**Methods**

We identified a set of diversity loci by separating California Lahontan trout into five groups: Truckee, Carson, Walker, Heenan Lake and Independence Lake. These groups reflect natural geographic structure and management units present in California Lahontan trout. As genetic loci identified as variable are influenced by individuals included and Minor Allele Frequencies (MAFs), we reduced the total samples analyzed to produce diversity loci to 20 individuals from each of the five groups. The 20 individuals from each unit were selected by omitting those with evidence of rainbow trout genetics and then selecting individuals closest to the median coverage of the total data set (generation of alignments previously described). Individuals selected were from across Truckee, Carson and Walker units, Heenan Lake individuals collected in 2010 and Independence Lake individuals in 2016 (Table SX).

The alignments of the 100 individuals were analyzed with ANGSD (Korneliussen et al., 2014) to call SNPs. Quality of called SNPs was ensured with a minimum number of 95% of individuals specified (-minInd 95), a minimum mapping and minimum base quality of 20 (-minMapQ 20 -min Q 20), a significance value for SNPs (-SNP\_pval 1e-6) and a posterior cutoff value of 0.95 (-postCutoff 0.95). Additional command line options specified were -GL 1, -doMajorMinor 1, -doMaf 1, -doPost 1, -doPlink2 and only chromosomes of the assembly were used. With default settings, ANGSD calls only biallelic SNPs. The resulting PLINK-formatted file was converted to a VCF-formatted file with PLINK (Purcell et al., 2007). We applied filters with VCFTools to restrict MAFs for each SNP to between 0.10 and 0.40 and removed loci out of Hardy-Weinberg Equilibrium (HWE) at a significance value of 0.05 (--hwe 0.05) (Danecek et al., 2011). We then produced a set of unlinked SNPS by removing SNPS with a R2 measurement of linkage disequilibrium greater than 0.1 within 10,000 bp of each other.

With this resulting ‘diversity SNP data set’ we examined the distribution of loci across chromosomes and verified that they were not physically close on each chromosome. The diversity SNP data set was imported into R with *vcfR* package (Knaus and Grünwald, 2017) and variance in the data set assessed with a Principal Component (PC) analysis. The covariance matrix was made with the dudi.pca function of the *adegenet* package (Jombart, 2008) with missing data filled with the mean of the locus. Evidence for neutral genetic structure with further assessed the iterative *k*-means clustering with the find.clusters() function of the *adegenet* package.

We investigated the diversity SNP data set with tools with the *snpR* package in R (<https://github.com/hemstrow/snpR>) for further evidence of genetic structing by generating estimates of pairwise Fst and calculated nucleotide diversity (theta pi).

**Results**

The aligned samples we obtained from individuals representing the five California management units (n=467) varied from a minimum number of filtered reads of 248,690 to a maximum of 5,696,479 with a mean of 1,814,644 and median of 1,603,772 with a standard deviation of 914,340. The 100 individuals selected for identification of diversity loci ranged from 888,018-2,301,530 filtered reads, with a mean of 1,542,032 median of 1,555,032, maximum of 2,301,530 and a standard deviation of 203,413. The initial genotype call with ANGSD produced 11,873 variants across chromosomes, filtering for a MAF of 0.10 to 0.40, removal of SNPs deviating from HWE and pruning of linked SNPs yields 485 SNPs (**Figures X and Y**). The mean MAF is 0.14 and median is 0.13.

Principal Component analysis separated the 100 individuals into four groups with the first two PCs (PC1 9.55% of variance, PC2 7.50% of variance), these groups are the Truckee, Carson, Walker and a combined Heenan Lake and Independence Lake group (**Figure Z**). Principal Component 3 (5.68% of variance) separates Walker and Carson samples from a larger grouping of Trucker, Independence Lake and Heenan Lake samples. Evidence for genetic structuring in the dataset through *k-*means clustering with the find.clusters() function (number of PCs included is 85) indicates clear support for *K* = 4 genetic clusters under the Bayesian Information Criterion (BIC), consisting of the Trucker, Carson, Walker and a combined Heenan Lake and Independence Lake genetic cluster (**Figure XX**).

Pairwise Fst values are the lowest between Heenan Lake and Independence Lake (0.011) and highest between the Truckee and Walker (0.141) and are presented in **Table X**. Nucleotide diversity (theta pi) is highest in Heenan Lake (median per-chromosome theta pi = 0.27) and lowest in Walker (0.13), with Carson intermediate (**Figure YY**).

**Tables and Figures**

**Table X** Pairwise Fst values from the diversity data set estimated between units. Lake is abbreviated L.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Carson** | **Walker** | **Heenan L.** | **Independence L.** |
| **Truckee** | 0.097 | 0.141 | 0.077 | 0.086 |
| **Carson** | - | 0.104 | 0.075 | 0.086 |
| **Walker** | - | - | 0.122 | 0.126 |
| **Heenan L.** | - | - | - | 0.011 |

**A graph showing the number of chromosome

Description automatically generated**

Figure X: Distribution of diversity loci across chromosomes in rainbow trout genome assembly.



Figure Y: Distribution of Minor Allele Frequencies across the diversity data set.

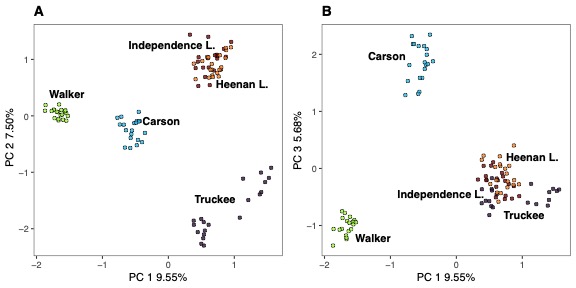


Figure Z: Principal Component (PC) analysis of the diversity data set showing the first three principal components. PC 1 and PC 2 are shown in panel A and PC 1 and PC 3 are shown in panel B. Lake is abbreviated L.



Figure XX: Bayesian Information Criterion (BIC) versus *K* genetic clusters from the diversity data set supporting *K* = 4 genetic clusters consisting of Truckee, Carson, Walker and combined Heenan Lake and Independence Lake.



Figure YY: Nucleotide diversity (Theta Pi) across the five units.

**Bibliography**

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Knaus, B.J., Grünwald, N.J., 2017. vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources 17, 44–53. https://doi.org/10.1111/1755-0998.12549

Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinform 15, 356. https://doi.org/10.1186/s12859-014-0356-4

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559–575. https://doi.org/10.1086/519795