Alaska Hatchery Research Group **Technical Document:**[1] #13

Title: Experimental design of pink salmon SNP development **Version:** 1.0
Authors: T. H. Dann, C. Habicht, W. D. Templin, L. W. Seeb and J. E. Seeb
Date:

Abstract

The fitness component of the Alaska Hatchery Research Program requires single nucleotide polymorphisms (SNPs) that are currently unavailable. The Gene Conservation Laboratory and the Seeb Laboratory of the University of Washington are using restriction site associated DNA sequencing to develop SNPs for pink salmon. Here we describe the experimental design we are using to develop SNPs, provide preliminary sequencing results, and propose SNP selection criteria to the Science Panel. DNA from 665 individual pink salmon sampled from 17 populations has been sequenced. Average retained reads was 1.6M for the 190 individuals that have had both rounds of sequencing, suggesting that we will achieve adequate sequence coverage to accurately estimate allele frequencies and identify variable SNPs useful for the fitness study. We propose a series of gating criteria and ranking measures to select SNPs for the fitness study and seek feedback from the Science Panel.

Background of AHRP

Extensive ocean-ranching salmon aquaculture is practiced in Alaska by private non-profit corporations (PNP) to enhance common property fisheries. Most of the approximately 1.7B juvenile salmon that PNP hatcheries release annually are pink salmon in Prince William Sound (PWS) and chum salmon in Southeast Alaska (SEAK; Vercessi 2014). The large scale of these hatchery programs has raised concerns among some that hatchery fish may have a detrimental impact on the productivity and sustainability of natural stocks. Others maintain that the potential for positive effects exists. To address these concerns ADF&G convened a Science Panel for the Alaska Hatchery Research Program (AHRP) whose members have broad experience in salmon enhancement, management, and natural and hatchery fish interactions. The AHRP was tasked with answering three priority questions:

I. *What is the genetic stock structure of pink and chum salmon in each region (PWS and SEAK)*?
II. *What is the extent and annual variability in straying of hatchery pink salmon in PWS and chum salmon in PWS and SEAK*?

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and other members of the Science Panel of the Alaska Hatchery Research Program. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division

III. *What is the impact on fitness (productivity) of natural pink and chum salmon stocks due to straying of hatchery pink and chum salmon*?

Introduction

Measuring the Impact on Fitness

To answer the third question, we need to know the origin and pedigree of each fish captured in select streams across multiple generations. **Origin** refers to the type of early life-history habitat (hatchery or natural) that a fish experienced. **Pedigree** refers to the family relationship among parents and offspring. **'Ancestral origin'** refers to the origin of an individual's ancestors (e.g., two parents of a single origin [hatchery/hatchery or natural/natural] or two parents of mixed origin [hatchery/natural]). These ancestral origins can be determined by combining information from three sources: identification of hatchery origin from otolith marks, pedigree from genetic data, and age from scales (for chum salmon from SEAK). By pairing these data within fish and across generations, we can estimate **reproductive success** (**RS**) among cross types (i.e. hatchery-hatchery, hatchery-natural, and natural-natural origin crosses). The AHRP is using the **relative reproductive success** (**RRS**) *of hatchery-origin fish to natural-origin fish as the measure of fitness in this study* (Tech Doc 1 – Shedd et al. 2014). The current design is for RRS to be estimated in six populations: Erb Creek, Hogan Bay, Paddy Creek, Spring Creek, Gilmour Creek and Stockdale Creek.

Identification of genetic markers

Estimating relative reproductive success via parentage analyses on the scale of this research program necessitates single nucleotide polymorphisms (SNPs). While microsatellites have historically been the marker-type of choice for parentage analysis due to their high variability and general availability, SNPs have recently received increased attention due to the potential for high-throughput screening, low genotyping error rates, and transferability among laboratories. With current technology at the ADF&G Gene Conservation Laboratory (GCL), genotyping cost per locus for microsatellites is an order of magnitude higher than for SNPs. With all parents sampled, theoretical work has shown that a set of 60-100 SNPs with minor allele frequency (MAF) > 0.3 allows for accurate pedigree reconstruction of large populations that contain thousands of potential mothers, fathers, and offspring (Anderson and Garza 2006). This theoretical work has been confirmed by empirical studies that have compared parentage analysis with both microsatellites and SNPs (Hauser et al. 2011, Tokarska et al. 2009, Anderson 2012). Hauser et al. (2011) compared 11 highly variable microsatellites specifically chosen for parentage analysis to 80 SNPs originally designed for genetic stock identification (GSI; high among-population variation). Over half of the SNPs had a MAF < 0.2, a level below which SNPs rapidly lose power in parentage analysis (Anderson and Garza 2006). Despite the limitations of the SNP marker set used by Hauser et al. (2011) with respect to parentage analysis, the authors found that assignment success was always higher for SNPs than for microsatellites across different parentage analysis software programs.

Accurately and confidently assigning offspring to parents requires many independently assorting alleles at genetic markers that are variable within the population being studied. Recent simulation work indicates that ~192 independent alleles with MAF greater than 0.3 can resolve parent-offspring relationships in study conditions expected for the AHRP (large populations, not all parents sampled; Shedd et al. 2015). The fitness study will also need to analyze many thousands of individuals in the laboratory, so genetic markers with high throughput capabilities are required. These factors make it clear that SNPs are the genetic marker to use for the fitness study. Other benefits of SNPs include the ability to select from 10,000s of potential markers, the reliability of genotype calls as a direct measure of sequence, reduced expense to genotype, and transferability among laboratories. Taqman assays (ADF&G's current SNP genotyping methodology) have been developed for 51 pink salmon SNPs that exhibit signatures of selection (University of Washington, unpublished), but these likely are not sufficient to resolve parentage in AHRP fitness streams. Thus, developing a panel of 192 SNPs that resolve parentage in both odd- and even-year lineages is a major objective of the fitness study.

Marker development

Restriction site associated DNA sequencing (RAD sequencing) has emerged as the best approach for SNP discovery and is the method used to develop SNPs for AHRP (Andrews et al. 2016). The approach consists of: 1) digesting DNA with restriction enzymes to isolate RAD tag sites among all individuals, 2) ligating a short sequence of individual-specific bases to the cut site to barcode each fragment to individual, 3) shearing DNA to reduce the length of sequence fragments for sequencing, 4) generating many copies of RAD tag sites via PCR amplification, 5) pooling individuals into libraries and 6) sequencing libraries to generate sufficient copies of RAD tag sites to confidently genotype individuals and discover variant bases (SNPs; Baird et al. 2008). Benefits of this approach include the standardized method that has been well vetted and documented in primary literature, the potential to select from among 10,000s of SNPs, and the transferability of information from SNPs developed via RAD sequencing across studies. The method has been successfully used in many taxa to address a variety of questions including identification of SNPs to distinguish closely related populations of Chinook salmon and genomic regions of divergence between ecotypes of threespine stickleback (Hohenlohe et al. 2010, Larson et al. 2014). The method has also proven successful in identifying SNPs exhibiting parallel selection between the two lineages of pink salmon (Seeb et al. 2014).

Selecting markers in context of genome organization

Constructing marker panels in the context of a linkage map is important to selecting independent markers tailored to their application. Linkage maps describe the order and relative spacing of markers along chromosomes (or linkage groups). The order and relative spacing are estimated from the frequencies of recombinations between loci and are often generated by examining segregation patterns of haploid or diploid families (i.e., Limborg et al. 2015). Selecting markers located on different chromosomes or distant enough from one another within chromosomes ensures statistical independence among markers, an assumption made by many common

analyses. For the purposes of parentage analyses, statistical independence of genotypes is important for statistical power as tightly linked markers provide redundant information for resolving parent-offspring relationships.

Goals of Technical Document

The goals of this technical document are to:

1) Describe the experimental design used to develop SNPs for pink salmon in Prince William Sound;
2) Provide results to date and an expected timeline for the remainder of SNP development process; and
3) Propose and, ask for input from the Science Panel on, the selection criteria for SNPs to be used for parentage analyses.

Methods

Identification of samples for RAD sequencing

A primary objective of our study is to identify a panel of genetic markers that maximizes parentage power in the six streams under analysis by the AHRP. Marker development for AHRP would ideally use individuals from fitness streams to identify highly variable markers. However, samples from individuals from nearby streams are a better choice for this particular study. The samples from fitness streams are of poor quality. Tissues were sampled from only post-spawning mortalities; such tissues often have degraded DNA that fail to provide quality RAD data (Graham et al. 2015). As a result we identified three criteria of representative populations to help select from available samples: 1) availability of quality tissue samples with paired data to identify natural-origin individuals and provide the potential to identify sex-linked SNPs discovered via RAD sequencing; 2) expectation of presence in fitness streams (i.e. one of the three hatchery populations potentially present in fitness streams; Habicht et al. 2000); and 3) expectation that population allele frequencies will be similar to natural fitness-stream populations. We considered tissue samples to be of good quality if they originated from spawning individuals of natural-origin pink salmon collected within the recent five years. When available, we also attempted to include early and late components of pink salmon runs to incorporate any temporal variability that may exist within the six fitness populations.

Results

Experimental design of RAD sequencing

The experimental design of our RAD sequencing effort was determined by the objectives and criteria above. The GCL has 250 collections of pink salmon populations (including hatcheries) from Prince William Sound; 134 of these originate from allozyme collections from the 1990s that likely include hatchery fish and/or degraded DNA. Seventy-six collections were sampled in 2013 and 2014, 20 of which are either hatchery or fitness stream populations. Of the remaining

56 collections, we selected 17 for RAD sequencing based upon the objectives and criteria above. We included the three hatchery populations expected to be observed in fitness streams: Armin F. Koernig Hatchery (AFK), Cannery Creek Hatchery, and Solomon Gulch Hatchery (aka VFDA). We attempted to select collections from West and East Prince William Sound to parallel the distribution of pink salmon populations throughout the Sound and represent allele frequencies expected in fitness streams (Figure 1; Table 1).

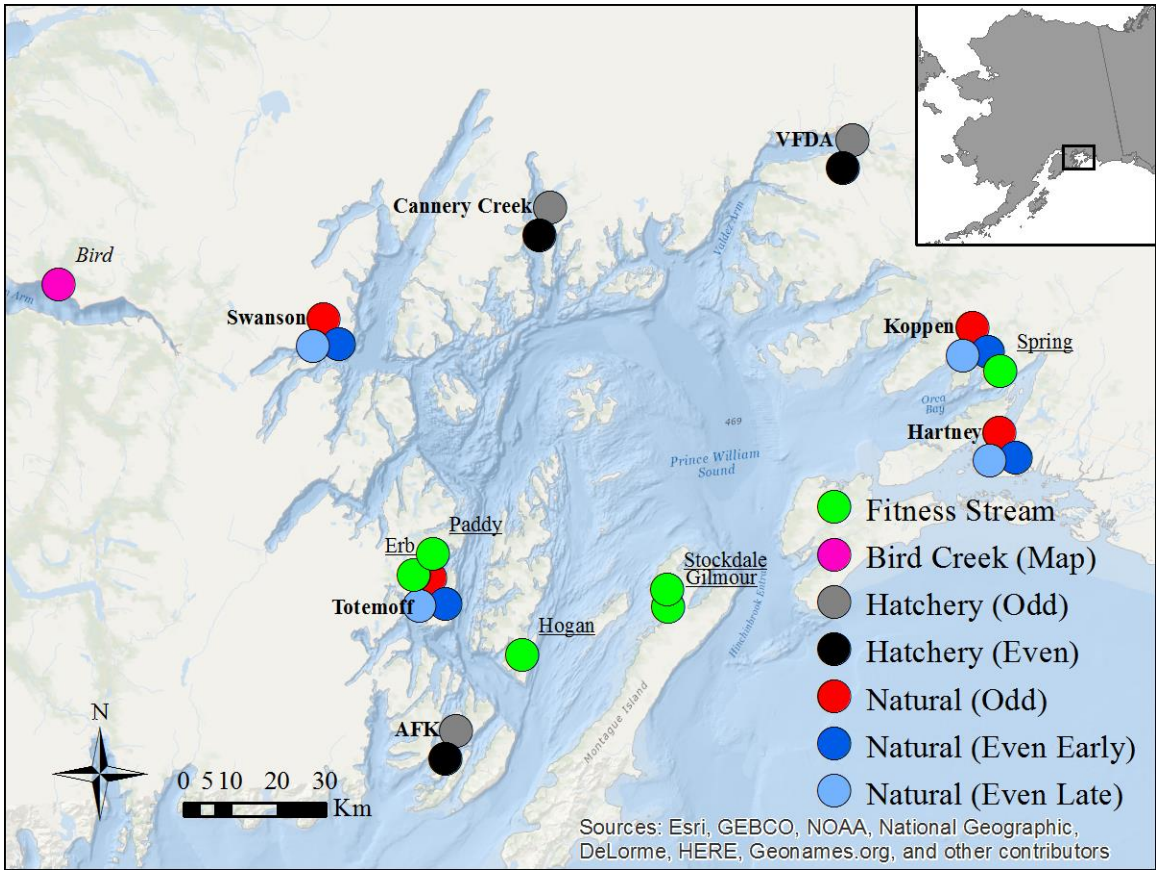


Figure 1. Locations of Alaska Hatchery Research Program fitness study streams (green), the population used for linkage map construction (Bird Creek from Cook Inlet), and hatchery (gray/black) and natural (red/blue) populations selected for RAD sequencing. Fitness study stream names are underlined, the name of the linkage map stream is italicized and RAD sequence stream names are in bold.

Table 1. Name, Gene Conservation Laboratory (GCL) code, collection date, latitude, longitude for collections used to develop SNPs. Also included are the number of samples available (Total) and included in RAD sequencing.

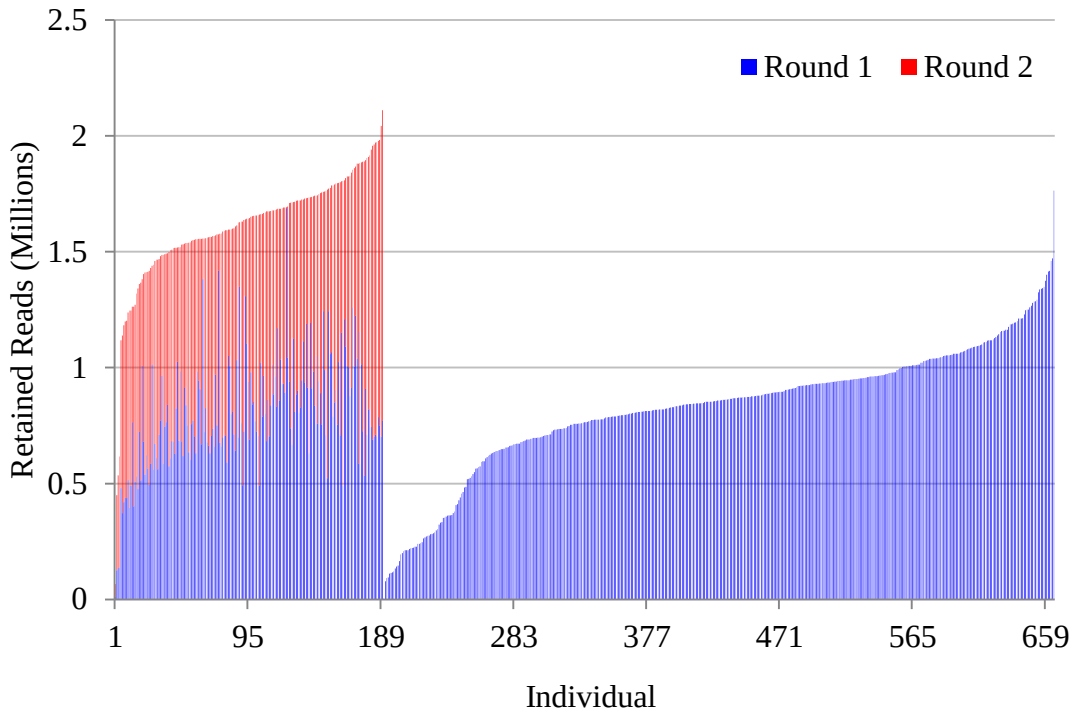
					Sample size	
Collection	GCL Code	Collection Date	Latitude	Longitude	Total	Sequenced
AFK Hatchery	PAFK13	8/22/2013	60.051	-148.065	200	37
Cannery Creek Hatchery	PCANN13	8/25/2013	61.019	-147.514	197	37
VFDA Hatchery	PVFDA13	8/9/2013	61.084	-146.304	200	37

Totemoff Creek	PTOTM13	8/27/2013	60.343	-148.088	96	37
Swanson Creek	PSWAN13	8/12/2013	60.849	-148.412	121	37
Koppen Creek	PKOPP13	7/30/2013	60.706	-145.898	216	37
Hartney Creek	PHART13	8/3/2013	60.502	-145.842	271	37
AFK Hatchery	PAFK14	8/25/2014	60.065	-148.065	200	37
Cannery Creek Hatchery	PCANN14	8/26/2014	61.019	-147.514	200	37
VFDA Hatchery	PVFDA14	7/31/2014	61.131	-146.348	200	37
Totemoff Creek Early	PTOTM14E	7/29/2014	60.343	-148.088	96	37
Totemoff Creek Late	PTOTM14L	8/26/2014	60.343	-148.088	102	37
Swanson Creek Early	PSWAN14E	7/28/2014	60.849	-148.412	120	37
Swanson Creek Late	PSWAN14L	8/25/2014	60.849	-148.412	125	37
Koppen Creek Early	PKOPP14E	8/2/2014	60.706	-145.898	120	35
Koppen Creek Late	PKOPP14L	8/31/2014	60.706	-145.898	124	38
Hartney Creek Early	PHART14E	8/4/2014	60.502	-145.842	46	37
Hartney Creek Late	PHART14L	8/21/2014	60.502	-145.842	109	37

Sequencing results

We have completed 64% of the sequencing and expect to complete the remainder in January, 2016. Our approach to collecting RAD data involves sequencing each library of 95 individuals twice at University of Oregon's Genomics and Cell Characterization Core Facility (https://gc3f.uoregon.edu/). The first round of sequencing used constant volumes of extracted DNA solution from each fish. The second round of sequencing is conducted after DNA concentrations are titrated to equalize the sequence coverage (number of retained reads) among individuals based on results from the first round. Equal coverage among individuals results in better coverage among markers and subsequently better overall estimates of allele frequencies for each population. Two libraries have been sequenced twice while the other five are in the process of being re-sequenced.

Initial results indicate that the goal of 1.5M retained reads will be reached for each individual. Individuals from the two libraries that have had both rounds of sequencing averaged 1.6M retained reads, while individuals from the remaining 5 libraries averaged 825K retained reads after one round (Figure 2). As expected, average retained reads/collection were unequal after the first round of sequencing (e.g., PVFDA13=232K, PAFK14=1,063K) but became much more equal after the second round of sequencing for the 6 completed collections (range 1,534-1,755K; Table 2).



174

175 Figure 2. The number of retained reads for 665 pink salmon sequenced to identify SNPs for the
176 Alaska Hatchery Research Program. Of the 665 fish, 190 have had both round 1 (blue) and 2
177 (red) of sequencing while the remaining 475 are expected to be completed in January.

178 Table 2. Number of individuals sequenced and average retained reads after the first, second, and
179 both rounds of sequencing.

Collection	n	Round 1	Round 2	Both
PAFK13	37	911,239	622,276	1,533,515
PCANN13	37	881,855	750,057	1,631,912
PHART13	37	848,992	793,622	1,642,613
PSWAN13	37	761,967	849,031	1,610,998
PKOPP13	37	600,574	1,008,059	1,608,633
PTOTM13	37	603,598	1,084,180	1,754,826
PVFDA13	37	232,305	NA	NA
PAFK14	37	1,063,014	NA	NA
PCANN14	37	930,839	NA	NA
PVFDA14	37	917,992	NA	NA
PHART14E	37	854,268	NA	NA
PHART14L	37	795,811	NA	NA
PKOPP14E	35	872,938	NA	NA
PKOPP14L	38	929,163	NA	NA
PSWAN14E	37	895,058	NA	NA

PSWAN14L	37	875,550	NA	NA
PTOTM14E	37	786,726	NA	NA
PTOTM14L	37	941,559	NA	NA

Proposed criteria to select SNPs for parentage purposes

We propose a series of gating criteria and ranking measures to select the set of 192 SNPs from the 10,000s available for the parentage panel. The gating criteria will reduce the pool of SNPs through exclusion based on adverse characteristics. As the first gating criterion, we will exclude loci that are duplicated as a result of the whole genome duplication common to the ancestor of all salmonids (Ohno 1970) to ensure that only unambiguous genotypes are used to determine parent-offspring relationships. Similarly, we will exclude loci that exhibit a minor allele frequency < 0.05 in any collection and those that fail to conform to Hardy-Weinberg expectations. Finally, to ensure independence of SNPs included in our panel, we will attempt to include only those SNPs with a minimum separation of 15 centiMorgans (cM). The cM is a measure of genetic distance between loci on a chromosome that is defined as the number of chromosomal crossovers (recombinations) expected among 100 individuals in a single generation and thus is a good measure to identify independence among loci. For example, 15 cM equates to 15 crossover events occurring for every hundred offspring in a generation between two loci. The current linkage map available for pink salmon, based upon odd-year populations from Washington State and Bird Creek, Alaska, characterizes 26 chromosomes with an average length of 129 cM (Limborg et al. 2014; University of Washington, unpublished). Given our goal of developing a panel of ~200 SNPs, we expect to select ~8 SNPs from each chromosome which translates into 16 cM spacing between SNPs if we assume similar chromosome lengths.

To choose from the remaining SNPs, we propose to use the same ranking measure proposed for the selection of chum salmon SNPs in Technical Document 2 (Shedd et al. 2014). Each SNP will be assigned a score based upon the mean and standard deviation (SD) of minor allele frequency (MAF) across the RAD collections following:

$$\text{score} = {2 \times (\text{mean MAF})} \Big/ {(1 + \text{SD of MAF})}$$

This measure standardizes scores between 0 and 1 and is an intuitive measure to base parentage power as MAF is the most important factor in a marker's power in parentage analyses (Anderson and Garza 2006).

We expect to identify hundreds of SNPs that can resolve parent-offspring relationships for AHRP and will select the final set based upon the gating and ranking criteria above as well as each marker's ability to be accurately genotyped using amplicon technology. The GCL is transitioning to the GT-seq methods (Campbell et al. 2014); ensuring that the final panel of

markers produces consistent sequence coverage among markers will result in more accurate genotype calls.

Questions for the AHRP

1. Are the proposed methods for selecting SNPs for parentage analyses adequate? Do you suggest other approaches to selecting SNPs?

AHRP Review and Comments

This technical document has been reviewed.

This document is acceptable to the AHRG.

There was once comment by Alex Wertheimer who stated:

The approach the authors have outlined for selecting SNPs for use in the parentage analysis is very well-justified and scientifically sound. The authors are taking advantage of advance techniques in identifying SNPs in the genome of pink salmon, and have devised criteria for selecting the "best" 192 SNP markers for use in the parentage analyses for the many thousands that are available. The authors' comprehensive description of the rationale and methodology continues the excellent job of documenting the scientific rigor brought to bear to address the priority questions and objectives of the Alaska Hatchery Research Program. I certainly have no recommendations on alternate approaches.

References

Anderson, E.C. 2012. Large-scale Parentage Inference with SNPs: an Efficient Algorithm for Statistical Confidence of Parent Pair Allocations. Statistical applications in genetics and molecular biology **11**(5).

Anderson, E.C., and Garza, J.C. 2006. The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics **172**(4): 2567-2582.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet **advance online publication**.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE **3**(10).

Campbell, N.R., Harmon, S.A., and Narum, S.R. 2014. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources: n/a-n/a.

Graham, C.F., Glenn, T.C., McArthur, A.G., Boreham, D.R., Kieran, T., Lance, S., Manzon, R.G., Martino, J.A., Pierson, T., Rogers, S.M., Wilson, J.Y., and Somers, C.M. 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Molecular Ecology Resources **15**(6): 1304-1315.

Habicht, C., Simpson, E.M., and Seeb, J.E. 2000. Broodstock acquisition and release sites for hatcheries producing pink salmon in Prince William Sound. Alaska Department of Fish and Game, Regional Information Report No. 5J00-07.

Hauser, L., Baird, M., Hilborn, R., Seeb, L.W., and Seeb, J.E. 2011. An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Resources **11**: 150-161.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., and Cresko, W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. PLOS Genetics **6**(2): e1000862.

Larson, W.A., Seeb, J.E., Pascal, C.E., Templin, W.D., and Seeb, L.W. 2014. Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. Canadian Journal of Fisheries and Aquatic Sciences **71**(5): 698-708.

Limborg, M.T., Waples, R.K., Allendorf, F.W., and Seeb, J.E. 2015. Linkage Mapping Reveals Strong Chiasma Interference in Sockeye Salmon: Implications for Interpreting Genomic Data. G3: Genes|Genomes|Genetics **5**(11): 2463-2473.

Limborg, M.T., Waples, R.K., Seeb, J.E., and Seeb, L.W. 2014. Temporally Isolated Lineages of Pink Salmon Reveal Unique Signatures of Selection on Distinct Pools of Standing Genetic Variation. Journal of Heredity **105**(6): 741-751.

Ohno, S. 1970. Enormous diversity in genome sizes of fish as a reflection of natures extensive experiments with gene duplicaiton. Transactions of the American Fisheries Society **99**(1): 120-&.

Seeb, L.W., Waples, R.K., Limborg, M.T., Warheit, K.I., Pascal, C.E., and Seeb, J.E. 2014. Parallel signatures of selection in temporally isolated lineages of pink salmon. Mol Ecol **23**(10): 2473-2485.

Shedd, K.R., Dann, T.H., Habicht, C., and Templin, W.D. 2014. Alaska Hatchery Reserach Program Technical Document 1: Defining relative reproductive success: which fish count? ADF & G Technical Document.

Shedd, K.R., Dann, T.H., Habicht, C., and Templin, W.D. 2014. Alaska Hatchery Reserach Program Technical Document 2: Parentage SNP selection - SEAK chum. ADF & G Technical Document.

Tokarska, M., Marshall, T., Kowalczyk, R., Wojcik, J.M., Pertoldi, C., Kristensen, T.N., Loeschcke, V., Gregersen, V.R., and Bendixen, C. 2009. Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. Heredity **103**(4): 326-332.

Vercessi, L. 2014. Alaska salmon fisheries enhancement program 2013 annual report. Fishery Mangement Report. Alaska Department of Fish and Game, Anchorage.