Chinook Baseline Expansion with Genome-Wide SNPs

Final Report to the Pacific Salmon Commission
April 9, 2013

by

Shawn Narum

Columbia River Inter-Tribal Fish Commission
Hagerman Fish Culture Experiment Station
3059-F National Fish Hatchery Road
Hagerman ID 83332

Summary:

The objective of this project was to characterize thousands of single nucleotide polymorphism (SNP) loci in collections of Chinook salmon from populations within and outside the Columbia River Basin to add to existing genetic baselines. A total of 48 populations were included in this project (22 from the coastwide range and 26 from the Columbia River) and analyzed with RAD-seq protocols to identify SNPs. More than 15,000 SNPs were discovered from the full data set and these markers will be highly useful for expanding genetic baselines in the Columbia River and coastwide.

Introduction

The use of SNPs is becoming increasingly popular for population analyses that were previously dominated by (neutral) μSAT markers. Many studies have compared the relative utility of both marker types (Liu et al. 2005; Morin et al. 2009, Smith and Seeb 2008, Hess et al. 2010), and in most observations SNPs perform equal that of μSATs, though a larger number of SNP loci are necessary to reach the same level of resolution. The use of SNPs provides many advantages over μSAT's, in that they are more prolific in the genome, with greater coverage for linkage analyses (Moen et al. 2008). More importantly, because SNPs may be located within functional genes, they are candidates for detecting positive selection or selective divergence shaping population differences. Chinook salmon in the CRB have been studied in great detail (Waples et al. 2004; Beacham et al. 2006; Narum et al. 2010), and our efforts are likely to provide additional information that will benefit and expound on the characterization the species (Matala et al. 2010). SNP discovery and assay development has provided the resources (e.g., Campbell and Narum 2008a) to include additional genetic markers to the existing GAPS microsatellite baseline (Seeb et al. 2007). New technology is available that allows for vastly increasing the number of SNPs that can be added to baseline populations. This new technology uses an approach called Restriction-site Associated DNA sequencing (RAD-seq) to discover and genotype thousands of SNPs in baseline samples. This should provide a nearly unlimited number of powerful markers for GSI purposes and greatly improve the resolution and accuracy of mixed stock analyses.

This study was intended to improve upon the existing genetic baseline, and follows from Recommendation 12 in the PSC Expert Panel Report (Hankin et al. 2005), that there is a need for support of an "immediate evaluation of a coordinated transition for all salmon species from GSI based on the use of microsatellite markers to GSI based on SNP markers."

Current Project Objectives

The objective of this project was to characterize thousands of single nucleotide polymorphism (SNP) loci in collections of Chinook salmon from populations within and outside the Columbia River Basin to add to existing genetic baselines. A total of 48 populations were included in this project (22 from the coastwide range and 26 from the Columbia River; Table 1) and analyzed with RAD-seq protocols to identify SNPs.

SNP Development Methods

Tissue samples from each individual were processed with Qiagen DNeasy® kits to extract DNA from fin clips stored in 100% ethanol. For many samples outside of the Columbia River Basin, DNA was received from WDFW as part of a previous project to genotype core

populations of Chinook salmon throughout the range (Table 1). In order to genotype tissues at thousands of SNPs, samples were prepared for library construction with restriction-site associated DNA (RAD) protocols (Baird et al. 2008) as described by Narum et al. (2013). Briefly, DNA was digested with Sbf1 and subsequently ligated with both a barcode adapter and an Illumina sequencing adapter. Barcoded adapters allowed an average of 48 individuals to be pooled in single libraries for sequencing on an Illumina HiSeq 2000 instrument with single-end 100 reads. Pooling strategy was determined based upon the number of expected reads from reagents available from Illumina for sequencing on the HiSeq 2000 instrument. Samples were sequenced to reach a target of 2 million reads per individual and data was analyzed with STACKS pipeline (Catchen et al. 2010) to identify and genotype SNP markers.

Results & Discussion

In order to develop an extensive number of SNP markers for baseline expansion in Chinook salmon, a RAD-seq approach was used on over 2,200 samples collected throughout the coastwide range. Collections of Chinook salmon were chosen to represent the major regions and genetic lineages identified in previous studies (e.g, Seeb et al. 2007; Moran et al. 2013). Of the total samples included, 86% reached the minimum target number of quality sequence reads (2 million; Table 1). This high level of data quality allowed for strong samples sizes for each collection to be included for the SNP discovery pipeline. Briefly, 100 bp reads were trimmed from the 3' end to 80bp to remove the portion of the read that is most prone to sequencing error and to reduce the probability of observing more than three SNPs in a single read which can increase the number of false SNPs discovered. Trimmed reads were filtered using quality scores to eliminate poor quality reads and those that contained one or more ambiguous base calls. From the 5' end of the read, the six base barcode sequence and partial *SbfI* site (TGCAGG) sequence were also removed after reads had been separated based on their unique barcode sequence. Samples from all collections were included in a sequence alignment catalog and the STACKS pipeline identified greater than 15,000 SNPs throughout the coastwide set of samples. Many SNPs were highly polymorphic suggesting that they will be highly informative to identify specific stocks.

The large number of SNPs identified in this project greatly expands the number of genetic markers that are available across the majority of this species' range. These SNP markers are expected to be useful for distinguishing specific stock structure, identifying origin of unknown fish for genetic stock identification, and investigating local adaptation of populations to their environments. The addition of these SNPs greatly increases the number of genetic markers over the 192 SNPs that are currently included in the Columbia River basin, and 96 SNPs coastwide.

Quality Control

Sequence data was tested under standard quality control procedures in CRITFC's genetic laboratory. This includes positive controls, known samples to estimate sequencing error rate, and identification of duplicated genes with doubled haploid individuals and Hardy-Weinberg tests. Loci and samples with excessive missing data were also excluded from the data set to reduce the potential for error.

Project Benefits / Monitoring and Evaluation

The objective of this study was intended to improve both regional and coastwide GSI applications. These additional SNPs will benefit all agencies as these markers are available for the entire genetics community.

Acknowledgements

We thank Stephanie Harmon, Travis Jacobson, Amanda Matala, Lori Maxwell, and Nate Campbell for assistance in the genetics lab. Ben Hecht provided support with bioinformatics and STACKS pipeline. Samples and DNA for genotyping were provided by multiple agencies including ADFG, OSU, NOAA, IDFG, WDFW, USFWS, and UW. DNA was extracted for several collections by WDFW staff.

Table 1. Collections of Chinook salmon for baseline expansion. 40 collections were analyzed with Pacific Salmon Commission funding and 8 with funding from other sources.

Region	Location	RAD_Library	avg.ind.bc.reads	n	#>2M
BC	Alsek	RL0053	3,166,134	47	41
CA	Battle	RL0032	3,810,169	46	46
Columbia	Capehorn	RL0091	2,898,011	38	35
Columbia	Catherine	RL0038	2,310,233	48	28
Columbia	Chamberlain	RL0065	4,041,135	48	48
Columbia	Clearwater	RL0046	3,863,806	47	46
OR	Cole	RL0036	4,619,507	46	46
Columbia	Cowlitz	RL0033	3,628,146	95	88
Columbia	Deschutes	RL0047	3,036,253	48	42
CA	Eel	RL0037	2,252,085	48	32
AK	George	RL0056	3,109,426	48	47
Columbia	Imnaha	RL0070	4,015,928	48	47
Columbia	JohnDay	RL0045/RL0090	2,463,960	60	34
Columbia	Johnson Cr.	RL0040/RL0087	3,116,023	96	70
AK	Kanektok	RL0028	2,250,702	47	35
AK	Kantishna	RL0027	2,549,968	48	34
AK	Karluk	RL0027	2,371,726	48	25
CA	Klamath	RL0063	2,639,940	48	37
Columbia	Lostine	RL0042	2,498,293	48	40
Columbia	LyonsFerry	RL0043	3,550,101	48	48
WA	Marblemount	RL0031	2,311,522	48	29
Columbia	Marsh	RL0041	2,445,498	48	33
Columbia	McCall	RL0035	4,338,513	47	43
OR	McKenzie	RL0048	3,035,607	48	48
Columbia	Methow	RL0088	3,167,218	48	47
AK	Montana	RL0030	3,266,491	48	40
BC	Morice	RL0055	3,278,556	48	48
Columbia	Newsome	RL0039	2,050,282	48	26
OR	Nestucca	RL0036	4,391,856	45	43
Columbia	Pahsimeroi	RL0089	2,967,177	47	43
Columbia	Priest	RL0033	4,471,061	46	46
AK	Pullen	RL0052	2,877,846	47	39
WA	Quinault	RL0035	4,307,694	48	48
Columbia	Rapid	RL0034	3,994,351	47	46
BC	Robertson	RL0054	3,667,535	48	48
OR	Rock	RL0069	3,856,080	48	48
Columbia	Sawtooth	RL0042	2,472,527	48	47

AK	Sinona	RL0026	2,502,662	48	45
AK	Soos	RL0026	2,799,241	48	48
Columbia	Spring	RL0060	4,107,685	48	48
AK	Tatsamenie	RL0029	2,762,028	47	39
AK	Togiak	RL0057	3,862,691	47	47
Columbia	Tucannon	RL0065/RL0089	2,164,992	96	51
Columbia	Warm Springs	RL0046/RL0088	2,558,580	96	63
Columbia	Wells	RL0058	4,022,373	48	45
Columbia	Wenatchee su/fa	RL0044/RL0090	2,304,650	96	63
Columbia	Wenatchee spring	2012	2,692,703	110	102
Columbia	Yakima	RL0045/RL0087	3,059,939	95	81

References

Baird NA, Etter PD, Atwood TS et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Beacham, T. D., J. R. Candy, K. L. Jonsen, J. Supernault, M. Wetklo, L. Deng, K. M. Miller, and R. E. Withler. 2006b. Estimation of stock composition and individual identification of Chinook salmon across the Pacific Rim using microsatellite variation. Transactions of the American Fisheries Society 135:861-888.

Campbell, N. R., and S. R. Narum. 2008a. Identification of novel SNPs in Chinook salmon and variation among life history types. Transactions of the American Fisheries Society 137:96-106.

Campbell, N. R., and S. R. Narum. 2008b. Quantitative PCR assessment of microsatellite and SNP genotyping with variable quality DNA extracts. Conservation Genetics DOI 10.1007/s10592-008-9661-7.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). Stacks: Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3*, **1**, 171-182.

Hankin, D. G., J. H. Clark, R. B. Deriso, J. C. Garza, G. S. Morishima, B. E. Riddell, C. Schwarz, and J. B. Scott. 2005. Report of the Expert Panel on the future of the coded wire tag recovery program for Pacific salmon. Pacific Salmon Comm.

Hess, J., A. P. Matala and S. R. Narum. 2010. Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. Molecular Ecology Resources 11 (Suppl. 1):1–13, doi: 10.1111/j.1755-0998.2010.02958.x

Liu, N., L. Chen, S. Wang, C. Oh and H. Zhao. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genetics 6(Suppl 1):S26.

Matala, A. P., J. Hess and S. R. Narum. 2010. Resolving adaptive and demographic divergence among Chinook salmon populations in the Columbia River Basin. Transactions of the American Fisheries Society. In Press.

Moen, T., B. Hayes, M. Baranski, P. R. Berg, S. Kjoglum et al., 2008. A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics 9: 223.

Moran et al. 2013…

Morin, P. A., K. K. Martien and B. L. Taylor. 2009. Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources 9:66-73.

Narum, S. R., J.E. Hess, and A.P. Matala. 2010 Examining genetic lineages of Chinook salmon in the Columbia River Basin. Transactions of the American Fisheries Society 139:1465-1477.

Narum, S. R., M. Banks, T.D. Beacham, M.R. Bellinger, M.R. Campbell, J. DeKoning, A. Elz, C.M. Guthrie III, C. Kozfkay, K.M. Miller, P. Moran, R. Phillips, L.W. Seeb, C.T. Smith, K. Warheit, S.F. Young, J.C. Garza. 2008. Differentiating salmon populations at broad and fine geographic scales with microsatellites and SNPs. Molecular Ecology 17:3464-3477.

Narum SR, Campbell NR, Meyer KA, Miller MR, Hardy RW (2013) Thermal adaptation and acclimation of ectotherms from differing aquatic climates. Molecular Ecology DOI: 10.1111/mec.12240.

Seeb, L. W, A. Antonovich, M.A. Banks, T.D. Beacham, M.R. Bellinger, S. M. Blankenship, M. Campbell, N.A. Decovich, J.C. Garza, C.M. Guthrie III, T. A. Lundrigan, P. Moran, S.R. Narum, J.J. Stephenson, K.J. Supernault, D.J. Teel, W.D. Templin,

J.K.Wenburg, S.F. Young, C.T. Smith. 2007. Development of a Standardized DNA Database for Chinook Salmon. Fisheries 32:540-552.

Smith CT, and L. W. Seeb. 2008. Number of alleles as a predictor of the relative assignment accuracy of short tandem repeat (STR) and single-nucleotide-polymorphism (SNP) baselines for chum salmon. Transactions of the American Fisheries Society, 137, 751–762.

Waples, R. S., D. J. Teel, J. M. Myers, and A. R. Marshall. 2004. Life-history divergence in Chinook salmon: historical contingency and parallel evolution. Evolution 58:386-403.