoring (Deagle et al. 2015; Leigh et al. 2018). Instead, a well-designed panel of single nucleotide polymorphisms (SNPs) derived from RAD sequencing can provide a useful alternative that enables rapid genotyping with minimal loss of the power and accuracy associated with larger data sets (Bootsma et al. 2020; Clemento et al. 2014; Narum et al. 2008; Smith et al. 2007). Ideally, SNPs selected for genetic monitoring are unlinked, non-paralogous, present in every population of interest in a study system, and can be reliably genotyped. These characteristics enable unbiased estimates of genetic diversity, effective population size, population assignment, and genetic population structure (Bootsma et al. 2020; Du et al. 2019; May et al. 2020; Rufo et al. 2019).

Though SNP panels are highly effective, developing panels for genetic monitoring may not be straightforward. For example, in cases where multiple small refuge populations have undergone significant drift due to isolation and low effective population sizes, they likely have low levels of diversity to begin with, and there may be a limited number of shared polymorphic sites for all the populations (Fraser 2008). This is because over time a high proportion of SNPs have drifted to fixation or loss. Though there may be novel SNPs generated through mutation, if they are only found in a single population, they are not helpful for monitoring all the refuge populations. In addition, natural "noise" generated in the evolution of the species is a factor that can generate inaccurate results. Whole-genome duplications have occurred in the recent ancestors of many fish families, resulting in pervasive paralogous loci (Campbell et al. 2019; Ferris and Whitt 1980). Sequencing reads from paralogous loci in different parts of the genome can collapse into a single SNP locus during assembly, causing fixed differences between paralogs to be erroneously identified as heterozygous genotypes (McKinney et al. 2017; Willis et

al. 2017). Collapsed paralogs can in turn artificially inflate heterozygosity estimates at the individual and population level if they are not properly filtered out of the RAD sequencing dataset (O'Leary et al. 2018; Willis et al. 2017).

The Paiute cutthroat trout (*Oncorhynchus clarkii seleniris,* PCT) provides a unique opportunity to overcome methodological obstacles articulated above for selecting SNP markers for use in monitoring multiple refuge populations. The Paiute cutthroat trout is a subspecies of cutthroat trout, belonging to the Salmonidae family, that experienced whole genome duplication approximately 88 million years ago, and therefore has a high proportion of paralogous genes in their genome (Macqueen and Johnston 2014). Paiute cutthroat trout were historically found in Lower Silver King Creek, which is a headwater of the East Fork of the Carson River, flowing through California and Nevada in the western United States (Finger et al. 2012). Non-native trout including rainbow trout (*O. mykiss*) and Lahontan cutthroat trout (*O. c. henshawi*) have been stocked in the native habitat of PCT, hybridizing with and out-competing them (Cordes et al. 2004). Nine small refuge PCT populations were established 1968-1998 (with founding population sizes ranging from 20 to 401 fish; see Finger, et al. 2012) to protect PCT from competition and hybridization (Table 2-1).

In 2017, a large wildfire burned the North Fork Cottonwood Creek watershed, and ultimately created dangerously low water levels in the stream. It was during this period that the first attempt was made at translocating PCT among the refuge populations, when 88 fish were translocated from North Fork Cottonwood Creek to Upper Silver King Creek (Table 2-1). The primary goal of this study is to analyze change in genetic diversity of the recipient population after this translocation, and to identify a panel of SNPs for future genetic monitoring of all nine refuge populations. We used BestRAD protocol (Ali et al. 2016) to prepare RAD sequencing

libraries for 854 individuals sampled across all refuge populations. We removed paralogous SNPs, filtered SNPs for depth of coverage, and removed F_{ST} outliers associated with RAD sequencing library batch differences. Using this set of unlinked and filtered SNPs, we developed a SNP panel to genetically monitor PCT refuge populations. Ultimately, this study provides a critical tool for managers to monitor multiple refuge populations at risk of extinction. The work may also be useful for researchers who encounter similar batch effects in RAD sequencing datasets more generally.

2. Methods

2.1 Sample collection, DNA sequencing, and quality filtering

Biologists from the California Department of Fish and Wildlife, U.S. Department of Agriculture Forest Service, and U.S. Fish and Wildlife Service collected fin clips from nine PCT refuge populations over multiple years from 1996 to 2021(Figure 2-1, Table 2-1). Five study populations are located within the Carson River Basin, including Upper Silver King Creek, Four Mile Canyon Creek, Fly Valley Creek, Corral Valley Creek, and Coyote Valley Creek (hereafter, we refer to these populations as 'within-basin'). The remaining four populations are Sharktooth Creek and Stairway Creek, which are located in San Joaquin River Basin, and North Fork Cottonwood Creek and Cabin Creek which are located in the Great Basin (hereafter we refer to these four populations as 'out-of-basin').

Fin clips were taken from live fish and were either dried on Whatman qualitative filter paper, placed in coin envelopes, or stored in ethanol at room temperature. We extracted DNA from fin clips with a magnetic bead-based protocol (Ali et al. 2016) and quantified using Quant-iT PicoGreen dsDNA Reagent (Thermo Fisher Scientific) with an FLx800 Fluorescence Reader

(BioTek Instruments). We prepared RAD libraries using the *Sbf1* restriction enzyme following the protocol in Ali et al. (2016) for ten 96-well plates, and then pooled and sequenced all ten libraries across three lanes of an Illumina NovaSeq at the DNA Technologies and Expression Analysis Core at the University of California Davis with paired-end 150-bp reads. We used fastq-multx 1.4.2 to demultiplex the sequencing data with an exact match with plate barcodes (https://github.com/brwnj/fastq-multx) using BarcodeSplitListBestRadPairedEnd.pl provided in Ali et al. (2016). Demultiplexed data was aligned to the rainbow trout reference genome downloaded from NCBI (Bioproject: PRJNA623027) by bwa-mem algorithm in the Burrows-Wheeler Aligner 0.7.17 (Li and Durbin 2009). We used SAMtools 1.7.2 (Li et al. 2009) to sort, remove duplicates, and count the mapped reads for each individual.

We next tested the maximum number of genotypes that could be successfully called for each individual. We ran called genotypes on all 854 individuals in SAMtools 1.7.2 genotype likelihood model (-GL 1) with a uniform prior (-doPost 2) to identify genotypes with mapping scores above 30 (-minMapQ 30), base scores above 20 (-minQ 20) but no minInd filter (-minInd 0). We used R to count the number of called SNPs in each individual and filtered out individuals with < 600,000 mapped reads and < 16,000 called SNPs (Figure 2-2).

2.2 Identifying and removing the batch effect.

2.2.1 Identifying the batch effect.

Initially, we performed a principal component analysis (PCA) on all the individuals that passed screening described above using PCAngsd (Meisner and Albrechtsen 2018) to assess overall quality. To generate the beagle input file in ANGSD 1.9, we calculated genotype likelihoods using the SAMtools model (-GL 1), inferred major and minor alleles directly from

genotype likelihoods (-doMajorMinor 1), estimated posterior genotype probability assuming a uniform prior (-doPost 2) and estimated allele frequencies assuming a fixed major allele and minor allele (-doMaf 1) (Korneliussen et al. 2014). We only used reads with mapping score > 30 (-minMapQ 30) and bases with a quality score 20 (-minQ 20). Furthermore, we only included SNPs with a minor allele frequency of at least 0.01 (-minMaf 0.01) that were represented in at least 60% of samples (-minInd 315).

We found PCA results did not reflect expected population genetic structure based on Finger et al. (2012). Rather, PCA revealed a strong signal associated with RAD library from which samples originated, especially for samples from Library 9 (Figure 2-3A, D). According to Finger et al. (2012), within-basin populations have more similar genetic variation than out-of-Basin populations, so we created two subsets of our data that each included samples in Library 9 and the other libraries that were sourced from the same populations in the similar year (Table 2-2). The first dataset included samples from all within-basin populations that were located both in Library 9 and the other libraries (Library 1, 6). The second dataset included individuals from one out-of-basin population, North Fork Cottonwood Creek (NFC), which were located in Library 9 and the other libraries (Library 7, 8). This allowed us to identify variants associated with library batch differences rather than a true biological signal. We called genotypes for within-basin populations and NFC separately using the same parameters as above using ANGSD 1.9 (Korneliussen et al. 2014). We only called genotypes with data for 60% of individuals (-minInd 222 for within-basin populations, -minInd 99 for NFC) and generated vcf files (-dovcf 1) for each subset.

2.2.2 Removing the batch effect caused by non-overlapping regions.

When we discovered SNP loci for each library independently, some SNPs only appeared in one or several libraries instead of all of them. To assess if these non-shared SNPs contributed to the observed batch effect, we used bedtools (Quinlan and Hall 2010) to identify SNPs in common between Library 9, which had the strongest batch effect, and the other libraries for NFC and within-basin populations separately. We combined all the intersected SNPs that were shared between NFC dataset and within-basin populations dataset, and then called genotypes for all range-wide samples at this combined SNP list, then performed PCA to assess if the batch effect was reduced (Figure 2-3 B, E).

2.2.3 Removing the batch effect caused by a coverage issue.

To investigate cause of the batch effect, we calculated pairwise F_{ST} for each SNP in the samples that were influenced by the batch effect (Table 2-2). We grouped individuals from the sample location (North Fork Cottonwood Creek) or from all within-basin populations that have minimal genetic difference (Finger et al. 2012) to identify SNPs that can be the false positive F_{ST} outliers. We used VCFtools 0.1.14 (Danecek et al. 2011) to compare Weir and Cockerham F_{ST} for every locus in our combined SNP list. We used intersected SNPs and identified SNPs with F_{ST} values >1.5x the interquartile range as outliers (Grubbs 1969). We extracted depth of each SNP for each individual using VCFR 1.12.0 (Knaus and Grünwald 2017), calculated the mean depth of each SNP, and conducted a Wilcoxon signed rank test in R package rstatix 0.7.0 (Kassambara 2020) to compare mean depth of F_{ST} outliers and non-outliers. To reduce the batch effect, we removed all SNPs that we identified as F_{ST} outliers or had the mean depth <5 or >50 (Figure 2-4, Figure 2-3C, F).

2.3 Selecting SNPs for genetic population structures and individual heterozygosity.

2.3.1 Selecting the SNPs

After excluding SNPs that were F_{ST} outliers or had mean depth < 5 or > 50, we performed genotype calling on remaining SNPs for each population or several populations together to evaluate if they share the same genetic population structure in PCA and Finger et al. (2012) (Figure 2-3C, F). We used the same criteria to select the SNPs that have the depth between 5 and 50 (-geno_mindepth 5 and -geno_maxdepth 50). Next, we used HDplot to remove paralogous SNPs called in each population separately to avoid the genetic population structure (McKinney et al. 2017). We selected SNPs with heterozygosity frequency (H) < 0.45 in each population (Danecek et al. 2021; Li 2011). This set of SNPs is referred to as RAD-seq SNPs. To select approximately 1000 SNPs for our genetic monitoring SNP panel, we identified SNPs with allele ratio (D) between 0.4 and 0.6 in each population. This set of SNPs is referred to as 'panel SNPs'. We used BCFtools 1.14 to prune the SNPs that had Lewontin's D > 0.1 to remove linkage disequilibrium and kept only SNPs called in all populations for both SNP sets.

2.3.2 Result validation

To validate accuracy of our panel SNPs, we compared population structure and individual heterozygosity results produced using the RAD-seq SNPs and the panel SNPs. We used a PCA to visualize population structure using each set of SNPs. We used realSFS in ANGSD 1.9 to calculate individual heterozygosity from the folded site frequency spectrum for each set of SNPs (Korneliussen et al. 2014). We calculated mean and variance of individual heterozygosity for each population (excluding sites with < 8 samples). We verified that RAD-seq

SNPs and panel SNPs produced similar individual heterozygosity results by performing a Wilcoxon signed rank test using the R package rstatix with a strict Bonferroni adjusted-P value (Kassambara 2020).

3. Results

3.1 Removing the batch effect.

After filtering individuals for sequencing and mapping quality, 524 out of 854 individuals were kept for further analysis (Figure 2-2). We used 212,592 SNPs in the initial range-wide PCA, where we observed a strong batch effect (Figure 2-3A, D). Genetic variation explained by PCA is extremely low due to the batch effect, noise, and temporal effect. Filtering our SNP dataset to only include loci in common between Library 9 and the other libraries resulted in a dataset with 95,932 SNPs, but PCA using these common SNPs revealed little improvement on the original batch effect (Figure 2-3B, E). We observed F_{ST} outliers in both NFC and within-basin populations (Figure 2-4A, B). We identified 6,533 out of 54,049 SNPs as F_{ST} outliers in NFC, and 6,680 out of 48,711 SNPs F_{ST} outliers in within-basin populations. These F_{ST} outliers had significantly lower depth than the non-outlier SNPs in both NFC and within-basin populations (Figure 2-5). After filtering for F_{ST} outliers and depth of coverage, our final dataset included 29,390 SNPs and a PCA showed the batch effect was substantially reduced (Figure 2-3 C, F). Although the batch effect is removed, the genetic variation explained by each PC is still low due to the existence of a temporal effect.

3.2 Selecting the SNPs to assess population structure and individual heterozygosity.

Based on the final PCA, we identified all within-basin populations as a single genetic group and Cabin Creek (CAB) as a single genetic group. For each of the sites NFC, Stairway Creek (SWC), and Sharktooth Creek (SHK), a strong temporal effect led us to separate samples collected before or after 2012 as two genetic groups in each population (Figure 2-3C). After the genotype calling, SWC, SHK in 2000, and NFC in 2011 were excluded for the further analyses due to the low number of genotypes called compared to the other populations. After merging all the non-paralogous SNPs selected by HDplot in each population, we selected 12,085 SNPs for the RAD-seq SNP dataset and 1,716 SNPs for the panel SNPs dataset. After pruning for linkage disequilibrium, our final RAD-seq SNP dataset included 6,187 SNPs and our final panel SNPs dataset included 1,114 SNPs.

3.3 Validating SNP panels.

3.3.1 PCA

PCAs generated by RAD-seq SNPs (Figure 2-6A) and panel SNPs (Figure 2-6B) were similar, but panel SNPs produced a clearer resolution of the expected population structure (Finger et al. 2012) and revealed less temporal genetic drift between the samples. In the PCA generated by RAD-seq SNPs, the primary result is very similar to results before applying HDplot and pruning linkage disequilibrium (Figure 2-3C). PCA generated by RAD-seq SNPs emphasized temporal genetic drift more than geographic relationships among sampling locations. The out-of-basin populations collected before and after 2012 were separated into two clusters by PC1, which explains 2.62% of the genetic variation in the dataset. within-basin populations and

out-of-basin populations are separated into two groups by PC2, which explained 1.61% of genetic variation. However, genetic population structure at a finer geographical scale, such as for each refuge population, is not observable in the RAD-seq PCA. In contrast, the PCA generated by panel SNPs is less impacted by the temporal effect, and the PCA shows genetic population structure among refuge populations besides the basin level only. Thus, the proportion of genetic variation explained by PC1 and PC2 were also increased compared to RAD-seq PCA as the paralogous SNPs and temporal effect were removed (Figure 2-3). Out-of-basin populations are more differentiated than within-basin populations. Besides the genetic population structure, 12 individuals from Upper Silver King Creek (USKC) sampled in 2021 were located intermediately between within-basin and out-of-basin populations in the RAD-seq PCA, and between NFC and USKC genetic clusters in the panel SNPs PCA. These individuals could be offspring resulting from interbreeding after the 2017 translocation from NFC into USKC, which will be tested in the next chapter.

3.3.2 Individual heterozygosity

To test if our SNP panel accurately estimates genetic diversity compared to the RAD-seq SNPs dataset, we compared individual heterozygosity values using both SNP sets. For most of our populations, we found no significant difference in mean heterozygosity estimates for Panel SNPs compared to RAD-seq SNPs; however we observed more variance in panel SNPs (Figure 2-6C, Table 2-1). However, we did find a significant difference in SWC samples during 2021. Mean individual heterozygosity was only 0.034 lower in SWC 2021, which is also a minimal difference in heterozygosity at the population level.

4. Discussion

4.1 The panel SNPs can successfully monitor genetic diversity.

In this study, we applied RAD sequencing to identify a panel of 1,114 unlinked, non-paralogous SNPs that can be used for long-term genetic monitoring and conservation of all nine refuge PCT populations. After comparing genetic diversity using our panel SNPs to the 6,187 unlinked, non-paralogous RAD-seq SNPs, we found panel SNPs generated comparable measurements of genetic diversity at the population level for all nine refuge populations. This SNP panel can be used to identify potential donor and recipient populations for future translocations and to monitor spatiotemporal genetic diversity changes, including following future translocation events.

4.2 Batch effect was successfully identified and removed.

In this study, we encountered a strong artificial effect of library preparation in RAD sequencing data, which interfered with data analysis to identify informative SNPs for genetic monitoring. Refuge population architecture, containing multiple refuge populations, ultimately creates analytical challenges that make batch effects more problematic. In each refuge population with low levels of genetic diversity or little genetic differentiation among populations, batch effects may present a stronger signal than the biological pattern, making it difficult to discern true genomic relationships among populations. Batch effects can arise in a variety of ways that are not mutually exclusive, including preparation of sequencing libraries by different personnel, errors in standardizing input DNA concentrations, or stochasticity in PCR steps during the process of next generation sequencing (De-Kayne et al. 2021; Leigh et al. 2018; O'Leary et al. 2018). Previous studies have demonstrated how batch effect can be caused by poly-G tails, base

quality score miscalibration, coverage differences, reference bias and alignment errors caused by difference in read type and read length, and DNA degradation (De-Kayne et al. 2021; Lou and Therkildsen 2022; O'Leary et al. 2018). In the process of identifying the cause of the batch effect, we speculate coverage difference is more likely to explain the batch effect in this study. We identified F_{ST} outliers in NFC as well as within-basin populations between Library 9 and the other libraries and the mean depth of the F_{ST} outliers was significantly lower than the non-outliers (Figure 2-5). This indicated that the batch effect was highly associated with SNPs in low depth that can likely cause false homozygosity during genotype calling, which is a common way of associating low depth and inaccuracy in genotype calling (Lou and Therkildsen 2022). To solve the batch effect, we removed all SNPs identified as F_{ST} outliers, and all the SNPs with a depth >50 or <5, which also modulates influence of the missing values in measurement of individual heterozygosity. We acknowledged that some F_{ST} outliers not caused by technical artifacts were also removed from the data set, but we aimed at selecting a neutral SNP panel to detect the population structures and heterozygosity in individual and population levels and therefore would have excluded these high F_{ST} sites anyway. Removing F_{ST} outliers associated with different batches (RAD sequencing libraries, in our case) can sufficiently remove the batch effect for populations with minimal population structure that are sequenced across multiple libraries. This method may be especially valuable for selecting SNPs when the samples from the same populations are partitioned into two libraries and only one of them is influenced by batch effect. If no samples in the library with a batch effect originate from the same location as the samples in the other libraries, performing genotype calling in one batch and applying the SNPs on another batch can also be an effective way to solve the issue (Lou and Therkildsen 2022).

4.3 The selected SNPs can generate a comparable result to RAD sequencing data.

We selected 1,114 SNPs for the genetic monitoring SNP panel and compared the result of PCA and individual heterozygosity to 6,187 SNPs representing the RAD sequencing data, after filtering the paralogous SNPs with low quality depth. In the PCA, panel SNPs reflected the population structures more clearly than the RAD-seq SNPs, while the RAD-seq SNPs reflected more on temporal differences within re-sampled populations. Samples collected before 2012 were removed after the depth threshold was added in the genotype calling, which may be caused by the lower depth due to DNA degradation or the rapid genetic differentiation due to a founder effect. Removal of lower depth samples in the same locations strengthened our ability to distinguish among populations. One common concern for SNP panels used for long-term genetic monitoring is fixation or loss of alleles in selected loci in one or more populations. This effect could ultimately weaken the genetic signal as SNPs become monomorphic for some analyses such as parentage analysis but would record the change of population heterozygosity. We recommend using all the SNPs to measure population heterozygosity but monitoring minor allele frequency over time to ensure effectiveness of analyses requiring additional genetic variation.

In the measurement of individual heterozygosity, Panel SNPs and RAD-seq SNPs have no significant difference in the mean individual heterozygosity in most of the populations. Although SWC in 2021 had significant differences, mean heterozygosity differences generated by either SNP set were minimal. Therefore, panel SNPs likely produce an unbiased measurement of genetic diversity at the population level. The unbiased result can be attributed to our use of the site frequency spectrum to estimate heterozygosity. In a previous study, allele frequency of SNPs was modified by its purpose: if the panel is designed for detecting population structure, SNPs

selected for the panel have high variance between populations (May et al. 2020). This selection step can break down site frequency spectrum of populations and thus generate a biased estimate of heterozygosity. However, maintaining site frequency spectrum also means keeping additional SNPs with low minor allele frequencies such as singletons and doubletons. These SNPs with lower minor allele frequency are less informative and ultimately require the panel to contain more SNPs to accomplish the goal of detecting population structure.

Development of the panel SNPs will allow for use of less expensive and technically demanding technologies to perform long-term monitoring of PCT. Compared to methods that requires a long time and high costs to sequence, these SNPs can be included in amplicon sequencing methods such as GT-seq (Campbell et al. 2015). Due to limits in number of SNPs that can be included in a GT-seq panel, we would need to run 2-3 panels of GT-seq to include all 1,114 SNPs. Hybridization capture methods are an alternative method that allow all the SNPs to be tested simultaneously. In hybridization capture methods, hybridization between sequencing reads and probes allows an effective and equal enrichment of SNPs located in introns to SNPs located in exons (Gasc et al. 2016). These non-primer methods provide a higher limit for number of SNPs than the primer-based method (GT-seq) as no PCR reaction needs are required to produce sequencing amplicons (Turner et al. 2009; Zhang et al. 2001).

5. Conclusion

In this study, we aimed at selecting non-paralogous, unlinked SNPs that exist in all nine refuge populations to develop a SNP panel for long-term genetic monitoring of PCT. In the process of performing RAD sequencing, we met a strong batch effect. To resolve the batch effect, we identified F_{ST} outliers in NFC or all within-basin populations sequenced in two RAD sequencing libraries, and the batch effect was successfully removed after filtering F_{ST} outliers

and the SNPs with extremely high or low read depth. After removing the paralogous and linked SNPs, we selected 1,114 SNPs for the panel and verified that these SNPs can generate unbiased results compared to non-paralogous and unlinked SNPs in RAD sequencing. We compared the result of the population heterozygosity calculated by these SNPs to 6,187 RAD-seq SNPs and found that the 1,114 SNPs indeed generate a comparable population heterozygosity to the RAD-seq SNPs. This SNP panel can be used for conservation management, specifically in genetic monitoring of population structure and heterozygosity metrics. However, efficacy of the method should be revisited over time as analyses that require more genetic variation, such as parentage analysis may ultimately emerge and be more effective.

6. Acknowledgements

Thanks to Sean O'Rourke and Mary E Badger help with extracting DNA and building RAD sequencing libraries. Thanks California Department of Fish and Wildlife personnel for collecting the samples in all nine refuge populations from 1996 to 2021.

7. References

Aitken, S. N., and M. C. Whitlock. 2013. Assisted Gene Flow to Facilitate Local Adaptation to Climate Change. Annual Review of Ecology, Evolution, and Systematics 44(1):367-388.

Ali, O. A., and coauthors. 2016. RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. Genetics 202(2):389-400.

Bootsma, M. L., and coauthors. 2020. A GT-seq panel for walleye (Sander vitreus) provides important insights for efficient development and implementation of amplicon panels in non-model organisms. Molecular Ecology Resources 20(6):1706-1722.

Campbell, N. R., S. A. Harmon, and S. R. Narum. 2015. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources 15(4):855-867.

Campbell, M. A., M. C. Hale, G. J. McKinney, K. M. Nichols, and D. E. Pearse. 2019. Long-Term Conservation of Ohnologs Through Partial Tetrasomy Following Whole-Genome Duplication in Salmonidae. G3 (Bethesda) 9(6):2017-2028.

Clemento, A., E. Crandall, J. Garza, and E. Anderson. 2014. Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (Oncorhynchus tshawytscha) in the California Current large marine ecosystem. Fishery Bulletin 112:112-130.

Danecek, P., and coauthors. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156-2158.

Danecek, P., and coauthors. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10(2).

De-Kayne, R., and coauthors. 2021. Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. Molecular Ecology Resources 21(3):653-660.

Deagle, B. E., C. Faux, S. Kawaguchi, B. Meyer, and S. N. Jarman. 2015. Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. Molecular Ecology 24(19):4943-4959.

Du, H., and coauthors. 2019. Target sequencing reveals genetic diversity, population structure, core-SNP markers, and fruit shape-associated loci in pepper varieties. BMC Plant Biology 19(1):578.

Ferris, S. D., and G. S. Whitt. 1980. Genetic Variability in Species with Extensive Gene Duplication: The Tetraploid Catostomid Fishes. The American Naturalist 115(5):650-666.

Finger, J. A., B. Mahardja, and B. May. 2012. Genetic management plan for the Paiute cutthroat trout (*Oncorhynchus clarkii seleniris*). G. V. Lab, editor.

Flanagan, S. P., and A. G. Jones. 2018. Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. Molecular Ecology Resources 18(2):264-280.

Fraser, D. J. 2008. How well can captive breeding programs conserve biodiversity? A review of salmonids. Evolutionary Applications 1(4):535-586.

Gasc, C., E. Peyretaillade, and P. Peyret. 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. Nucleic Acids Research 44(10):4504-4518.

Grubbs, F. E. 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics 11(1):1-21.

Hoban, S., and coauthors. 2021. Global Commitments to Conserving and Monitoring Genetic Diversity Are Now Necessary and Feasible. BioScience 71(9):964-976.

Kassambara, A. 2020. Pipe-Friendly Framework for Basic Statistical Tests [R package rstatix version 0.6.0].

Knaus, B. J., and N. J. Grünwald. 2017. vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources 17(1):44-53.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15(1):356.

Leigh, D. M., H. E. L. Lischer, C. Grossen, and L. F. Keller. 2018. Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. Molecular Ecology Resources 18(4):778-788.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987-2993.

Li, H., and coauthors. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

Lou, R. N., and N. O. Therkildsen. 2022. Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. Molecular Ecology Resources 22(5):1678-1692.

Macqueen, D. J., and I. A. Johnston. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proceedings of the Royal Society B: Biological Sciences 281(1778):20132881.

May, S. A., G. J. McKinney, R. Hilborn, L. Hauser, and K. A. Naish. 2020. Power of a dual-use SNP panel for pedigree reconstruction and population assignment. Ecology and evolution 10(17):9522-9531.

McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources 17(4):656-669.

Meek, M. H., and W. A. Larson. 2019. The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. Molecular Ecology Resources 19(4):795-803.

Meisner, J., and A. Albrechtsen. 2018. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210(2):719-731.

Meretsky, V. J., and coauthors. 2006. New Directions in Conservation for the National Wildlife Refuge System. BioScience 56(2):135-143.

Narum, S. R., and coauthors. 2008. Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. Molecular Ecology 17(15):3464-3477.

O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy. 2018. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology 27(16):3193-3206.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-842.

Rufo, R., F. Alvaro, C. Royo, and J. M. Soriano. 2019. From landraces to improved cultivars: Assessment of genetic diversity and population structure of Mediterranean wheat using SNP markers. PloS one 14(7):e0219867.

Smith, C. T., and coauthors. 2007. Impacts of Marker Class Bias Relative to Locus-Specific Variability on Population Inferences in Chinook Salmon: A Comparison of Single-Nucleotide Polymorphisms with Short Tandem Repeats and Allozymes. Transactions of the American Fisheries Society 136(6):1674-1687.

Templeton, A. R., R. J. Robertson, J. Brisson, and J. Strasburg. 2001. Disrupting evolutionary processes: The effect of habitat fragmentation on collared lizards in the Missouri Ozarks. Proceedings of the National Academy of Sciences 98(10):5426-5432.

Turner, E. H., S. B. Ng, D. A. Nickerson, and J. Shendure. 2009. Methods for Genomic Partitioning. Annual Review of Genomics and Human Genetics 10(1):263-284.

Waters, C. D., and coauthors. 2015. Effectiveness of managed gene flow in reducing genetic divergence associated with captive breeding. Evolutionary Applications 8(10):956-971.

Willis, S. C., C. M. Hollenbeck, J. B. Puritz, J. R. Gold, and D. S. Portnoy. 2017. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. Molecular Ecology Resources 17(5):955-965.

Zhang, D. Y., M. Brandwein, T. Hsuih, and H. B. Li. 2001. Ramification amplification: A novel isothermal DNA amplification method. Molecular Diagnosis 6(2):141-150.

8. Figures and Tables

FIGURE 2-1. Maps showing relative geographical location of each refuge population. (A) Two out-of-basin refuge populations are located in the Cottonwood Creek and Cabin Creek; (B) Within-basin refuge populations include: Upper Silver King Creek, Four Mile Canyon Creek, Fly Valley Creek, Corral Creek and Coyote Creek. (C) Two out-of-basin refuge populations are located in Stairway Creek and Sharktooth Creek. Sampling years are included on the stream label for each refuge population.

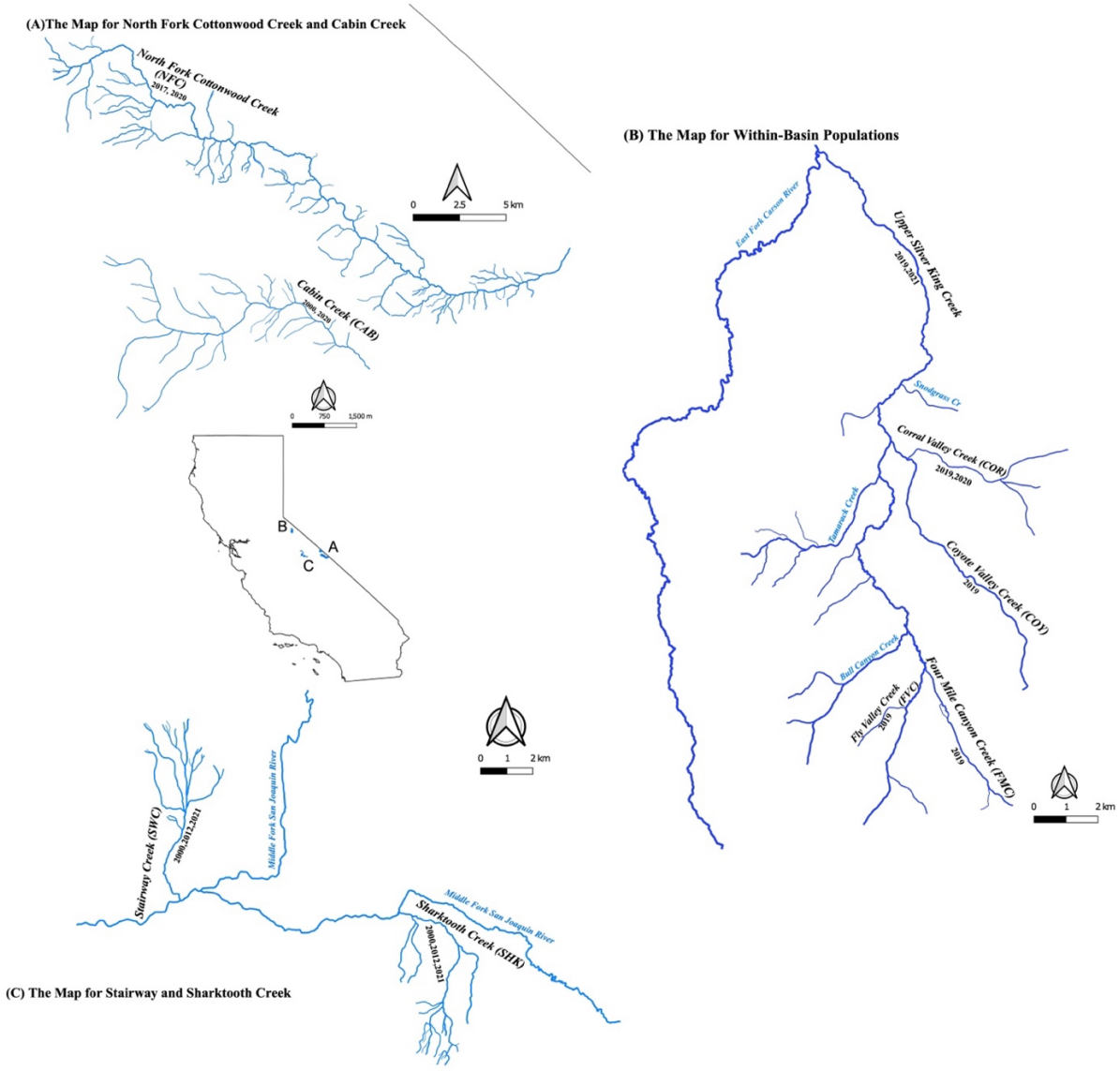


FIGURE 2-2. Called SNPs plotted as a function of mapped reads highlights individuals passing the quality filtering procedure. Individuals from each library were selected by number of called SNPs and the number of the mapped reads. Color represents libraries where the individual was obtained. The dashed line denotes the threshold for number of called SNPs and number of

mapped reads. The inset figure zooms into the distribution of called SNPs and mapped reads near the threshold to aid interpretability.

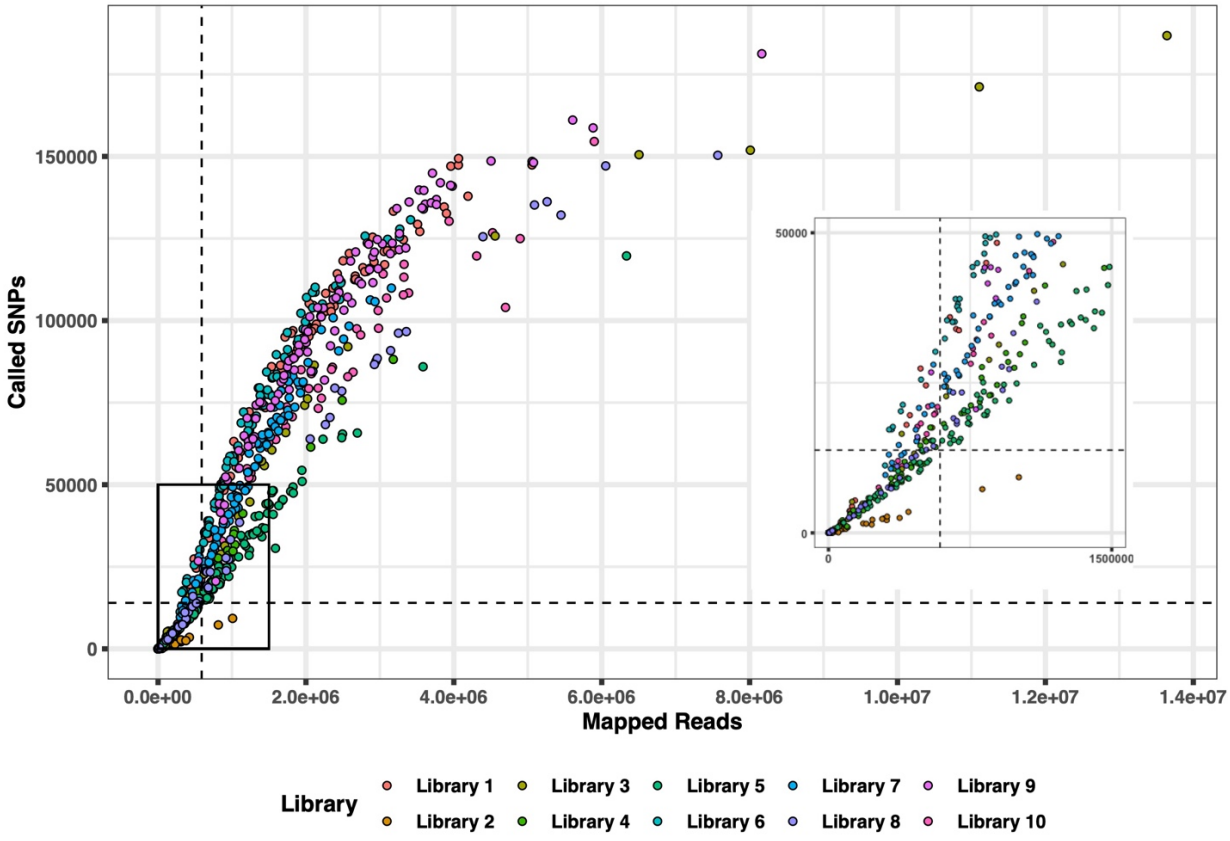


FIGURE 2-3. PCAs with samples coded by population and sequencing library. PCAs colored by population structures are generated using SNPs (A) with no filter, (B) after removing non-overlapping SNPs, (C) after removing F_{ST} outliers. PCAs colored by RAD sequencing libraries, are generated by SNPs (D) with no filter, (E) after removing non-overlapping SNPs, (F) after removing F_{ST} outliers. Shapes indicate sampling years in all PCAs and individuals most impacted by the batch effect were circled using purple ovals.

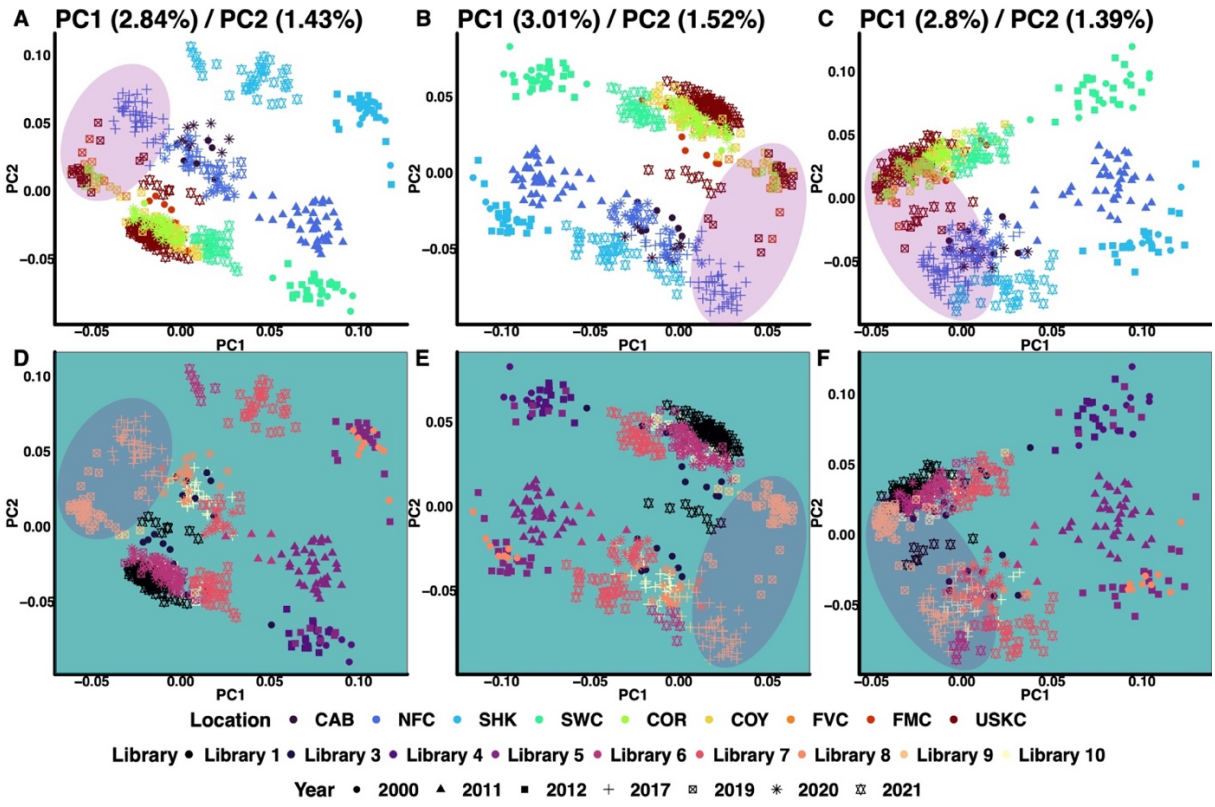


FIGURE 2-4. F_{ST} of SNPs (black and gray points) vary alongside position in the genome. However, F_{ST} values also vary for the shared SNPs between Library 9 and other libraries for (A) NFC, (B) within-basin populations.

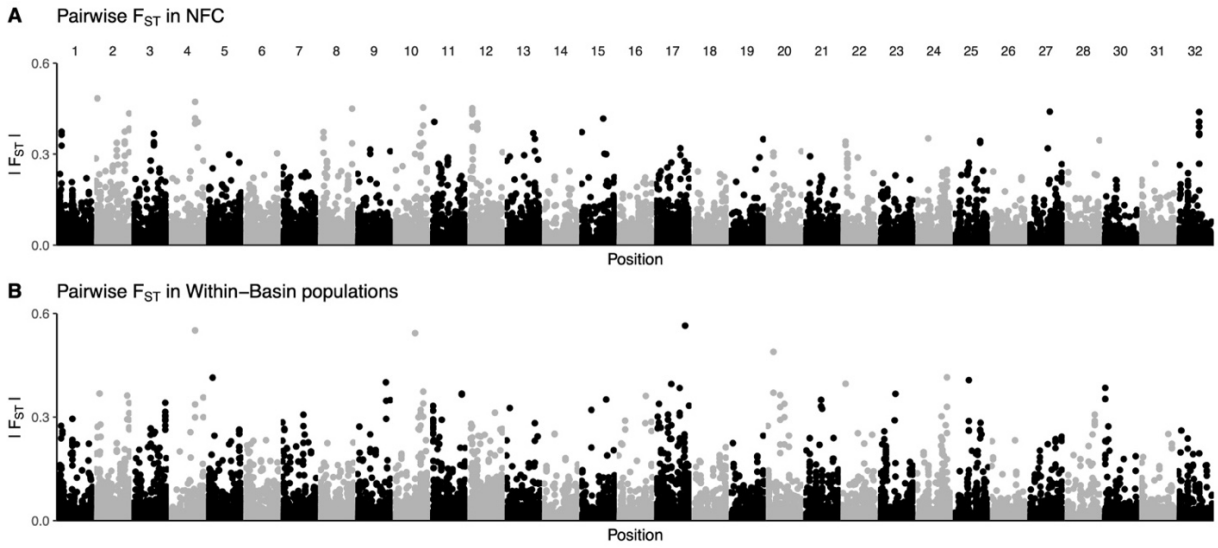


FIGURE 2-5. Mean depth in F_{ST} outliers and non-outlier SNPs. Distribution of mean depth was plotted for F_{ST} outliers and non-outlier SNPs in (A) NFC in Library 9, (B) NFC in other libraries, (C) within-basin in Library 9, (D) within-basin in other libraries. (E) Significant differences in depth between F_{ST} outliers and non-outlier SNPs in NFC. (F) Significant differences in depth between F_{ST} outliers and non-outlier SNPs for within-basin populations (P-values: ****=0.0001).

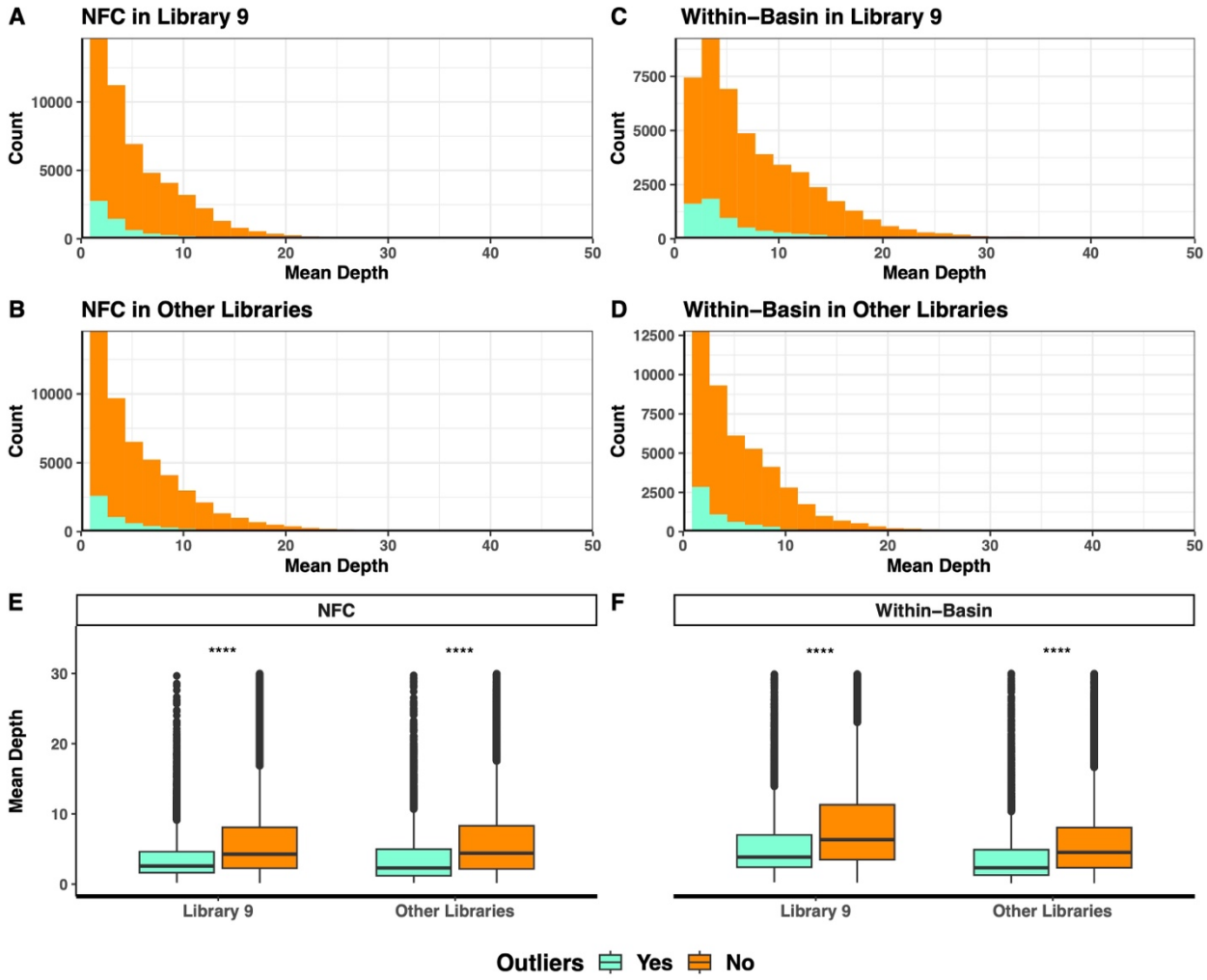


FIGURE 2-6. Validation of selected SNPs. (A) PCA generated by unlinked, non-paralogous SNPs in RAD sequencing. (B) PCA generated by unlinked, non-paralogous SNPs in panel SNPs. (C) Comparison of individual heterozygosity estimated by RAD-seq SNPs and panel SNPs. Significance is labeled with * and P-value > 0.05 was not labeled.

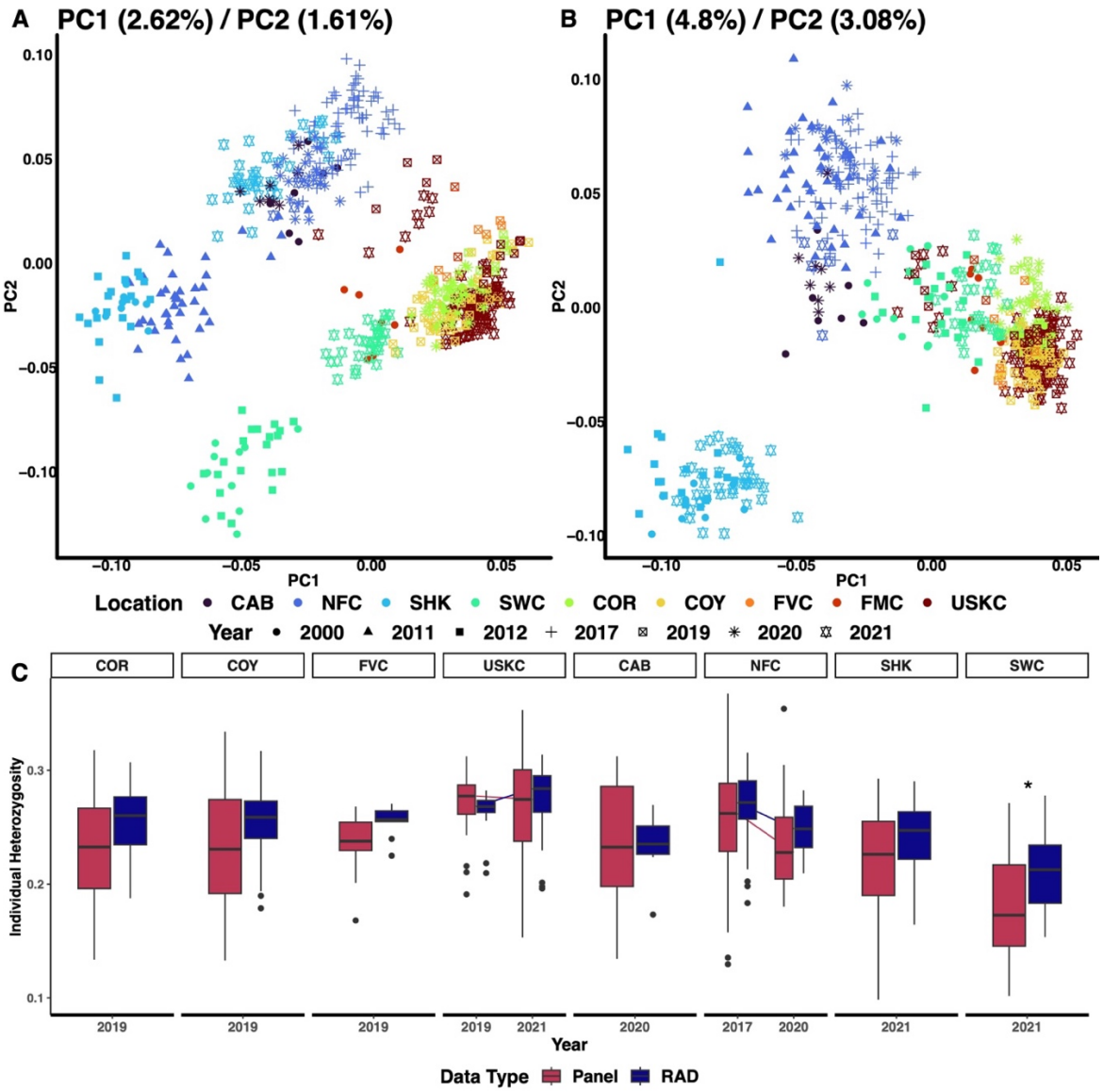


TABLE 2-1. Refuge population sampling information and individual heterozygosity values. Number of SNPs discovered in each population for RAD-seq and panel SNPs before and after filtering paralogous SNPs by HDplot, mean and variance of individual heterozygosity generated using RAD-seq and panel SNPs are listed in the table. NFC 2020 and 2021, and COR 2019 and 2020 were merged into one population separately (labeled as NFC 2020 and COR 2019 in the analyses) because of the low sampling size in NFC 2021 and COR 2020.

Basin	Type	Location	Sample collection year	Sample size	Selected SNPs		Mean		Variance	
					Before HDplot	After HDplot	RAD-seq SNPs	Panel SNPs	RAD-seq SNPs	Panel SNPs
Carson River Basin	within-basin	Corral Valley Creek (COR)	2019	10	8423	305	0.256	0.231	0.00078	0.0021
			2020	39						
		Coyote Valley Creek (COY)	2019	55			0.255	0.232	0.00099	0.0025
		Four Mile Canyon Creek (FMC)	2019	2			NA			
		Fly Valley Creek (FVC)	2019	9			0.256	0.234	0.00022	0.001
		Upper Silver King Creek (USKC)	2019	19			0.264	0.269	0.00037	0.00011
			2021	87			0.276	0.267	0.00067	0.0023
Great Basin	out-of-basin	Cabin Creek (CAB)	2000	6	filtered out		NA			
			2020	8	6823	662	0.233	0.235	0.00082	0.0035
		North fork Cottonwood Creek (NFC)	2017	75	8054	621	0.270	0.258	0.00074	0.002
			2020	14			0.249	0.234	0.00044	0.0018
			2021	6						
San Joaquin Basin		Sharktooth Creek (SHK)	2000	9	filtered out		NA			
			2012	20						
			2021	37	6898	589	0.216	0.241	0.0011	0.002
		Stairway Creek (SWC)	2000	11	filtered out		NA			
			2012	20						
			2021	32	5383	365	0.177	0.211	0.0012	0.0019

TABLE 2-2. The sampling populations in each RAD sequencing library. Populations highlighted in red represent samples influenced by the batch effect, and populations highlighted in yellow denote samples not influenced by the batch effect but were used as the comparison group in pairwise F_{ST}.

Library	Population	Sampling Year	Sample size	Sample size after filter
Library 1	Upper Siver King Creek	2021	96	87
Library 2	Upper Siver King Creek	2021	20	0
	North Fork Cottonwood Creek	1996	37	0
	Cabin Creek	2000	39	0
Library 3	Cabin Creek	2000	11	7
	Four Mile Canyon Creek	2000	9	7
	Fly Valley Creek	2000	31	0
	Corral Valley Creek	2000	39	3
	Coyote Valley Creek	2000	1	0
	Stairway Creek	2000	4	1
	Upper Siver King Creek	2000	1	1
Library 4	Stairway Creek	2000	43	10
		2012	25	13
Library 5	Stairway Creek	2012	14	7
	Sharktooth Creek	2012	40	20
	North Fork Cottonwood Creek	2011	42	35
Library 6	North Fork Cottonwood Creek	2011	8	5
	Coyote Valley Creek	2019	35	32
	Corral Valley Creek	2020	44	39
	Sharktooth Creek	2021	9	8
Library 7	Sharktooth Creek	2021	31	29
	Stairway Creek	2021	40	32
	North Fork Cottonwood Creek	2020	19	14
		2021	6	6
Library 8	North Fork Cottonwood Creek	2020	10	10
	Cabin Creek	2020	8	8

	Sharktooth Creek	2000	28	9
Library 9	North Fork Cottonwood Creek	2017	48	47
	Coyote Valley Creek	2019	15	15
	Corral Valley Creek	2019	8	7
	Upper Silver King Creek	2019	18	17
	Four Mile Canyon Creek	2019	2	2
	Fly Valley Creek	2019	5	5
Library 10	Coyote Valley Creek	2019	12	8
	Corral Valley Creek	2019	8	3
	Upper Silver King Creek	2019	6	5
	North Fork Cottonwood Creek	2017	37	28
	Fly Valley Creek	2019	5	4
		Total	854	524

Chapter 3: The role of genetic monitoring of refuge populations to save Paiute Cutthroat Trout

Abstract

Genetic monitoring is commonly employed in conservation management to track the health and robustness of refuge populations, especially when multiple populations persist. Paiute cutthroat trout (*Oncorhynchus clarkii seleniris*) is a subspecies of cutthroat trout listed as 'Endangered' under the US Endangered Species Act by the US Fish and Wildlife Service (USFWS). The subspecies was historically confined to only Lower Silver King Creek, California; thus it has long been at high risk of extinction. Nine refuge populations (five located within the Carson River drainage and four outside the Carson River drainage) were developed by managers using translocation after non-native salmonids invaded their small natal footprint. Because founding populations for the refuges were naturally small, potential loss of genetic diversity and rapid genetic divergence has always been a conservation concern. For example, populations may lack genetic diversity such that they fail to adapt to novel environments (to which they were translocated) over time. In 2017, 88 PCT from North Fork Cottonwood Creek were rescued and translocated to Upper Silver King Creek. This provided an opportunity to assess how gene flow generated via the translocation contributed to genetic diversity in the recipient population. In this study, we applied 1,114 single nucleotide polymorphisms that had been identified for genetic monitoring of Paiute cutthroat trout, to estimate genetic population structure and population heterozygosity on all refuge populations as well as parentage assignment on translocated individuals. We found refuge populations located outside Carson River drainage have more genetic differentiation than those located within the Carson River

drainage. Every refuge population had similar population heterozygosity levels, but Upper Silver King Creek had the highest and Stairway Creek had lowest population heterozygosity. We found 12 hybrid offspring in Upper Silver King Creek after translocation and the parents of seven individuals were successfully identified. Overall, selected SNPs can successfully identify hybrid offspring generated by translocation and can assess the change of genetic diversity in the recipient population after the translocation. These techniques will be useful for conservation management of PCT, and similar efforts may be helpful to monitoring spatiotemporal variation in genomics of other native and declining trout species.

1. Introduction

The Paiute cutthroat trout (*Oncorhynchus clarkii seleniris,* PCT) is a subspecies of cutthroat trout that is endemic to California and has the narrowest range of any cutthroat trout subspecies; it is found only in the headwaters of the East Fork Carson River, California (Cordes et al. 2004; Finger et al. 2012; Alber 2020). Originally PCT occupied no more than 16 km of Lower Silver King Creek (LSKC) below Llewellyn Falls and above a barrier falls in Silver King Canyon gorge (Behnke 2002). Isolation of PCT from its sister subspecies, Lahontan cutthroat trout (*O. c. henshawi,* LCT) is estimated to be from 0.26-0.85 million years ago (Saglam et al. 2017). This relatively long isolation from LCT resulted in loss of body spotting in PCT; thus PCT became the only cutthroat trout that has almost completely lost body spotting (Behnke 1965). Since at least 1924, Paiute cutthroat trout intensely threatened by introduction of non-native salmonids such as rainbow trout (*Oncorhynchus mykiss,* RT) and Lahontan cutthroat trout (Corde et al. 2004, Finger et al. 2010). These non-native salmonids compete and hybridize with Paiute cutthroat trout, which significantly impair PCT populations and make them vulnerable to extinction. PCT was listed as 'endangered' in 1967 under the federal Endangered Species

Preservation Act of 1967 (USFWS 1967) but was down-listed to 'threatened' (USFWS 2004) to allow more flexible management, including increased translocation experimentation (Finger et al. 2012).

Starting in 1946, in order to protect PCT from introgression, individuals were translocated out of their native habitat in Silver King Creek to establish refuge populations, resulting in nine total populations that persist today (Figure 2-1, Table 3-1, Finger et al. 2012). While the existence of multiple refuge populations helps to protect PCT from extinction, each individual refuge population was established with relatively few individuals and continues to support just small populations (founding size from 20-145, Finger et al. 2012). In turn, these small, isolated populations are threatened by a series of interacting factors including loss of genetic diversity, inbreeding depression, reduced population fitness, and lower effective population sizes, all of which can ultimately promote extinction (Whiteley et al. 2015). Small and isolated populations are also more vulnerable to catastrophic environmental events such as wildfires and drought (Caughley 1994; Hedrick et al. 1996). Thus, each refuge population is still at high risk of extinction and requires ongoing science-based management and monitoring. Moving individuals among refuge populations via translocation is one management option strongly considered because translocated individuals will generate hybrid F1 offspring and transfer new genetic variation to the recipient population (Whiteley et al. 2015). This assisted gene flow can ultimately slow the erosion of genetic diversity and possibly increase genetic and phenotypic diversity for individual refuge populations (Whiteley et al. 2015; Kolodny et al. 2019). However, genetic rescue may be complicated by genetic differentiation from bottlenecks and local adaptation following the translocation (Weeks et al. 2011; Furlan et al. 2020). Thus, populations that have divergent genetic population structure or possess local adaptations should

be avoided for translocation because they can lead to outbreeding depression and failed translocation efforts (Gilk et al. 2004; Furlan et al. 2020).

Over the last 60 years there have been numerous management efforts aimed at promoting recovery of PCT (Cordes et al. 2004) including a recent rescue leading to a translocation. In 2017, a wildfire burned inside of the North Fork Cottonwood Creek (NFC) watershed, one of the out-of-basin populations located in physiographical province Great Basin, California, and the water level of North Fork Cottonwood Creek was extremely low (Alber 2020). To rescue this refuge population from drought, 88 PCT that were ready to spawn were translocated from North Fork Cottonwood Creek to Upper Silver King Creek (USKC). Yet it remains unclear whether this translocation was at all successful. A translocation or genetic rescue can only be considered successful if translocated individuals breed with the recipient population, and genetic monitoring is required to detect interbreeding. After the translocation, the success of generating F1 hybrid offspring by the translocated individuals should be evaluated to guarantee enhanced genetic diversity in the recipient population. In this study, we used a panel of SNPs discovered by Su et al. (In prep) to measure genetic diversity of all the refuge populations and to monitor change in genetic diversity before and after the translocation to Upper Silver King Creek. Characterizing genetic population structure and diversity of all refuge populations would be helpful in selecting donor populations for future translocation efforts.

2. Methods

2.1 Sampling, DNA sequencing, and quality filtering

Biologists from California Department of Fish and Wildlife (CDFW), US Forest Service (USFS), and USFWS collected fin clips from all nine refuge populations over a period of multiple years (Table 3-1). Five of the refuge populations (Upper Silver King Creek, Four Mile

Caynon Creek, Fly Valley Creek, Coyote Valley Creek, Corral Valley Creek) were located within the Carson River Drainage (hereafter called "within-basin" populations). The remaining four populations are split into the San Joaquin River Basin, Sharktooth Creek and Stairway Creek, and the Great Basin, North Fork Cottonwood Creek and Cabin Creek (hereafter we refer to these four populations as "out-of-basin").To monitor effects of translocation, 24 approximately one-year-old (< 60 mm) individuals were sampled from Upper Silver King Creek to detect F1s (fish with 1 NFC parent and 1 USKC parent) in 2019 (Wong, 1975; Titus et al. 2009), and Upper Silver King Creek was fully resampled for the first time after NFC translocation in 2021.

To collect the DNA, adipose fin clips were taken from live fish and were either dried on Whatman qualitative filter paper, placed in coin envelopes, or stored in ethanol at room temperature. DNA was extracted from fin clips with a magnetic bead-based protocol (Ali et al. 2016) and quantified using Quant-iT PicoGreen dsDNA Reagent (Thermo Fisher Scientific) with an FLx800 Fluorescence Reader (BioTek Instruments). We prepared RAD libraries using the *Sbf1* restriction enzyme following Ali et al. (2016) for ten 96-well plates, and then pooled and sequenced all ten libraries across three lanes of an Illumina NovaSeq at the University of California Davis DNA Technologies and Expression Analysis Core with paired-end 150-bp reads. We used fastq-multx 1.4.2 to demultiplex sequencing data with an exact match with well and plate barcodes (https://github.com/brwnj/fastq-multx). Demultiplexed data were aligned to the rainbow trout reference genome downloaded from NCBI (https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_013265735.2/) using the bwa-mem algorithm in the Burrows-Wheeler Aligner 0.7.17 (Li and Durbin 2009). We used SAMtools

1.7.2 (Li 2011) to sort, remove duplicates, and count all mapped reads. We only selected individuals with > 500,000 mapped reads for downstream analysis (N=524, Table 3-1).

2.2 Detecting the genetic population structure of nine refuge populations.

2.2.1 PCA

To investigate range-wide population genetic structure of the refuge populations, a Principal Component Analysis (PCA) was conducted with all the samples (range-wide PCA). We used the SNP panel from Su et al. (in prep) (-rf) and filtered SNPs with minor allele frequencies < 0.05 (-minMaf 0.05) to generate a beagle input in ANGSD 1.9 (Kalyaanamoorthy et al. 2017). Afterward, we used PCAngsd (Meisner and Albrechtsen 2018) to generate a genetic covariance matrix. We used eig() function in R 4.1.2 to compute the eigen matrix and used ggplot2 (Wickham 2016) to visualize dimensionality and variation within the context of the PCs.

2.2.2 Range-wide Admixture Analysis

We further characterized genetic structure of PCT by running a range-wide admixture analysis on the nine refuge populations. We used the same beagle file generated for PCAngsd to perform the admixture analysis. We used NGSadmix (Skotte et al. 2013) to run 10 iterations of each K value (the number of genetic groups) from 1 to 10. After using Evanno method (Evanno et al. 2005) to select optimal K from 1 to 10, we visualized the admixture plot for the selected optimal K value and its adjacent K values.

2.2.3 F_{ST}

To quantify degree of genetic differentiation among refuge populations, we estimated pairwise F_{ST} values. We used ANGSD 1.9 to perform genotype calling on all SNPs and

generated a VCF file (-dovcf 1). We used VCFtools 0.1.14 (Danecek et al. 2011) to calculate the Weir and Cockerham F_{ST} for every SNP. We considered missing values and the negative F_{ST} values as 0 and calculated the mean of the F_{ST} values for all the SNPs for each population. The pairwise F_{ST} between different years for the same location was also calculated if the sampling year of one location is > 1. To compare if genetic differentiation is greater among within-basin populations, out-of-basin populations, or between them, pairwise comparisons were categorized into "In ~ In" (within-basin population and within-basin population), "In ~ Out" (within-basin population and out-of-Basin population), "Out ~ Out" (out-of-basin population and out-of-basin population). The heatmap of the pairwise F_{ST} and the boxplot of each category were visualized using R package ggplot2 (Wickham 2016).

2.3 Estimating population heterozygosity for all refuge populations.

To estimate levels of genetic diversity within each population, we calculated mean of the individual heterozygosity. To calculate individual heterozygosity, we generated the folded site allele frequency file (.saf) for all SNPs in each individual using ANGSD (Kalyaanamoorthy et al. 2017). Single site allele frequency files were used as the input to generate the site frequency spectrum for each individual, and the individual heterozygosity was estimated based on the site frequency spectrum. We excluded populations with fewer than eight individuals sampled for this analysis and performed Wilcoxon signed rank test with Bonferroni adjustment to detect significant differences among the refuge populations. To detect change in heterozygosity following translocation, we compared heterozygosity between non-hybrid individuals and hybrid individuals identified by the admixture analysis in USKC 2019 and 2021 and performed a Wilcoxon signed rank test with Bonferroni adjustment using R package rstatix (Kassambara 2020). Mean individual heterozygosity was calculated by mean() function in R (R Core Team

2023), and the boxplot of the individual heterozygosity was visualized in R package ggplot2 (Wickham 2016).

2.4 Identifying offspring-parent relationships after translocation.

2.4.1 USKC/NFC Admixture analysis

To detect any F1 hybrid offspring after the translocation from NFC to USKC, we performed another admixture analysis that only included the donor and recipient populations: NFC in 2017, USKC in 2019, and USKC in 2021. Similar to range-wide admixture analysis, we used the SNP panel to generate a beagle file in ANGSD and used NGSadmix to generate the q-value matrix for the selected populations when K = 2. We identified F1 hybrid offspring as individuals with more than 40% of their ancestry from the recipient populations.

2.4.2 Parentage Analysis

To detect any F1 offspring resulting from the NFC translocation event amongst USKC fish sampled in 2019 and 2021, we used the software program COLONY2 to conduct parentage analysis (Jones and Wang 2010). To generate the input file for COLONY2, we selected all individuals that passed the sequencing and mapping filters in the donor and recipient populations and performed genotype calling in ANGSD 1.9 using the SNP panel and used write_colony_input(), a function of the software program snpR (Hemstrom and Jones 2023).

COLONY2 requires genotypes from potential parents and offspring to conduct parentage analysis. For potential parents, we used translocated NFC individuals (N=75) and non-admixed USKC individuals (N=16) which were non-admixed shown in admixture analysis. We included all sampled fish as potential mothers and fathers, as we did not know sex of individuals. For

potential offspring in this analysis, we used genotypes of all individuals collected in USKC in 2021 (N=87) and any individuals identified as admixed (using the NGSAdmix analysis) collected in USKC in 2019. For negative controls (to ensure that COLONY was not selecting spurious parents), we included as potential offspring two individuals collected from Coyote Valley Creek in 2019. We set the probability of parents being present in our dataset to 0.5. We also used the full likelihood method, set genotype error rate = 0.1 and dropout rate = 0.01 for one medium length run. We allowed outbreeding, and selected random mating, and polygamy for the mating system.

3. Results

3.1 Within-basin populations are more similar to each other than out-of-basin populations.

3.1.1 PCA

In the PCA of all the populations, the first three PC axes explain more genetic variation than the other PCs (Supplemental Figure S.3-1). Thus, we plotted PC1 and PC2 as well as PC1 and PC3 (Figure 3-1A, B). PC1 explains 5.18% of the genetic variation and PC2 explains 3.05% of the genetic variation. In PC1 and PC2, four out-of-basin populations, North Fork Cottonwood Creek, Cabin Creek, Sharktooth Creek, and Stairway Creek are separated from each other as well as from the within-basin SKC watershed populations (Figure 3-1A). Compared to the other three out-of-basin populations, the cluster of Stairway Creek is relatively closer to within-basin populations (Figure 3-1A). Cabin Creek, and North Fork Cottonwood Creek are more closely related to each other than the distance of them to Sharktooth Creek (Figure 3-1A). Within-basin populations are more closely related to each other than out-of-basin populations (Figure 3-1A).

PC3, which explains 1.71% of genetic variation, reveals more genetic differentiation within the within-basin populations (Figure 3-1B). Corral Valley Creek is more unique from the other within-basin populations, while Four Mile Canyon Creek, Fly Valley Creek, Coyote Valley Creek, and Upper Silver King Creek are still clustered (Figure 3-1B).

3.3.2 Range-wide Admixture Analysis

In the admixture analysis, the optimal K range-wide is 1 (Supplemental Figure S.3-2). As the delta K values of K = 2, 3, 4 are higher than the delta K values of the other Ks, we visualized the admixture analysis from K=2-4. When K = 2, three of four out-of-basin populations, Sharktooth Creek, North Fork Cottonwood Creek, Cabin Creek are separated into one cluster, and all five the within-basin populations, Four Mile Canyon Creek, Fly Valley Creek, Coyote Valley Creek, Corral Valley Creek, Upper Silver King Creek, are separated into another cluster (Figure 3-1C). However, Stairway Creek showed more admixture between the two clusters when K =2. Most of the Upper Silver King Creek populations are partitioned into the same cluster as the other within-basin populations, but 12 Upper Silver King Creek individuals had the admixture of both clusters. When K = 3, Sharktooth Creek forms a distinct cluster itself from all the other out-of-basin populations. When K = 4, Corral Valley Creek separates from the other within-basin populations, and Stairway Creek, Four Mile Canyon Creek, Fly Valley Creek, and Coyote Valley Creek also have a high proportion of admixed ancestry in this cluster.

3.3.3 F_{ST}

F_{ST} values are similar to results in PCA and admixture analysis. Within-basin populations are more similar to each other than out-of-basins as the pairwise F_{ST} values between two within-basin populations (range: 0.013 - 0.036 mean: 0.025) are significantly smaller than F_{ST} values of two out-of-basin populations (range: 0.030 - 0.068 mean: 0.053; P-value: 1.1e-5) or between one

within-basin population and one out-of-basin population (range: 0.034 - 0.065 mean: 0.052; P-value: 2e-08) (Figure 3-1D, E). mean of pairwise F_{ST} values of two out-of-basin populations are not significantly different from the mean of pairwise F_{ST} values of one within-basin population and one out-of-basin population (Figure 3-1E). Similar to the PCA and admixture analysis, Sharktooth Creek has the highest F_{ST} values to all the other populations (range: 0.048 - 0.068, mean: 0.060). Upper Silver King Creek in 2021 has the lowest F_{ST} to all the other populations (range: 0.013 - 0.062, mean: 0.042). The F_{ST} caused by the temporal effect (the different sampling years in the same location) was relatively small compared to the spatial effect. The F_{ST} value between Upper Silver King Creek in 2019 and 2021 was 0.013 and the F_{ST} value between NFC in 2017 and 2020 was 0.023.

3.2 Successful identification of parents for seven out of 12 hybridized individuals.

3.2.1 USKC/NFC Admixture Analysis

To identify hybridized F1 offspring following translocation, we ran admixture analysis on the donor and recipient populations involved in this translocation, and ultimately identified 12 hybridized individuals. Four individuals were collected in 2019 and the other eight were collected in 2021. Using 40% ancestry from the donor population as the threshold for F1 offspring in the recipient population, we identified seven out of the 12 individuals as potential offspring from the F1 generation between donor and recipient populations.

3.3.2 Parentage Analysis

After running the parentage analysis, we successfully identified the parent from seven of nine F1 offspring. Three out of four 2019 individuals (USKC 9, 11, 10) and one 2021 individual

were identified as having the same parent (NFC 84) with an accuracy of 94%. One 2019 individual (USKC 4), and the other two 2021 individuals (USKC 7 and 65) were identified as each having distinct parents. All individuals with < 40% of ancestry from the donor population failed to have any parents identified.

3.3 Population heterozygosity is similar across all nine refuge populations.

We found that all refuge populations with a sampling size > 8 had similar values of mean heterozygosity, ranging 0.21-0.27 (Figure 3-2A). The population with the lowest mean heterozygosity was Stairway Creek, which was significantly lower than all the other refuge populations except Cabin Creek, and the population with the highest mean heterozygosity was Upper Silver King Creek, which was significantly higher than most of the other refuge populations (Supplemental Table S.3-1). North Fork Cottonwood Creek also had significantly elevated population heterozygosity compared to most of the populations (Supplemental Table S.3-1). However, fish were collected from North Fork Cottonwood Creek in two different years, and we found mean heterozygosity decreased from 2017 to 2020. To detect change in mean heterozygosity after the translocation from North Fork Cottonwood Creek to Upper Silver King Creek, we compared non-hybridized individuals to hybridized individuals identified by the admixture analysis in USKC 2019 and 2021 and found no significant increase in mean heterozygosity in hybrid individuals (Figure 3-2B; P-value: USKC 2019 and hybridized individual: 0.94; USKC 2021 and hybrid individual: 0.83).

4. Discussion

4.1 Population Structure and Translocation History.

Results of our PCA, admixture analyses, and F_{ST} all support one another. When K = 2, out-of-basin populations form a distinct cluster from the within-basin populations, as all the out-of-basin populations are separated from each other and from all the within-basin populations that are clustered together along PC1 and PC2 (Figure 3-1 A, C). Similarly, the pairwise F_{ST} values between two within-basin populations were significantly lower than the pairwise F_{ST} values of two out-of-basin populations or one within-basin and one out-of-basin population (Figure 3-1 D, E) When K = 3, Sharktooth Creek separated from the other out-of-basin populations and formed another cluster, and Sharktooth Creek was also the most distant from the other populations along PC1 and PC2 (Figure 3-1 A, C). At the same time, Sharktooth Creek also had the highest F_{ST} values from all the other populations (Figure 3-1 D). When K = 4, Coyote Valley Creek separated from other within-basin populations, which was also concordant with patterns along PC1 and PC3 (Figure 3-1 B, C).

Observed genetic population structure supports recent translocation histories of fish in these systems to re-establish refuge populations after the removal of hybrids between PCT and non-native trout. Similarities among five within-basin populations can be explained by shared source populations and frequent translocations after multiple chemical treatments to remove non-native fish. After the introduced RT and PCT/RT hybrids were removed successfully by the chemical treatment in 1977, Fly Valley Creek was the direct or indirect source population used to re-establish Coyote Valley Creek, Corral Valley Creek, and Upper Silver King Creek. Coyote Valley Creek population was re-established from the Fly Valley Creek population after successive chemical treatments in 1964 and 1977, and 54 PCT were transferred from Fly Valley

Creek to Coyote Valley Creek in 1989 again. In 1994-1998, pure PCT were relocated back into Upper Silver King Creek from Coyote Valley Creek and Fly Valley Creek. After the successful chemical treatments in 1987 and 1988 in Corral Valley Creek, the source populations for the re-established population were also Fly Valley Creek and Coyote Valley Creek. These cycled donor-recipient relationships among within-basin populations homogenized any genetic divergence that might have arisen due to genetic drift in these small populations.

Compared to the within-basin populations, out-of-basin locations have comparatively fewer translocations. Cordes et al. (2004) and the PCT Recovery Plan (Alber 2020) provide a more detailed list of translocations, founding events, and chemical treatments 1912-2004. The North Fork Cottonwood Creek population was originally established using PCT from Upper Silver King Creek, Corral Valley Creek, and Coyote Valley Creek in 1946. North Fork Cottonwood Creek was then utilized as a source population to establish the other two out-of-basin populations, Cabin Creek and Sharktooth Creek in 1968. Stairway Creek was established in 1972 from 77 fish in Delaney Creek, which was previously sourced with 43 Four Mile Canyon Creek fish and three Fly Valley Creek fish. The admixture analysis also reflected the admixed ancestry between Stairway Creek and Four Mile Canyon Creek when K = 2 and closer genetic similarity of Stairway Creek and within-basin populations compared to the other three out-of-basin refuge populations, which also reflected the different source population used to establish Stairway Creek from the other three refuge populations.

4.2 Parentage analysis and post 2017 translocation genetic monitoring.

After 88 fish were translocated from North Fork Cottonwood Creek in 2017 to Upper Silver King Creek, we observed 12 F1 individuals collected in 2019 and 2021. We used parentage analysis to determine how many individuals from the donor population, North Fork

Cottonwood Creek, had spawned, but only seven F1s were assigned to North Fork Cottonwood Creek fish with high confidence. Interestingly, three 2019 and one 2021 F1 hybrids shared the same parent, which indicated that this individual mated more than once, or that one 2019 offspring was not sampled in 2019 but sampled in 2021. However, five fish from Upper Silver King Creek failed to have parents identified in the parentage analysis. One potential reason is that their parents are one of the 11 North Fork Cottonwood Creek individuals that did not pass our filtering threshold for inclusion. Moreover, we found most of the individuals that didn't assign to parents had lower admixed ancestry proportions from donor population, commonly ranging from 0.35 to 0.40. These individuals could be F2 hybrids as three of four fish with no parent identified are 2021 samples, and the one-year-old juveniles in 2019 (< 60 mm) can be already sexually mature and reproduce in 2021, according to the three-year life span for most individuals (Wong, 1975). As the whole population was not thoroughly sampled in 2019, it is likely that their parents were missed. Another alternative hypothesis is that the false negative results were generated by the reduction of the admixed proportion due to the unbalanced dam and sire genome inheritance caused by the recombination exhibited in 1,114 selected SNPs. Thus, the heterozygous sites caused by the hybridization were under-represented by the selected SNPs and resulted in no parent identified in the analysis.

Another concern for the parentage analysis was the false positive as the some of the individuals in the donor population also had the ancestry from the recipient population (range from 0%-31%) due to the shared polymorphism and demographic history. Thus, individuals that had the admixed ancestry proportion from 35%-40% can be the offspring of two translocated NFC individuals from North Fork Cottonwood Creek. However, although 1 NFC individual had 31% admixed ancestry from the recipient population, all other individuals had less than 25%

ancestry from the recipient population. On the other hands, PCA could distinguish the NFC individuals from the hybrid offspring individuals clearly, and no mismatches were found among PCA, admixture analysis, and parentage analysis. Thus, although the donor and recipient populations are not clearly separated by the admixture analysis, the false positive results in parentage analysis were less likely than false negative results caused by the filtering and sampling alone.

In theory, a sudden jump in population heterozygosity will be observed within a couple of generations following a translocation event, because new alleles were introduced from the donor population (NFC) to the recipient population (USKC). However, we did not observe such an increase in heterozygosity in Upper Silver King Creek fish at the time of sampling. There are many potential reasons for this result; the first being that it is possible that very few fish from North Fork Cottonwood Creek were able to successfully spawn in Upper Silver King Creek, thereby reducing the genetic rescue effect. We did not sample any fish in Upper Silver King Creek with two parents from North Fork Cottonwood Creek after translocation, as might be expected if survival of translocated fish was high, though this could also be explained by incomplete re-sampling of fish in Upper Silver King Creek. Secondly, North Fork Cottonwood Creek had similar heterozygosity at an individual and population level as Upper Silver King Creek, so it may have had limited ability to boost heterozygosity in Upper Silver King Creek. This finding was further supported by the insignificant difference in population heterozygosity between hybrids and non-hybrid individuals in Upper Silver King Creek (Figure 3-3). Though we did not find very many F1s in our samples, collectively the population heterozygosity in Upper Silver King Creek was the highest among all the refuge populations.

4.3 Identifying potential donor populations for future translocations.

Based on our analyses, all populations have similar population heterozygosity, and genetic differentiation was relatively weak, even if genetic differentiation was unequal among refuge populations. Translocations among populations can be an effective management strategy to boost population heterozygosity safely for PCT. At the same time, because we successfully detected F1 hybrid offspring after the translocation of fish from North Fork Cottonwood Creek to Upper Silver King Creek, we may conclude that genetic differentiation between within-basin and out-of-basin populations is not an absolute barrier for translocations between them. As the North Fork Cottonwood Creek is the donor population of Cabin Creek and Sharktooth Creek, Cabin Creek can also be a potential donor population. More caution is needed for using Sharktooth Creek as a donor population because it has the highest genetic differentiation from all the other populations. Stairway Creek was the out-of-basin population that was genetically relatively similar to the within-basin populations and had the lowest population heterozygosity. Stairway Creek should not be used for translocations unless managers can ensure a translocation can bring new genetic variation to the recipient population. On the other hand, census size of the refuge population should also be considered for the sustainability of each refuge population as translocating the adults at spawning age out of the refuge population could compromise the reproductive capacity of the population, which may ultimately reduce the size and genetic diversity of the donor population.

The translocation in Westlope cutthroat trout (*Oncorhynchus clarkii lewisi*) can serve as a guideline for the future translocation of PCT. Westlope cutthroat trout is also a subspecies of cutthroat trout that has widespread fragmented habitats, resulting in the decrease in the genetic variation in each population. Kovach et al. 2022 confirmed that the assisted gene flow generated

by the translocation successfully increased the genetic variation in the recipient populations. However, the effectiveness of assisted gene flow created by the translocation is highly dependent on the between-population genetic variation as a donor population will bring more genetic variation if it is more differentiated from a recipient population. However, PCT has more limited between-population genetic variation. Thus, the expected increase in genetic variation will also be lower than Westlope cutthroat trout. In Kovach et al. 2022, they translocated the donor population that has high genetic differentiation in a very small migration rate to the recipient population, which can limit the potential outbreeding depression, if it happens, in a small proportion of the population. Sharktooth Creek has the highest genetic differentiation to all the other refuge populations, and the assisted gene flow in a small migration rate can allow the increase of genetic variation in PCT in an effective and relatively safe way.

Translocation is an important strategy to save PCT and similar species from the loss of genetic diversity and population divergence, but conservation managers often have limited time and resources to make decisions. Initial monitoring after the 2017 NFC translocation suggested a high feasibility to the translocation. However, diversity metrics for each of the refuge populations may change over time. We recommend continued genetic monitoring to further evaluate translocation events, both with our subset of markers on a frequent basis, but also with reduced representation sequencing or whole genome resequencing on an occasional basis. This approach will enable managers to evaluate population responses to environmental and demographic factors (e.g., drought, fire, bottlenecks, etc.) and make more informed decisions for future management actions.

5. Conclusion

Overall, we used 1,114 SNPs selected by Su et al. (In prep) to quantify genetic population structure and population heterozygosity on all nine refuge populations. Within-basin populations are relatively genetically similar, while out-of-basin populations are more differentiated from each other. Heterozygosity levels among all nine refuge populations are apparently quite similar to each other. Upper Silver King Creek has the highest population heterozygosity (0.27) while Stairway Creek has the lowest population heterozygosity (0.21). At the same time, we also used these SNPs to monitor the change of the genetic diversity and parent-offspring relationships in Upper Silver King Creek after 88 PCT individuals were translocated from North Fork Cottonwood Creek. We found 12 individuals are the hybrid offspring that have 35-60% genomic component from the North Fork Cottonwood Creek, and the parents of seven offspring were successfully identified. However, we didn't detect any significant difference of the individual heterozygosity between the hybrid offspring and all the other individuals in 2019 and 2021. These results will aid conservation management of PCT, and will assist in establishing a science-based framework for assessing the efficacy of future conservation actions.

6. References

Alber, L. 2020. Paiute cutthroat trout reintroduction. Natural Resource Agency, Heritage and Wild Program, editor.

Ali, O. A., and coauthors. 2016. RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. Genetics 202(2):389-400.

Behnke, R. J. 1965. A systematic study of the family Salmonidae with special reference to the genus *Salmo*. University of California, Berkeley, California.

Caughley, G. 1994. Directions in Conservation Biology. Journal of Animal Ecology 63(2):215-244.

Cordes, F. J., A. J. Israel, and M. Bernie. 2004. Conservation of Paiute cutthroat trout: the genetic legacy of population transplants in an endemic California salmonid. California Fish and Game 90(3):101-118.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14(8):2611-20.

Finger, J. A., B. Mahardja, and B. May. 2012. Genetic management plan for the Paiute cutthroat trout (*Oncorhynchus clarkii seleniris*). G. V. Lab, editor.

Furlan, E. M., and coauthors. 2020. Assessing the benefits and risks of translocations in depauperate species: A theoretical framework with an empirical validation. Journal of Applied Ecology 57(4):831-841.

Gilk, S. E., and coauthors. 2004. Outbreeding Depression in Hybrids Between Spatially Separated Pink Salmon, *Oncorhynchus gorbuscha*, Populations: Marine Survival, Homing ability, and Variability in Family Size. Environmental Biology of Fishes 69(1):287-297.

Harig AL and Fausch KD. 2002. Minimum habitat requirements for establishing translocated cutthroat trout populations Ecological Applications 12: 535-551.

Hedrick, P. W., R. C. Lacy, F. W. Allendorf, and M. E. Soulé. 1996. Directions in Conservation Biology: Comments on Caughley. Conservation Biology 10(5):1312-1320.

Hemstrom, W., and M. Jones. 2023. snpR: User friendly population genomics for SNP data sets with categorical metadata. Molecular Ecology Resources 23(4):962-973.

Jones, O. R., and J. Wang. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10(3):551-555.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14(6):587-589.

Kassambara, A. 2020. Pipe-Friendly Framework for Basic Statistical Tests [R package rstatix version 0.6.0].

Kolodny, O., and coauthors. 2019. Reconsidering the management paradigm of fragmented populations. bioRxiv:649129.

Kovach RP, Leary RF, Bell DA*, et al.* 2022. Genetic variation in westslope cutthroat trout reveals that widespread genetic rescue is warranted Canadian Journal of Fisheries and Aquatic Sciences 79: 936-946.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987-2993.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-60.

Meisner, J., and A. Albrechtsen. 2018. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210(2):719-731.

Saglam, I. K., and coauthors. 2017. Genomic Analysis Reveals Genetic Distinctiveness of the Paiute Cutthroat Trout *Oncorhynchus clarkii seleniris*. Transactions of the American Fisheries Society 146(6):1291-1302.

Skotte, L., T. S. Korneliussen, and A. Albrechtsen. 2013. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics 195(3):693-702.

Stead JE, Boucher VL, Moyle PB, Rypel AL. 2022. Growth of Lahontan cutthroat trout from multiple sources re-introduced into Sagehen Creek, CA. PeerJ 10:e13322 https://doi.org/10.7717/peerj.13322

Team, R. C. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena, Austria.

Titus, R.G., and L. Calder. 2009. Age and growth of Paiute cutthroat trout, *Oncorhynchus clarkii seleniris*, in the Silver King Creek drainage, California. Technical Completion Report, California Department of Fish and Game, Sacramento, California. 15 pp.

USFWS. 1967. Endangered Species List - 1967. Federal Register.

USFWS. 2004. Revised recovery plan for the Paiute cutthroat trout (Oncorhynchus clarkii seleniris). Portland, Oregon.

Weeks, A. R., and coauthors. 2011. Assessing the benefits and risks of translocations in changing environments: a genetic perspective. Evolutionary Applications 4(6):709-725.

Whiteley, A. R., S. W. Fitzpatrick, W. C. Funk, and D. A. Tallmon. 2015. Genetic rescue to the rescue. Trends in Ecology & Evolution 30(1):42-49.

Wickham, H. 2016. Ggplot2: Elegant graphics for data analysis. Springer International Publishing, Cham, Switzerland.

Wong, D. M. 1975. Aspects of the life history of the Paiute cutthroat trout, Salmo clarki seleniris Snyder, in North Fork Cottonwood Creek, Mono County, California, with notes on 34

behavior in a stream aquarium. M.S. thesis, California State University, Long Beach, California.

7. Figures and Tables

FIGURE 3-1. Genetic population structure of PCT using selected SNPs. PCAs of all samples, with color coding population and shape coding year. (A) PC1 and PC2, (B) PC1 and PC3. (C)Range wide admixture analysis of all the samples from K=2 to K = 4. (D) Pairwise FST values between refuge populations or different sampling years in the same refuge population. (E) The comparison of the FST values between two within-basin populations (In ~ In), two out-of-basin populations (Out ~ Out), and one within-basin and one out-of-basin population (In ~ Out). P-values denote results from Wilcoxon signed rank tests for each pair.

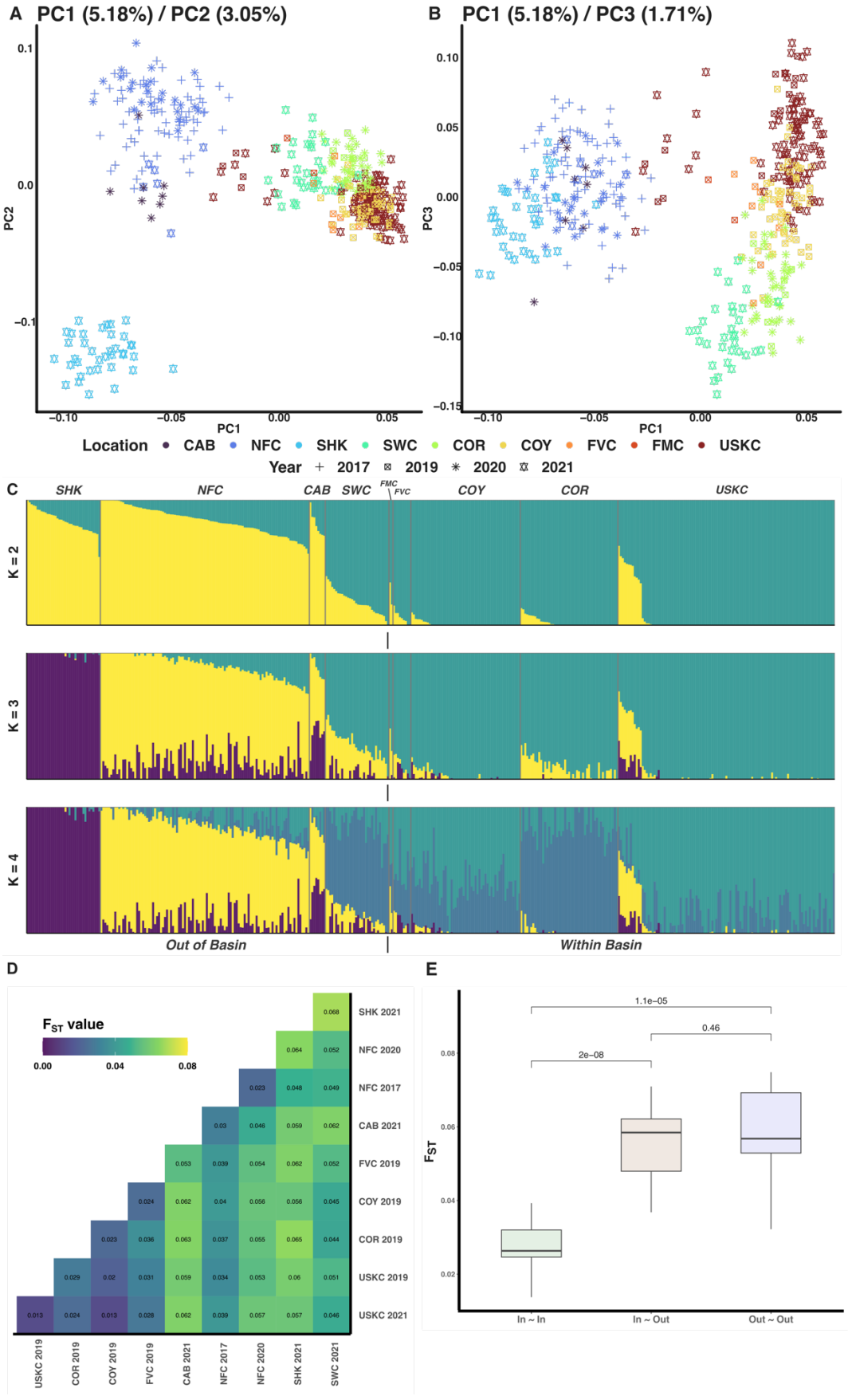

A PC1 (5.18%) / PC2 (3.05%)
B PC1 (5.18%) / PC3 (1.71%)
PC1
PC2
PC3
Location CAB NFC SHK SWC COR COY FVC FMC USKC
Year 2017 2019 2020 2021
C
SHK NFC CAB SWC FMC FVC COY COR USKC
K = 2
K = 3
K = 4
Out of Basin
Within Basin
D
FST value
0.00 0.04 0.08
SHK 2021
NFC 2020
NFC 2017
CAB 2021
FVC 2019
COY 2019
COR 2019
USKC 2019
USKC 2021
SWC 2021
E
FST
1.1e−05
0.46
2e−08
In ~ In
In ~ Out
Out ~ Out

FIGURE 3-2. Parent-offspring relationship within USKC/NFC after the translocation. (A) Admixture analysis of USKC/NFC when K =2. Blue represents ancestry in the donor population, and yellow represents ancestry in the recipient population. (B) The parent-offspring relationship between the USKC/NFC. Fish with the same color represents parent-offspring relationships. Accuracy of the parentage assignment is labeled above the fish while proportion of admixture from the donor population is labeled below the fish. Fish in grey denote hybrids with no parent identified.

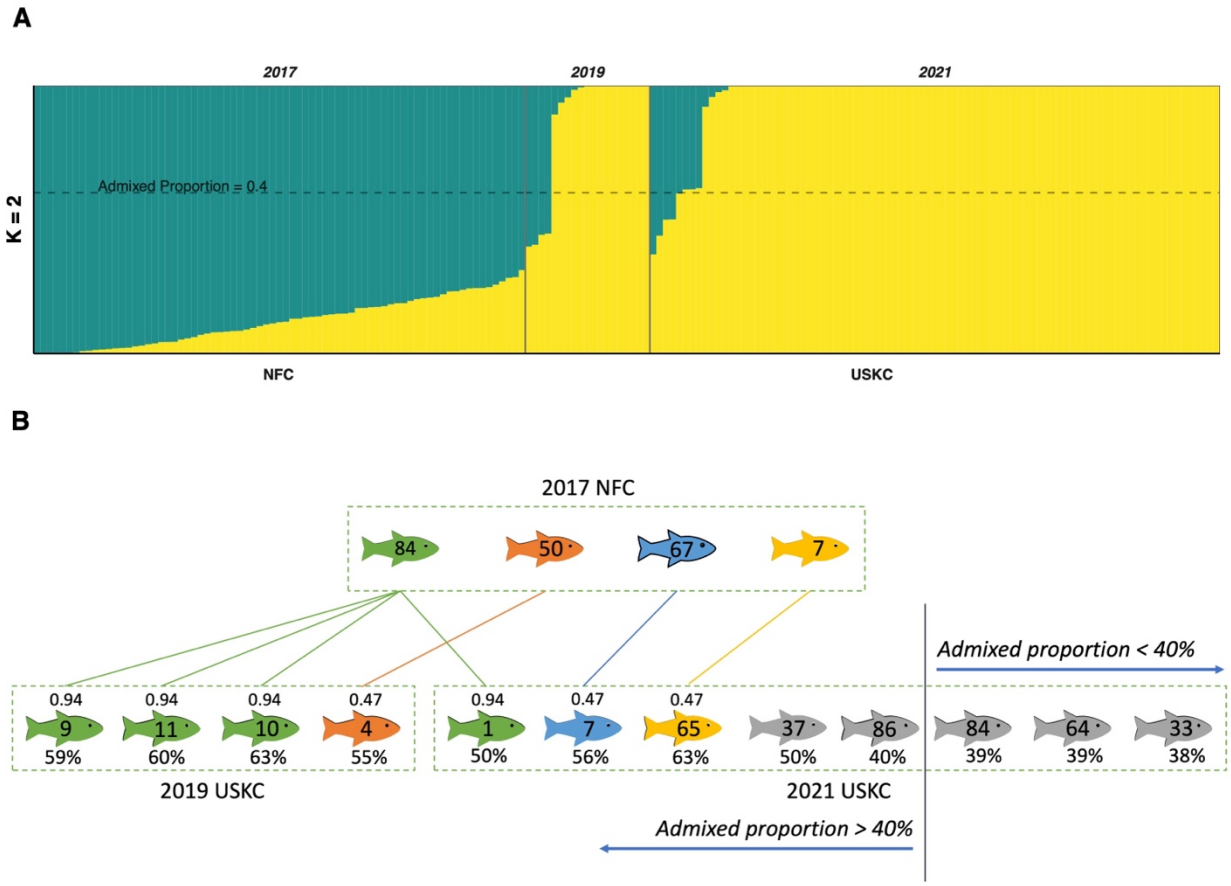


FIGURE 3-3. Heterozygosity estimates for each PCT refuge population. (A) Distribution of individual heterozygosity of each refuge population. (B) Distribution of individual heterozygosity in non-hybrid individuals and hybrids following translocation. P-values listed for each comparison reflect the Bonferroni adjustment.

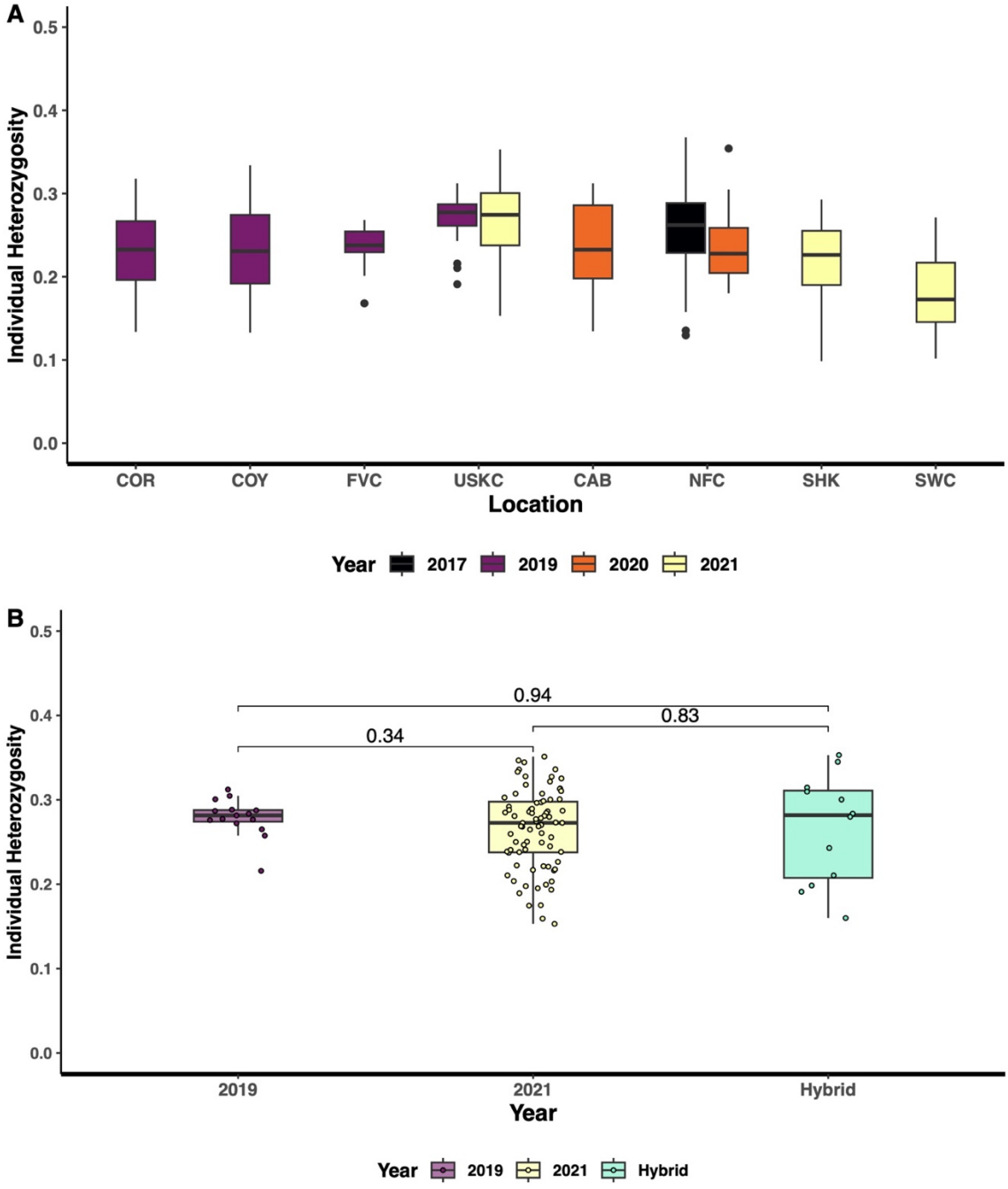
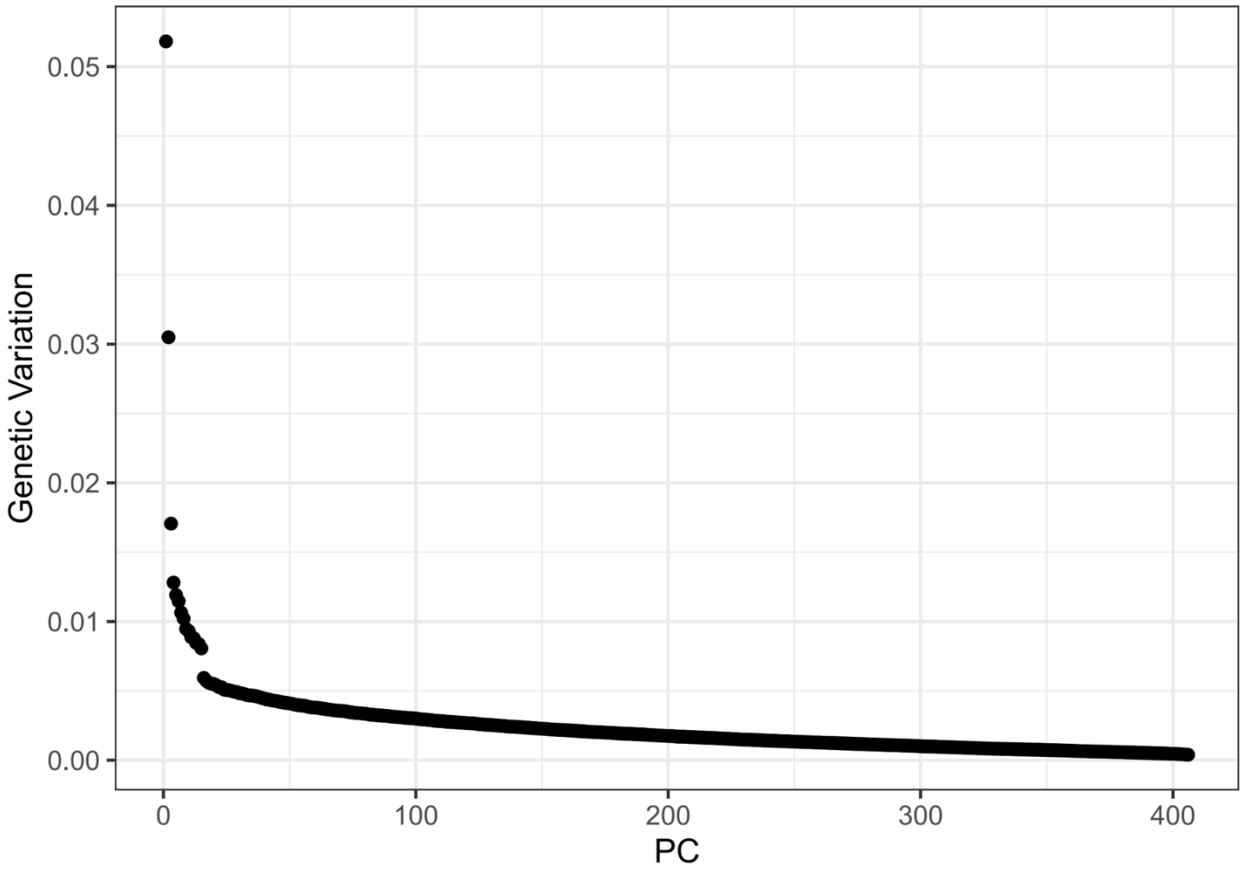


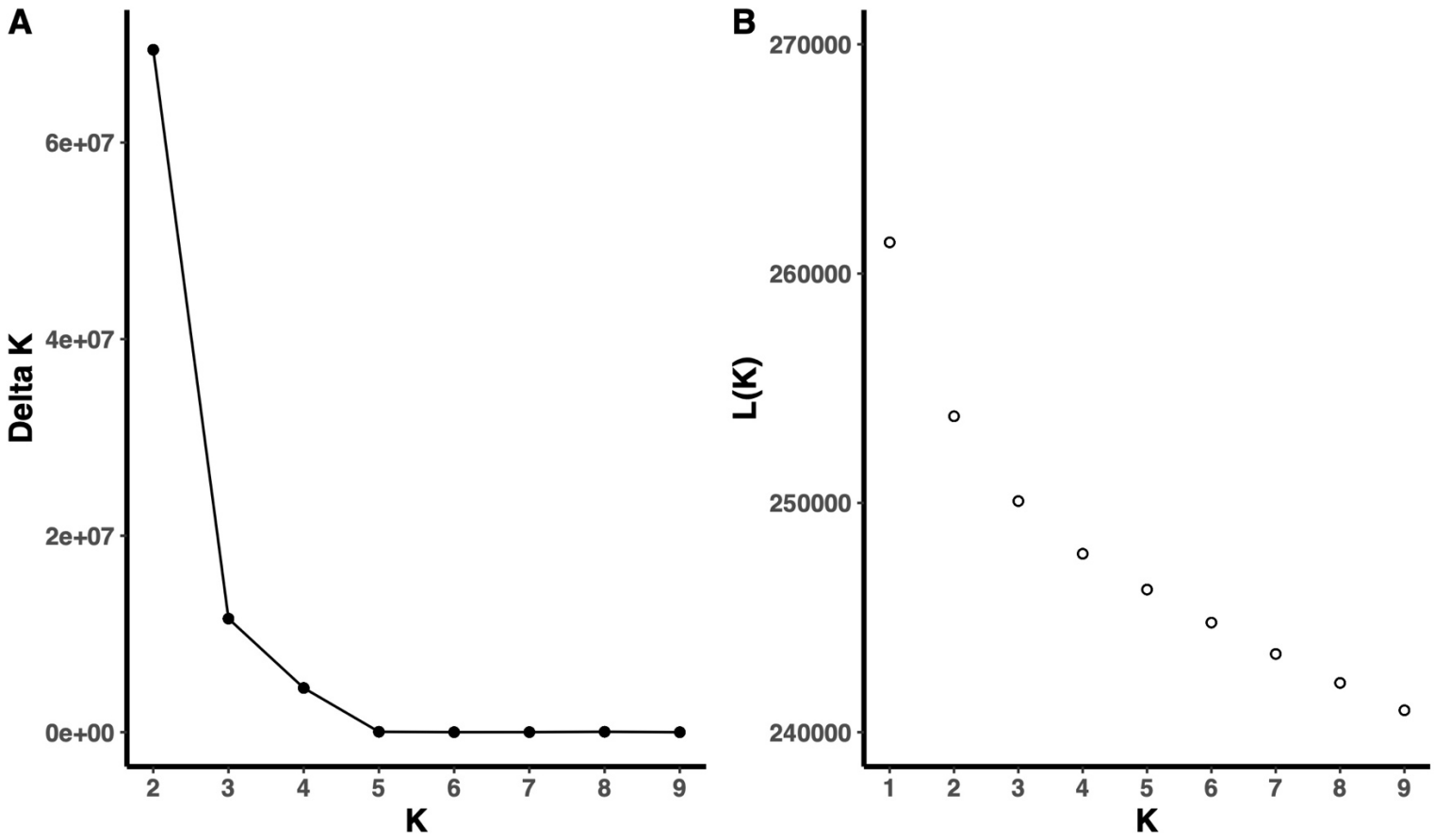
TABLE 3-1. Details on samples used in analyses of population heterozygosity including sampling location, sample size before/after mapping and sequencing quality control, and the mean and variance of individual heterozygosity. NFC 2020 and 2021, and COR 2019 and 2020 were merged into one population separately (labeled as NFC 2020 and COR 2019 in the analyses) because of low sampling size in NFC 2021 and COR 2020.

Basin	Type	Location	Sample collection year	Sample Size	Sample size after filtering	Mean	Variance
Carson River Basin	within-basin	Corral Valley Creek (COR)	2019	16	10	0.236	0.002
			2020	44	49	0.23	0.0022
		Coyote Valley Creek (COY)	2019	73	55	0.232	0.0025
		Four Mile Canyon Creek (FMC)	2019	2	2		NA
		Fly Valley Creek (FVC)	2019	10	9	0.234	0.001
		Upper Silver King Creek (USKC)	2019	20	19	0.269	0.00011
			2021	115	87	0.267	0.0023
Great Basin	out-of-basin	Cabin Creek (CAB)	2020	8	8	0.235	0.0035
		North fork Cottonwood Creek (NFC)	2017	86	75	0.258	0.002
			2020	35	19	0.234	0.0018
			2021	6	6		
San Joaquin Basin		Sharktooth Creek (SHK)	2021	39	37	0.241	0.002
		Stairway Creek (SWC)	2021	41	32	0.211	0.0019

SUPPLEMENTAL FIGURE S.3-1. Proportion of genetic variation explained by multiple PCs. Each point represents a PC, and the first three PCs explain more genetic variation than the latter PC axes.



SUPPLEMENTAL FIGURE S.3-2. Optimal K selected by Evanno method. (A) Delta K was calculated from K = 2 to K = 9. (B) Abstract value of mean log likelihood of each K calculated from K = 1 to K = 9.



SUPPLEMENTAL TABLE S.3-1.The pairwise comparison of population heterozygosity in each population each year. The sample size of population 1 (column 1) and population 2 (column 2) were recorded in sample size 1 and sample size 2, respectively. P-values, adjusted P-values, and significant levels after adjustment were also recorded for each comparison. P-value cutoffs: ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001.

	Population 1	Population 2	**Sample Size 1**	**Sample Size 2**	**P-value**	**Adjusted P-value**	**Significance level of adjusted P-value**
1	CAB 2020	COR 2019	16	98	0.367	1	ns
2	CAB 2020	COY 2019	16	110	0.397	1	ns
3	CAB 2020	FVC 2019	16	18	0.251	1	ns
4	CAB 2020	NFC 2017	16	150	0.006	0.27	ns
5	CAB 2020	NFC 2020	16	48	0.673	1	ns
6	CAB 2020	SHK 2021	16	74	0.748	1	ns
7	CAB 2020	SWC 2021	16	64	0.002	0.09	ns
8	CAB 2020	USKC 2019	16	38	0.008	0.36	ns
9	CAB 2020	USKC 2021	16	174	0.001	0.045	*
10	COR 2019	COY 2019	98	110	0.999	1	ns
11	COR 2019	FVC 2019	98	18	0.763	1	ns
12	COR 2019	NFC 2017	98	150	0.000139	0.006255	**
13	COR 2019	NFC 2020	98	48	0.345	1	ns
14	COR 2019	SHK 2021	98	74	0.02	0.9	ns
15	COR 2019	SWC 2021	98	64	9.36E-11	4.212E-09	****
16	COR 2019	USKC 2019	98	38	0.002	0.09	ns
17	COR 2019	USKC 2021	98	174	3.5E-08	1.575E-06	****
18	COY 2019	FVC 2019	110	18	0.813	1	ns
19	COY 2019	NFC 2017	110	150	0.000176	0.00792	**
20	COY 2019	NFC 2020	110	48	0.424	1	ns
21	COY 2019	SHK 2021	110	74	0.025	1	ns
22	COY 2019	SWC 2021	110	64	1.01E-10	4.545E-09	****
23	COY 2019	USKC 2019	110	38	0.002	0.09	ns
24	COY 2019	USKC 2021	110	174	9.13E-08	4.1085E-06	****
25	FVC 2019	NFC 2017	18	150	0.007	0.315	ns
26	FVC 2019	NFC 2020	18	48	0.592	1	ns
27	FVC 2019	SHK 2021	18	74	0.179	1	ns
28	FVC 2019	SWC 2021	18	64	4.73E-06	0.00021285	***
29	FVC 2019	USKC 2019	18	38	0.000212	0.00954	**
30	FVC 2019	USKC 2021	18	174	0.000489	0.022005	*
31	NFC 2017	NFC 2020	150	48	3.89E-05	0.0017505	**

32	NFC 2017	SHK 2021	150	74	4.68E-09	2.106E-07	****
33	NFC 2017	SWC 2021	150	64	1.12E-19	5.04E-18	****
34	NFC 2017	USKC 2019	150	38	0.685	1	ns
35	NFC 2017	USKC 2021	150	174	0.036	1	ns
36	NFC 2020	SHK 2021	48	74	0.281	1	ns
37	NFC 2020	SWC 2021	48	64	3.92E-08	1.764E-06	****
38	NFC 2020	USKC 2019	48	38	5.08E-05	0.002286	**
39	NFC 2020	USKC 2021	48	174	1.69E-07	7.605E-06	****
40	SHK 2021	SWC 2021	74	64	7.59E-06	0.00034155	***
41	SHK 2021	USKC 2019	74	38	8.68E-07	3.906E-05	****
42	SHK 2021	USKC 2021	74	174	1.53E-12	6.885E-11	****
43	SWC 2021	USKC 2019	64	38	5.37E-13	2.4165E-11	****
44	SWC 2021	USKC 2021	64	174	9.16E-22	4.122E-20	****
45	USKC 2019	USKC 2021	38	174	0.095	1	ns