**Task 3: Examining domestication selection in the conservation hatchery for Delta Smelt**

*Background*

Delta smelt have been cultured in a conservation hatchery system at the Fish Conservation and Culture Lab at UC Davis (FCCL). The FCCL uses a pedigree-based breeding system that applies parentage analysis and the addition of wild fish to keep the hatchery fish genetically similar to the wild population (Fisch et al. 2013; Lindberg et al. 2013). The pedigree-based management applied at the FCCL aims to both minimize average co-ancestry (i.e., kinship) in the refuge population and maintain their genetic diversity. Despite the intense management at the FCCL, Finger et al.2 showed strong evidence of genetic adaptation to captivity in the hatchery population. For example, offspring from crosses with two cultured parents with high levels of hatchery ancestry (e.g., a high domestication index; DI) are statistically more likely to survive to adulthood than those with wild parents or cultured-born parents with low hatchery ancestry (low DI), and this trend has increased over time. At present, the number of wild fish that are available to incorporate as broodstock in the FCCL refuge population is nearly zero, and without this wild input the rate of domestication will increase. Understanding the mechanism behind hatchery adaptation will allow for the FCCL to adjust their protocol to minimize the rate of hatchery adaptation, and thus improve the likelihood that future wild population supplementation will be successful.

We hypothesized: 1) there are heritable genetic changes associated with domestication selection in the FCCL, 2) domestication adaptation happens during the first few generations in the hatchery, and 3) the locus or loci under domestication selection contribute to the reproductive success of smelt in the hatchery.

*Experimental design*

At the UC Davis Genomic Variation Lab, we accessed archived FCCL delta smelt samples (fin clips) from eight generations from 2008-2015. We grouped 240 individuals (not related at least at F1) for sequencing based on their domestication index (DI) and offspring survival rate (Figure 1). DI is calculated based on the number of generations that an individual's genome spent in captivity and is equal to the average of parents DI plus 1 (because another year has passed in captivity). Based on this, smelt are grouped into low (0 < DI ≤ 2), medium (3 ≤ DI ≤ 4), and high (DI≥7) DI groups. We created this study design in part to examine how quickly adaptation is happening in captivity. At the FCCL, the rate of survival of offspring from a family (Offspring survival rate of a family) serves as a proxy for fitness in captivity and is calculated based on the total number of an individual offspring from a family that survive to be tagged at adulthood. The relative reproductive success was normalized by taking the ratio of each individual’s offspring survival rate to the mean of that in the corresponding generation and grouped as low (≤ 0.3) and high (≥ 0.7).

*Summary of previous results*

Previously, we reported some results based on the incomplete genome assembly. To do that, we aligned and conducted population genetic analyses from the four groups of delta smelt with varying DI level: wild, low, medium, and high. Restriction-site associated DNA (RAD) sequencing by PstI enzyme was used based on the protocol described in Ali et al.1 to collect genomic data. In order to detect outlier SNPs associated with domestication selection, genome-wide SNP FST (FST calculated per each SNP across genome) was calculated between wild and the different DI groups, FST is a measure of genetic differences between groups.

We first produced a RAD sequencing derived partial assembly (RAD assembly) and reported some outlier SNPs (90 SNPs) as candidate loci associated with domestication selection. Next, we produced a linked-read draft assembly that was orders of magnitude more contiguous than the previous RAD sequencing derived assembly. We found several FST peaks across the genome as candidate loci under domestication selection. Although with the increase in the reference genome contiguity, candidate loci appeared as elevated peaks, we were still observing multiple peaks across the genome. Because the draft assembly was not yet at chromosome level contiguity, we argued that with the chromosome-level reference genome, we may find that the peaks currently located on different scaffolds may belong to the same region. If the peaks located on different scaffolds come from one region, it would signify domestication at the hatchery is a monogenic trait (a trait that is determined by alleles of a single gene) under selection. If the FST peaks continue to be distributed at different locations spread across the genome, it would signify that domestication is a polygenic trait (a trait that is caused by effects of many genes) and multiple genes or loci throughout the genome are under selection.

Since then, we have produced a high-quality draft genome assembly to explore genomic signature of domestication selection and whether or not it is a monogenic or polygenic trait. The following is our updated analysis and results, and interpretation.

*Methods*

First, we aligned from our four groups sequences to the reference genome using BWA software. The output files from the alignments were Sequence Alignment Map (SAM) files, which were then converted to Binary Alignment Map (BAM) files using SAMtools4. SAMtools was then used to sort reads by name, remove reads that did not have a paired end, remove PCR duplicates, and index the BAM files. After aligning to the draft assembly, we began our analyses. All population genetic analyses were conducted using ANGSD3 which analyzes raw RAD sequencing data based on a probabilistic framework in the form of genotype likelihoods (GL). For the analyses, we used ANGSD’s implementation of a SAMtools genotype likelihood model (-GL 1) with a minimum base quality of 20 (-minQ 20) and minimum mapping quality of 20 (-minMapQ 20).

We wanted to standardize aligned read counts across individuals to account for variability that could bias downstream analysis. To do this, we used SAMtools-view to examine alignment quality, and observed high variance in quality and number of aligned reads. Due to the very high variability we observed (~20,000 – ~15,000,000) within individuals from the four groups, we evaluated the distribution of aligned-read counts and genotype call counts for individuals in each of the four groups (Figure 1). We then used this distribution to choose an aligned-read per individual ranging between 3,000,000bp to 6,000,000bp. Accordingly, individuals aligned-read counts totaling lower than our preferred range were removed, and individuals with aligned-read counts higher than our range had their reads subsampled using SAMtools-view.

As a result of the aligned read count filtration, the number of individuals in each group was reduced (Table 1). This resulted in relatively low sample numbers in the low and high DI groups compared to medium and wild. Because substantial differences in group sample sizes will affect the significance in FST pairwise comparisons, we decided to group individuals and make two comparisons: 1) wild/low DI group vs. medium/high DI group, and 2) wild group vs. hatchery (low/medium/high DI) group. With the former, we expected to detect loci that changed gradually in captivity while with the latter we expected to detect loci that change rapidly in captivity.

Next, we performed a pairwise FST sliding window analysis to detect SNPs associated with domestication selection. Sliding window FST is the FST analysis performed interactively over a certain length along a sequence or chromosome5. Since our SNP FST analyses were very noisy, and also because when selection affects a single SNP it also affects region where it is located, we decided to conduct a sliding window analysis which examines SNP outliers within their genomic regions. To conduct the sliding window analysis, we used snpR package6 which is a software package designed for whole-genome analysis with SNP data. In snpR, we performed FST sliding window function (calc\_smoothed\_averages) using 200kb window length (sigma=200, step = 50) to find differentiated regions on the genome. We then applied the bootstrapping function (do\_bootstraps) to test for the significance of those regions (at p≤0.001).

*Results and discussion*

The resultsof our twopairwise FST sliding window analyses are shown in figure 2. In the both comparisons (1 and 2), there are highly and moderately elevated regions with significant FST windows. Those elevated regions, have high genetic differences between the two groups in each comparison, and are candidate regions associated with domestication selection. In comparison 1, wild/low DI group vs. medium/high DI group, two elevated regions are significant: one on chromosome 15 and one on chromosome 22 (lg 15 and lg22). In comparison 2, wild vs. low/medium/high DI group, there are multiple significant peaks on chromosomes 4, 10, 13, 15, 16, 19, 22, 23, and 26. Interestingly, the two elevated regions on chromosomes 15 and 22 are present in the both comparisons which makes them more likely as candidate regions under the selection.

To further explore those candidate regions, we looked at the distribution of the windows across each chromosome. We observed that outlier windows clustered in one region on each chromosome (Figure 3).This strengthens the probability of the region being under selection (i.e. the existence of peak of windows around significant windows indicates regional selection). Moreover, looking at the high resolution of the regions in the both comparisons (Figure 3a and b) shows the same pattern of the peak distribution in each of the comparisons (comparison 1 in red and comparison 2 in blue) which is strong evidence that they are located at the exact same region on the genome and are under the same selective pressure.

Our analyses have provided strong evidence that heritable genetic changes are happening in captivity. We identified at least two regions on two different chromosomes as candidate regions under domestication selection. This can signify that domestication is a polygenic (i.e. domestication is caused by more than one region in the genome). In addition, since each of the two candidate regions are present in the both comparisons, it seems domestication selection is happening in the early generations. This is because even when the low DI group is separated from the wild group in comparison 2, those regions on lg15 and lg22 are elevated. Additionally, in comparison 2 other elevated regions show significant FST windows. Although these regions can be potentially under domestication selection, more investigation is required.

*Next steps*

For further research, we will explore genes and molecular pathways associated with the two candidate regions discovered in comparison 1 and we will do more investigation of the other regions discovered in comparison 2 that may potentially be under domestication selection. Using the reference genome, we will also run a genome-wide association study (GWAS) to assess association of our candidate loci with the recovery rate at the FCCL. If there is an association between domestication selection and recovery rate, we expect to observe the same loci as the domestication selection candidate loci.

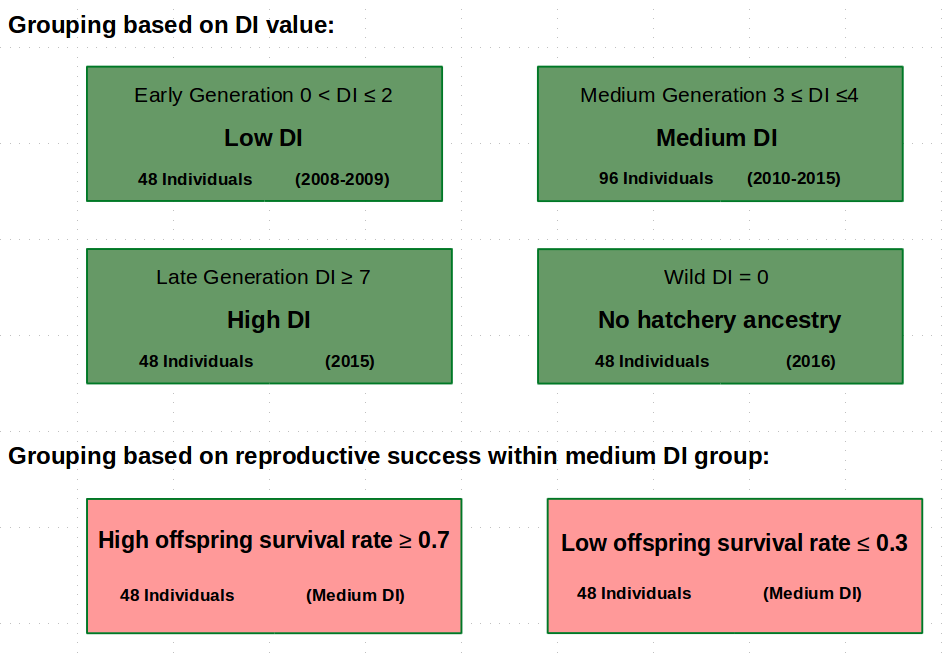
**

Figure 1. Experimental design. DI = Domestication Index. Green boxes show the four groups based on their DI value: low DI (0 < DI ≤ 2, 48 individuals), medium DI (3 ≤ DI ≤ 4, 96 individuals), high DI (DI≥7, 48 individuals), wild (DI = 0, 48 individuals). Pink boxes show two groups based on their recovery rate within the medium DI group: high offspring recovery rate (≥ 0.7, 48 individuals) and low offspring recovery rate (≤ 0.3, 48 individuals).

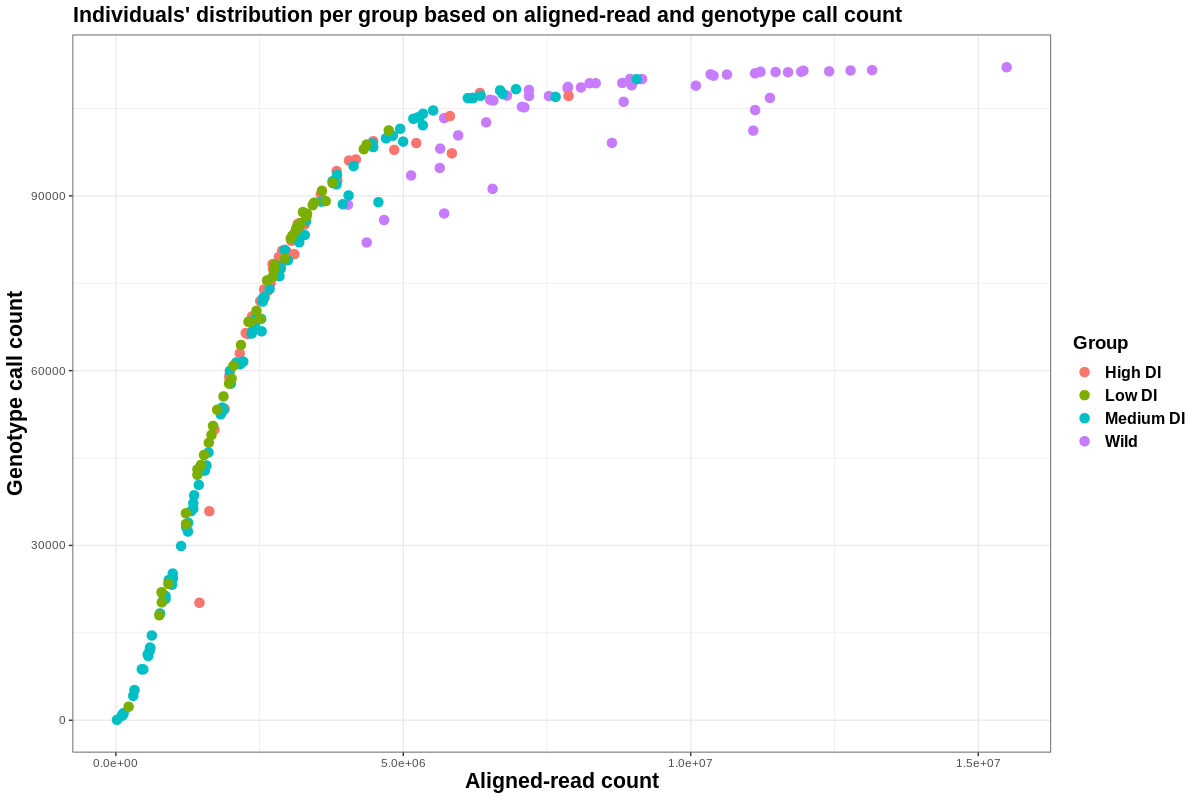
**

Figure 1. Distribution of individuals in the four groups based on their aligned-read and genotype call count (individuals with lower number of aligned-read count have lower number of genotype count). Each dot represents an individual and color represents a group. X-axis is number of aligned-read and y-axis is the number of genotype calls for each individuals. Based on the distribution, an aligned-read per individual ranging between 3,000,000bp to 6,000,000bp were choose.

Table 1. Number of individuals before and after filtration based on their aligned-read quality. The low and medium groups had many individuals removed after filtration, requiring us to standardize the number of aligned reads across groups.

|  |  |  |
| --- | --- | --- |
| Group | Number of samples before filtering | Number of samples after filtering |
| Wild | 48 | 48 |
| Low DI | 48 | 19 |
| Medium DI | 96 | 34 |
| High DI | 48 | 22 |

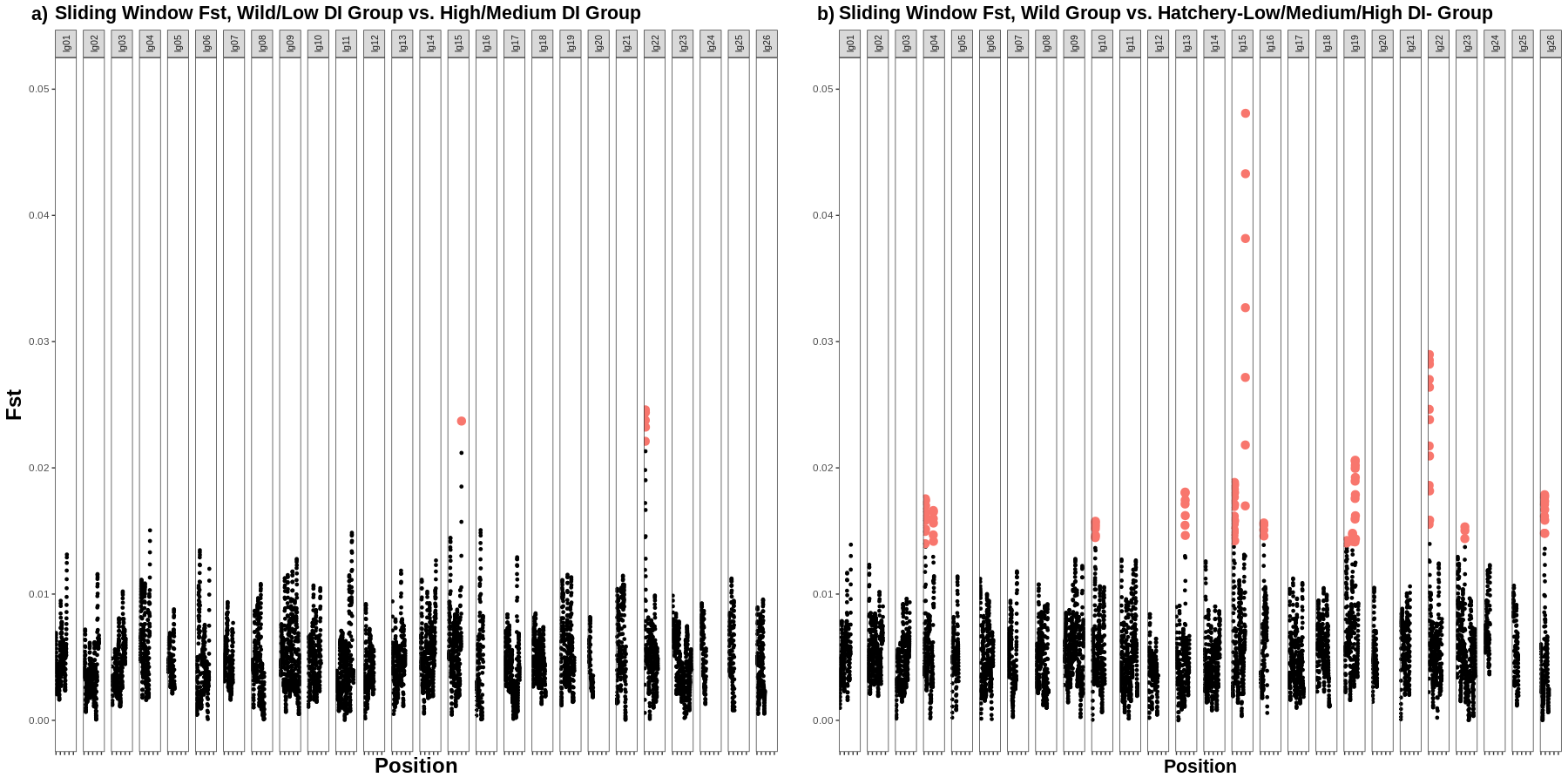


Figure 2. Pairwise FST sliding window analysis. X-axis: position on a chromosome, Y-axis: FST value of each window. Each dot represents a window and red dot represents significant windows (p≤0.001). a) comparison 1 between wild/low DI group vs. medium/high DI group. Elevated regions on chromosomes 15 and 22 are significant (lg 15 and lg22). b) comparison 2 between wild groups and hatchery (low/medium/high DI) group. Elevated regions on chromosomes 4, 10, 13, 15, 16, 19, 22, 23, and 26 are significant. Elevated regions on chromosomes 15 and 22 presenting in the both comparison are more likely to be under domestication selection.

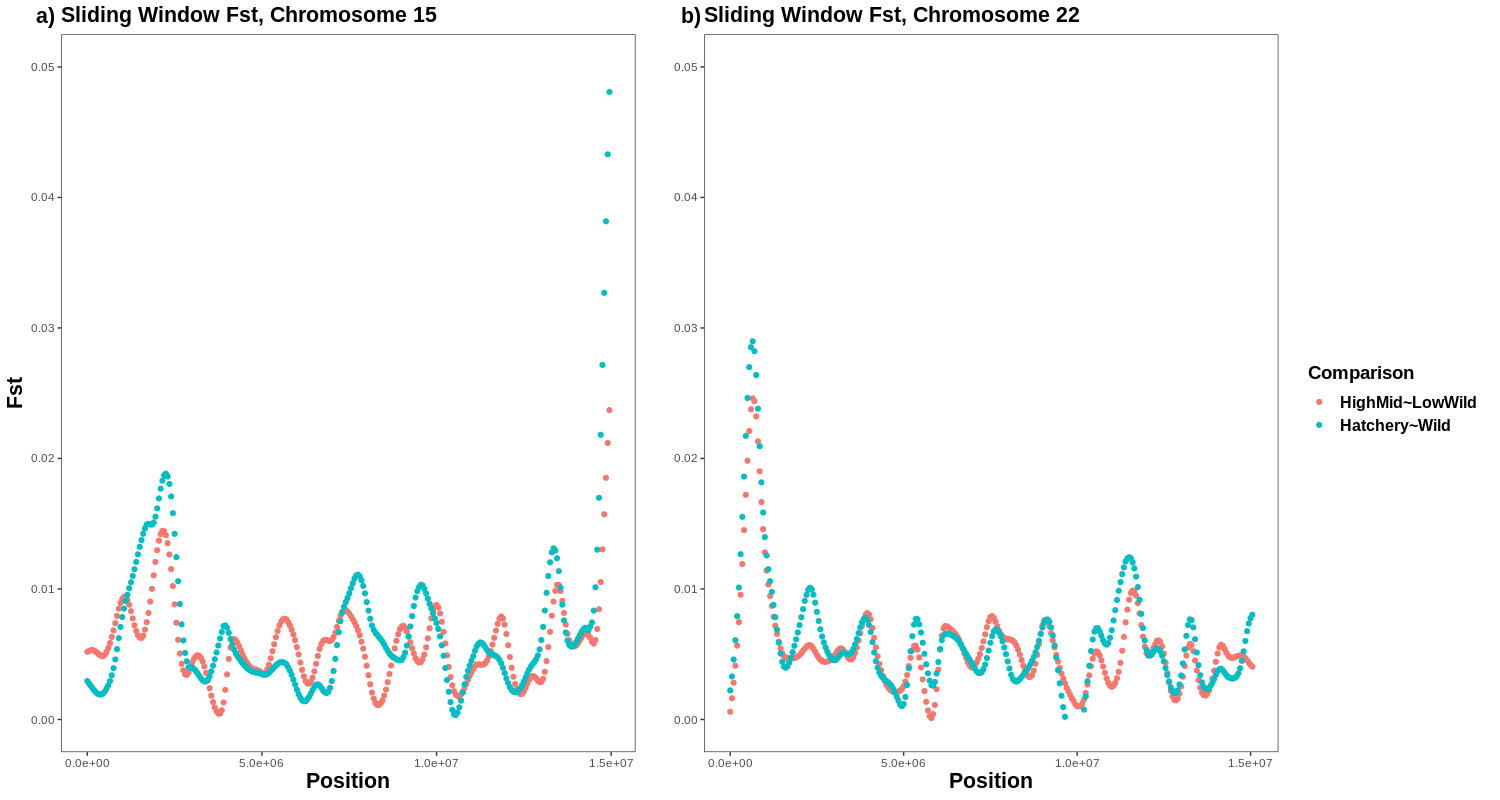


Figure 3. Higher resolution of pairwise FST sliding window analysis of the candidate regions in the two comparisons. X-axis: position on a chromosome, Y-axis: FST value of each window. Red represent comparison 1 and blue represents comparison 2. a) distribution of windows on chromosome 15 (lg15). b) distribution of windows on chromosome 22 (lg22). The peaks are located near to the edge of the chromosomes and window distribution follows the same pattern in the both comparisons. This strengthens the probability of the region being under selection.

*References*

1. *Ali OA, O’Rourke SM, Amish SJ, et al. Rad capture (Rapture): Flexible and efficient sequence-based genotyping. Genetics. 2016;202(2):389-400. doi:10.1534/genetics.115.183665.*
2. *Finger AJ, Mahardja B, Fisch KM, et al. A conservation hatchery population of delta smelt shows evidence of genetic adaptation to captivity after 9 generations. J Hered. 2018;109(6):689-699. doi:10.1093/jhered/esy035.*
3. *Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15(1):1-13. doi:10.1186/s12859-014-0356-4.*
4. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.
5. *Zhu, Z., Wang, Y., Zhou, X. et al. SWAV: a web-based visualization browser for sliding window analysis. 2020, Sci Rep 10, 149.*
6. *William Hemstrom, Melissa Jones. snpR: user friendly population genomics for SNP datasets with categorical metadata. 2021, Authorea. February 06.*