Alaska Hatchery Research Program **Technical Document:**[1] 2

Title: Parentage SNP selection – SEAK chum **Version:** 1.0
Authors: K. Shedd, T. H. Dann, C. Habicht, and W. D. Templin
Date: May 27, 2014

Abstract

Uncertainty about the impact of hatchery salmon on the productivity and sustainability of natural stocks in Prince William Sound (PWS) and Southeast Alaska (SEAK) was the impetus for the Alaska Hatchery Research Program (AHRP). One major portion of this project is designed to use genetic data to perform parentage analysis in order to create pedigrees and assess the impact on fitness (productivity) of natural pink *Oncorhynchus gorbuscha* and chum *O. keta* salmon stocks due to straying of hatchery pink and chum salmon. **Single nucleotide polymorphisms (SNPs)** have been identified as the marker type for the parentage analysis. Markers are being developed for pink salmon and markers are available for chum salmon. However, the marker suite for chum salmon has yet to be determined. Here we describe our intended process to select the set of SNPs for chum salmon that provides the maximum resolution possible for parentage analysis to meet the objectives of the AHRP.

Background of AHRG

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 PNP hatcheries release annually are pink salmon in Prince William Sound (PWS) and chum salmon in Southeast Alaska (SEAK; Vercessi 2013). 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. ADF&G convened a Science Panel (Alaska Hatchery Research Group; AHRG) whose members have broad experience in salmon enhancement, management, and natural and hatchery fish interactions. The AHRG 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)*?

[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 Alaska Hatchery Research Group. 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.

II. *What is the extent and annual variability in straying of hatchery pink salmon in PWS and chum salmon in PWS and SEAK*?

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 AHRG is using the **relative reproductive success** (**RRS**) *of hatchery-origin fish to natural-origin fish as the measure of fitness in this study* (Shedd et al. 2014).

Problem: Which Markers to Use for Parentage Analysis

The reconstruction of pedigrees via parentage analysis using genetic markers is based on simple **Mendelian inheritance**, where an offspring inherits one of two alleles from each parent. While the concept is simple, the implementation of exclusion-based parentage analysis can be challenging in open systems, where not all parents are sampled, and genetic information is limited and/or subject to genotyping error. For this purpose, there are a wide variety of statistical likelihood methods that utilize either a frequentist-likelihood or Bayesian approach to assess the probability of parent-offspring relationships. Nevertheless, any parentage analysis is subject to the limitations of the genetic marker set employed. Thus, it is important to 1) select informative markers; 2) select robust markers that produce highly accurate and consistent genotypes under a wide range of tissue qualities, and 3) determine the requisite number of markers for successful parentage analysis.

While **microsatellites** have historically been the maker-type of choice for parentage analysis due to their high variability and general availability, SNPs have recently received increased attention due to their 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. Theoretical work has shown that a set of 60-100 SNPs with minor allele frequency (MAF) $\geq$ 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). 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.

The GCL has 188 SNPs available for chum salmon (Table 1). These 188 SNPs have been previously narrowed down to 96 SNPs (the maximum number of SNPs that can be run on a single high-throughput SNP chip), however this set of 96 SNPs was optimized for GSI in western Alaska using high-quality samples as part of The Western Alaska Salmon Stock Identification Program (WASSIP) (DeCovich et al. 2012), not parentage analysis in Southeast Alaska using carcass tissues. In order to make the final selection of the best 96 SNPs for parentage analysis for the AHRP, the GCL proposes to determine the performance for all 188 SNPs on a sample from all 4 pedigree streams and then empirically determine the set of SNPs required for optimal success in parental assignments of 2014 alevin to 2013 adults in Fish Creek.

Goals of Technical Document

Two goals of this technical document are to:

1) Propose and document the method for selecting markers to be used in parentage analysis.
    i. Determine population genetic summary statistics for all 188 SNPs for the 4 chum salmon pedigree streams sampled in SEAK.
    ii. Determine laboratory performance for all 188 SNPs for chum salmon carcasses sampled in SEAK.
    iii. Determine the required number of SNPs necessary for robust, accurate parentage analysis of alevin and adult chum salmon in SEAK using Fish Creek as the test population.
2) Request a decision by the AHRG on these methods prior to August 2014.

Methods

Phase 1: Ranking SNPs

Suitability of SNPs for parentage analysis for AHRP: All 188 SNP markers will be assayed in 95 randomly selected adult individuals sampled in 2013 from each of the 4 pedigree streams (Figures

I. Figure 1).

1. Unranked measures: Measures in this section will be given veto power and markers will be discarded if they do not pass the following tests.
    a. Hardy-Weinberg Equilibrium (HWE): Conformance to HWE will be measured with Genepop version 4.0.11 (Rousset 2008). Markers out of HWE at α =0.05 in any of the 4 populations or exhibiting overall significance, measured across all 4 populations, at α =0.01 will be dropped. An overall p-value will be calculated according to Fisher's method for combined probability test.
    b. Linkage Disequilibrium: Linkage disequilibrium will be measured with Genepop version 4.0.11 (Rousset 2008). Marker pairs that exhibit linkage disequilibrium at α =0.05 in 3 or more of populations examined will be considered "associated" and the SNP with the lesser average MAF among populations will be dropped.
    c. Laboratory Performance: Only markers that have an overall relative scoring success rate of >80% (relative to the best-performing marker, to account for poor tissue quality) and a discrepancy rate of <2% across all 4 populations will be retained.
2. Ranked measures: The measures in this section of the selection process will be scored between 0 and 1 (worst to best) using the equation:

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

    a. "Mean MAF" is the mean MAF calculated across the 4 pedigree streams and is our primary metric of interest,
    b. "SD of MAF" is the standard deviation of MAF among the 4 pedigree streams. This attributes a "cost" to including markers that are highly variable among populations (i.e., useful in some but not others).
    c. Each SNP that passed the unranked measures above will be assigned a rank based upon its score for this measure. This order of ranks will be used in subsequent measures below.

Phase 2: Evaluating SNP Sets

II. Empirical test of SNPs for parentage analysis: The SNP markers that pass the unranked, veto portion of Phase 1 will then be assayed for all remaining adults collected from Fish Creek in 2013 and all alevin collected from Fish Creek in 2014 in order to perform parentage analysis, as these are currently the only chum samples available for parentage analysis.
    1. Parentage analysis: Parentage analysis will be performed by assigning alevin to adults using the highest ranked markers (according to Equation 1) in sets of SNPs that are efficiently screened using GCL laboratory methods: 24, 48, 96, 120, 144, and all SNPs marker sets.

2. Cost/benefit analysis: We will examine the relationship between increasing number of markers used in a set (cost) and success in parentage analysis. Dependent variables will include:
    a. Number of offspring assigned to two parents
    b. Number of offspring assigned to one parent
    c. Number of parents with one or more successfully assigned offspring
    d. Mean log-odds (LOD) score ratios between most likely parent and the next most likely parent from Cervus3 (Hauser et al. 2011) as a measure of confidence in parentage assignment.
        i. Cervus3, a likelihood parentage assignment program, provides LOD score for each parent-offspring assignment.
        ii. A LOD score is the natural log of the likelihood ratio (i.e. probability of a putative-parent-offspring pair being related divided by the probability that they are unrelated).
        iii. For a given offspring, the LOD score is computed for multiple putative parents.
        iv. The ratio of LOD scores between the most likely parent and the second most likely parent gives a proxy for the level of confidence in assigning the most likely parent.
        v. The distribution of LOD ratios can be compared between marker sets to assess the level of confidence in correct parent assignments.
    e. Parentage error rates (putative)
        i. Error rates defined by comparing assignments of a set to assignments made with all available markers (Gold Standard).
        ii. Type I error – assigning an untrue parent
        iii. Type II error – failing to assign a true parent when the true parent is present in the sample

Final Considerations

III. Final considerations: The candidate SNPs will be ordered from best to worst with respect to the measures in the ranked portion of Phase 1 (Equation 1). Given that the GCL is optimized to use 96 SNP markers in a set, the top 96 candidates from Phase 1 will be selected, unless Phase 2 suggests that equal assignment power can be obtained with 48 SNPs or that more than 96 SNPs are necessary to acquire adequate power.

Questions for the AHRG

1. Are the proposed methods for ranking markers appropriate and sufficient? Are there other considerations that should be assessed as well?
2. Are the proposed methods for determining marker set appropriate?

AHRG Review and Comments

This technical document was discussed at the December 12, 2014 meeting of the AHRG. In addition it was reviewed by email exchange prior to the meeting.

The proposed methods are acceptable.

This document is acceptable to the AHRG.

References

Anderson, E. C., and J. C. Garza. 2006. The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. Genetics 172:2567-2582. http://www.genetics.org/cgi/content/abstract/172/4/2567

DeCovich, N. A., J. R. Jasper, S. M. Turner, C. Habicht, and W. D. Templin. 2012. Western Alaska Salmon Stock Identification Program Technical Document 23: Chum salmon SNP selection results. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-25, Anchorage. http://www.adfg.alaska.gov/FedAidPDFs/RIR.5J.2012.25.pdf

Elfstrom, C. M., C. T. Smith, and L. W. Seeb. 2007. Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. Molecular Ecology Notes 7(6):1211-1215 (5). http://www3.interscience.wiley.com/journal/120808786/abstract?CRETRY=1&SRETRY=0

Hauser, L., M. C. Baird, R. Hilborn, L. S. Seeb, and J. E. Seeb. 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(Supplement 1):13. http://www.ncbi.nlm.nih.gov/pubmed/21429171

Petrou, E. L., L. Hauser, R. S. Waples, W. D. Templin, D. Gomez-Uchida, and L. W. Seeb. 2013. Secondary contact and changes in coastal habitat availability influence the nonequilibrium population structure of a salmonid (Oncorhynchus keta). Molecular Ecology 22(23):5848-5860 (13). http://onlinelibrary.wiley.com/doi/10.1111/mec.12543/pdf

Rousset, F. 2008. GENEPOP 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8(1):103-106 (4). http://www.arlis.org:2074/ehost/pdfviewer/pdfviewer?vid=4&hid=10&sid=c9d59b7a-94df-4e14-925f-df0bb3d382e3%40sessionmgr4

http://genepop.curtin.edu.au/

Seeb, J. E., C. E. Pascal, E. D. Grau, L. W. Seeb, W. D. Templin, T. Harkins, and S. B. Roberts. 2011. Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. Molecular Ecology Resources 11(2):335-348 (14). http://onlinelibrary.wiley.com/doi/10.1111/j.1755-0998.2010.02936.x/abstract

Shedd, K. R., T. H. Dann, C. Habicht, and W. D. Templin. 2014. Alaska Hatchery Reserach Program Technical Document 1: Defining relative reproductive success: which fish count? ADF & G Technical Document:10.

Smith, C. T., J. Baker, L. Park, L. W. Seeb, C. M. Elfstrom, S. Abe, and J. E. Seeb. 2005a. Characterization of 13 single nucleotide polymorphism markers for chum salmon. Mol. Ecol. Notes:259-262. http://www.researchgate.net/publication/227610256_Characterization_of_13_single_nucleotide_polymorphism_markers_for_chum_salmon/file/79e4150ed8f04e96d3.pdf

Smith, C. T., C. M. Elfstrom, J. E. Seeb, and L. W. Seeb. 2005b. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Molecular Ecology 14:4193-4203. http://doc.nprb.org/web/publication/project_0205-0303_seeb_mol_ecol_2005.pdf

Tokarska, M., T. Marshall, R. Kowalczyk, J. Wójcik, C. Pertoldi, T. Kristensen, V. Loeschcke, V. R. Gregersen, and C. Bendixen. 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. http://www.nature.com/hdy/journal/v103/n4/full/hdy200973a.html

Vercessi, L. 2013. Alaska salmon fisheries enhancement program 2012 annual report. Alaska Department of Fish and Game, Fishery Management Report No. 13-05, Anchorage. http://www.adfg.alaska.gov/FedAidPDFs/FMR13-05.pdf

Tables

Table 1.–188 chum salmon SNPs available for use in SEAK chum salmon parentage analyses

Assay	Source[a]	Assay	Source[a]	Assay	Source[a]
Oke_PPA2-635	A	Oke_gdh1-234	B	Oke_ras1-249	A
Oke_ACOT-100	B	Oke_gdh1-62	B	Oke_RFC2-618	C
Oke_AhR1-278	A	Oke_GHII-3129	A	Oke_RH1op-245	C
Oke_AhR1-78	A	Oke_glrx1-78	B	Oke_ROA1-209	B
Oke_APOB-60	B	Oke_GNMT-100	B	Oke_RPN1-80	B
Oke_arf-319	C	Oke_GnRH-373	E	Oke_RS27-81	B
Oke_ATP5L-105	B	Oke_GPDH-191	C	Oke_RS27-94	B
Oke_ATP5L-248	B	Oke_GPH-105	A	Oke_RS9-379	B
Oke_azin1-90	B	Oke_GPH-78	A	Oke_RSPRY1-106	D
Oke_brd2-118	D	Oke_H2AX-72	B	Oke_serpin-140	C
Oke_brp16-65	B	Oke_hmgb1-66	B	Oke_slc1a3a-86	B
Oke_CATB-60	B	Oke_hnRNPL-239	A	Oke_sylc-90	B
Oke_ccd16-77	D	Oke_HP-182	A	Oke_TCP1-78	A
Oke_CCT3-143	A	Oke_HSP90BA-299	A	Oke_TCTA-202	B
Oke_CCT3-220	A	Oke_IGFI.1	C	Oke_TCTA-99	B
Oke_CD123-62	B	Oke_il-1racp-67	C	Oke_Tf-278	A
Oke_CD81-108	B	Oke_IL8r2-406	B	Oke_thic-84	B
Oke_CD81-173	B	Oke_IL8r-272	E	Oke_txnrd1-74	B
Oke_cjo57-86	B	Oke_KPNA2-87	A	Oke_u0602-244	D
Oke_CKS1-70	B	Oke_lactb2-71	B	Oke_U1001-79	D
Oke_CKS1-94	B	Oke_lamp2-138	B	Oke_U1002-165	D
Oke_CKS-389	E	Oke_LAMP2-186	B	Oke_U1002-262	D
Oke_CO1A1-72	B	Oke_mcfd2-86	B	Oke_U1008-83	D
Oke_CO1A1-76	B	Oke_METK2-97	B	Oke_U1010-154	D
Oke_col1a2-62	B	Oke_mgll-49	B	Oke_U1010-251	D
Oke_Cr30	E	Oke_MLRN-63	B	Oke_U1012-241	D
Oke_Cr386	E	Oke_Moesin-160	C	Oke_U1012-60	D
Oke_ctgf-105	A	Oke_nc2b-148	B	Oke_U1015-255	D
Oke_CTR2-82	B	Oke_ND3-69	E	Oke_U1016-154	D
Oke_DBLOH-79	B	Oke_ndub3-58	B	Oke_U1017-52	D
Oke_DCXR-87	B	Oke_NHERF-123	B	Oke_U1018-50	D
Oke_DM20-548	E	Oke_NHERF-54	B	Oke_U1019-218	D
Oke_e2ig5-50	B	Oke_NUPR1-70	B	Oke_U1020-75	D
Oke_EF2-394	B	Oke_PDIA3-475	B	Oke_U1021-102	D
Oke_EIF4EB	C	Oke_PDIA3-82	B	Oke_U1022-114	D
Oke_eif4g1-43	B	Oke_pgap-111	B	Oke_U1022-139	D
Oke_f5-71	B	Oke_pgap-92	B	Oke_U1023-147	D
Oke_FANK1-166	B	Oke_pnrc2-78	B	Oke_U1024-113	D
Oke_FANK1-96	B	Oke_psmd9-188	B	Oke_U1025-135	D
Oke_FBXL5-61	B	Oke_psmd9-57	B	Oke_U1027-89	D
Oke_gdh1-191	B	Oke_rab5a-117	B	Oke_U1028-100	D

224 Table 1.–page 2 of 2.

Assay	Source[a]	Assay	Source[a]	Assay	Source[a]
Oke_U1031-132	D	Oke_U2026-64	B	Oke_U2057-80	B
Oke_U1103-150	B	Oke_U2029-79	B	Oke_U212-87	C
Oke_u1-519	E	Oke_U2031-37	B	Oke_U216	C
Oke_U2001-629	B	Oke_U2032-74	B	Oke_u217-172	C
Oke_U2002-200	B	Oke_U2033-122	B	Oke_u200-385	C
Oke_U2003-142	B	Oke_U2034-55	B	Oke_U302-195	A
Oke_U2005-62	B	Oke_U2035-54	B	Oke_U502-241	A
Oke_U2006-109	B	Oke_U2037-76	B	Oke_U503-272	A
Oke_U2007-190	B	Oke_U2038-32	B	Oke_U504-228	A
Oke_U2010-94	B	Oke_U2040-77	B	Oke_U505-112	A
Oke_U2011-107	B	Oke_U2041-84	B	Oke_U506-110	A
Oke_U2015-151	B	Oke_U2042-61	B	Oke_U507-286	A
Oke_U2016-118	B	Oke_U2043-51	B	Oke_U507-87	A
Oke_U2017-87	B	Oke_U2045-43	B	Oke_U509-219	A
Oke_U2019-112	B	Oke_U2047-49	B	Oke_U510-204	A
Oke_U202	C	Oke_U2048-91	B	Oke_U511-271	A
Oke_U2020-51	B	Oke_U2049-99	B	Oke_U514-150	A
Oke_U2021-86	B	Oke_U2050-101	B	Oke_UBA3-245	D
Oke_U2022-101	B	Oke_U2052-56	B	Oke_uqcrfs-69	B
Oke_U2023-99	B	Oke_U2053-60	B	Oke_XBP1-82	B
Oke_U2024-93	B	Oke_U2054-58	B	Oke_zn593-152	B
Oke_U2025-86	B	Oke_U2056-90	B		

225 a A= Elfstrom et al. 2007; B= Petrou et al. 2013; C= Smith et al. 2005b; D= Seeb et al. 2011; and E= Smith et al.
226 2005a.
227

Figures

Figure 1.–Map of 4 chum salmon pedigree streams in SEAK.