Western Alaska Salmon Stock Identification Program **Technical Document:** [1] 9

Title: Chum salmon SNP discovery – First method **Version:** 1.0

Authors: J. E. Seeb, C. E. Pascal, E. D. Grau, L. W. Seeb, W. D. Templin, T. Harkins, S. B. Roberts

Date: December 14, 2010

Introduction:

Early in the development process for the Western Alaska Salmon Stock Identification Project (WASSIP) it was clear that the resolution possible to distinguish among regional areas for chum salmon spawning in western Alaska regional areas (Norton Sound, lower Yukon and Kuskokwim rivers, and Bristol Bay) was not going to be sufficient to meet the standards set by the Advisory Panel with available genetic markers, including the recently developed SNP markers (see Technical Document 4 and Seeb et al. 2011b for the current panel of 53 SNPs). These four regional areas define important units for management, yet when treated as separate reporting groups the four areas could not be distinguished sufficiently using the 53-marker set. The Department began the process of discovering additional SNP markers for chum salmon through a contract with International Program for Salmon Ecological Genetics (IPSEG; http://www.fish.washington.edu/research/ipseg/research.html) at the University of Washington. These efforts were based on cDNA sequences from two chum salmon sampled from the Susitna (Cook Inlet) and Delta (Yukon River) rivers. This process has been published in *Molecular Ecology Resources* (Seeb et al. 2011a) which is provided here as Technical Document 9 (Appendix A). This process added 37 validated SNPs to those already available for used in chum salmon for WASSIP. The SNPs discovered through this and other efforts will be assayed in 30 populations and a subset of the best 96 SNPs will be used for mixed stock analysis in WASSIP.

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program Technical Committee. 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.

Literature cited:

Seeb, J. E., C. E. Pascal, E. D. Grau, L. W. Seeb, W. D. Templin, S. B. Roberts, and T. Harkins. 2011a. Transcriptome sequencing and high-resolution melt analysis advance SNP discovery in duplicated salmonids. Molecular Ecology Resources doi: 10.1111/j.1755-0998.2010.02936.x.

Seeb, L. W., W. D. Templin, S. Sato, S. Abe, K. I. Warheit, and J. E. Seeb. 2011b. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Molecular Ecology Resources xxx: xx-xx

Specific questions for the Technical Committee:

This document is provided for informational purposes and we have no specific questions. However, any comment or review that you might have would be appreciated.

Appendix A: Seeb et al. 2011.

Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) doi: 10.1111/j.1755-0998.2010.02936.x

PERMANENT GENETIC RESOURCES ARTICLE

Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids

J. E. SEEB,* C. E. PASCAL,* E. D. GRAU,* L. W. SEEB,* W. D. TEMPLIN,† T. HARKINS‡§ and S. B. ROBERTS*

*School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA,
†Gene Conservation Laboratory, Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK 99518, USA,
‡454 Life Sciences, Roche Diagnostics Corporation, 9115 Hague Road, Indianapolis, IN 46256, USA

Abstract

Until recently, single nucleotide polymorphism (SNP) discovery in nonmodel organisms faced many challenges, often depending upon a targeted-gene approach and Sanger sequencing of many individuals. The advent of next-generation sequencing technologies has dramatically improved discovery, but validating and testing SNPs for use in population studies remain labour intensive. Here, we detail a SNP discovery and validation pipeline that incorporates 454 pyrosequencing, high-resolution melt analysis (HRMA) and 5′ nuclease genotyping. We generated 4.59×10^8 bp of redundant sequence from transcriptomes of two individual chum salmon, a highly valued species across the Pacific Rim. Nearly 26 000 putative SNPs were identified—some as heterozygotes and some as homozygous for different nucleotides in the two individuals. For validation, we selected 202 templates containing single putative SNPs and conducted HRMA on 10 individuals from each of 19 populations from across the species range. Finally, 5′ nuclease genotyping validated 37 SNPs that conformed to Hardy–Weinberg equilibrium expectations. Putative SNPs expressed as heterozygotes in an ascertainment individual had more than twice the validation rate of those homozygous for different alleles in the two fish, suggesting that many of the latter may have been paralogous sequence variants. Overall, this validation rate of 37/202 suggests that we have found more than 4500 templates containing SNPs for use in this population set. We anticipate using this pipeline to significantly expand the number of SNPs available for the studies of population structure and mixture analyses as well as for the studies of adaptive genetic variation in nonmodel organisms.

Keywords: 454, chum salmon, high-resolution melt analysis, homeologue, next-generation sequencing, paralogous sequence variant

Received 30 November 2009; revision received 10 September 2010; accepted 30 September 2010

Introduction

For aquatic organisms, particularly salmonids, population genetics has been widely used for resource management (Waples *et al.* 2008). Microsatellites were extensively applied to population and conservation genetic studies over the last decade because of their high variability and power to resolve population structure (Narum *et al.* 2008). However, several properties associated with microsatellites including complicated mutation rates, presence of null alleles, high potential genotyping error rate (Seeb *et al.* 2007; Stephenson *et al.* 2009) and low throughput have led salmonid researchers to seek alternative markers. Currently, investigators working with a variety of organisms including salmonids are choosing to develop population data sets using single nucleotide polymorphisms (SNPs) given their lower error rates, increased automation of sample processing, potential for genome-wide scans of either selectively neutral or adaptive variation and ability to use historical and low-quality material (Everett *et al.* 2011; Hemmer-Hansen *et al.* 2011; Olsen *et al.* 2011).

Until recently, SNP discovery has been slow, often depending upon the targeted-gene approach and Sanger (chain-termination) sequencing (Smith *et al.* 2005a,b; Elfstrom *et al.* 2007). Recent advances in molecular chemistries now provide vast availability to sequence information from any organism through next-generation sequence technologies; the most commonly available platforms generate gigabases of short-read sequence in a

Correspondence: James E. Seeb, Fax: +206 543 5728;
E-mail: jseeb@uw.edu
§Present address: Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA.

single run. These technologies remove many of the impediments to characterizing large numbers of SNPs in nonmodel organisms (Primmer 2009; Storfer *et al.* 2009), although assembly and alignment of the relatively short reads can be challenging (Everett *et al.* 2011). The combination of new genetic technologies allowing the sequencing of the entire transcriptome along with novel algorithms for sequence analyses provides a catalyst for a wide variety of population genetics and genomics studies (Hudson 2008; Hale *et al.* 2009; Slate *et al.* 2009). Hauser & Seeb (2008) predicted that these technological gains would allow the development of thousands of SNPs as genetic markers for population identification and mixture analyses as well as for the studies of adaptive genetic variation. Further, the availability of large numbers of SNPs and the identification of outlier loci with high levels of differentiation hold particular promise for applications such as compositional analyses of admixtures of closely-related populations, assignment tests of individuals to population of origin or verification of specific selection pressures and local adaptation (e.g. Vasemagi & Primmer 2005; Smith & Seeb 2008; Karlsson *et al.* 2011).

Concomitant with the access to this wealth of DNA sequence comes the nontrivial confirmation of reliable and assayable SNP markers, derived from the assembly and alignment of short reads, which may require several validation steps. False positives may result from sequencing error (Harismendy *et al.* 2009) or the occurrence of paralogous sequence variants (PSVs) that occur in gene duplicates (Renaut *et al.* 2010). Assembling DNA sequence reads in salmonids can be especially daunting because of the large number of paralogous sequences that remain from the tetraploid event 25 MYA (Allendorf & Danzmann 1997; Koop *et al.* 2008). While stringent assemblies eliminate many PSVs, high stringency (i.e. ≥0.98) will eliminate highly polymorphic sequences from the assembly such as the MHC classes of genes that are especially informative in population genetics. For example, Miller & Withler (1996) report that SNPs in *MHCII* occur at frequencies approaching 0.1. Finally, valid SNPs discovered in transcriptomes may not lend themselves to high-throughput genotyping if they are located proximal to intron/exon boundaries that disrupt PCR.

Here, we outline a pathway for SNP discovery and validation of high-throughput assays using chum salmon (*Oncorhynchus keta*). Chum salmon has the widest natural distribution of all of the Pacific salmon species, ranging northward from Korea in the western Pacific Ocean, through the Chukchi Sea in the Arctic, and south to the northwest United States in the eastern Pacific Ocean. The species provides a significant component of many commercial and subsistence fisheries in Asia and North America. Some wild populations on both sides of the Pacific Ocean have become progressively more threatened as a result of hatchery production, habitat loss and climate change (see Farley *et al.* (2009) and references therein). Consequently, management agencies increasingly rely upon molecular markers to clarify population structure, study oceanic distribution, characterize migration patterns and identify the population components of fishery harvests in attempts to conserve depressed populations (Sagarin *et al.* 2009).

Differentiating closely related stocks of chum salmon that were connected historically but currently distributed in separate large river systems has been an ongoing challenge (Seeb *et al.* 2004; Smith & Seeb 2008). Developing DNA markers for chum salmon that are easily standardized and transferred across laboratories is particularly important because of the species' broad distribution, susceptibility to bycatch and complicated management guided by international treaties. Laboratories in Korea, Japan, Russia, Canada, and the USA use SNPs to study the populations of salmon, but high-throughput assays for only 60 loci are currently available for chum salmon across the Pacific Rim (Seeb *et al.* 2011).

Many laboratories genotype salmonid populations using 96-SNP arrays (Seeb *et al.* 2009), so additional SNPs will increase both efficiency and resolution. In this study, we use 454 FLX pyrosequencing to characterize a portion of the chum salmon transcriptome; we conducted three sequencing runs on cDNA from testes from each of two chum salmon from different populations (cf., Ellegren 2008; Vera *et al.* 2008) where we detected thousands of putative SNPs. We present a stringent validation pathway including PCR test, high-resolution melt analysis (HRMA; Wu *et al.* 2008; see also McGlauflin *et al.* 2010), Sanger resequencing and population genotyping where we develop 37 new high-throughput assays for the commonly used 96-array. We believe that this discovery and validation pipeline will have general applicability for SNP discovery in a broad range of nonmodel organisms.

Methods

Our approach first includes discovery and annotation followed by a four-step validation process. High-throughput genotyping assays were developed for population screening for the fourth step, and final validation that a variant was a true SNP was confirmed by goodness-of-fit to Hardy–Weinberg equilibrium (HWE) expectations.

Samples and sequencing

Two male chum salmon were collected—one from the Susitna River (August 2007) and one from the Delta River (November 2007) in Alaska (Fig. 1). Testes tissue from freshly harvested fish was immediately frozen on dry ice

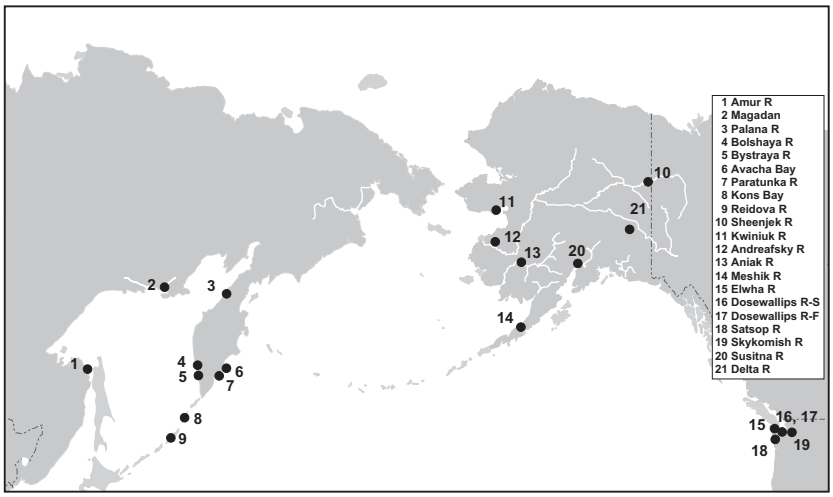

Fig. 1 Location of test populations of chum salmon. Individuals included in the 454 FLX runs originated from the Susitna (20, SUSI) and Delta (21, DELT) rivers in Alaska.

and stored at −80 °C. Total RNA was extracted following a method combining Tri Reagent (Sigma) and RNeasy Mini (Qiagen) columns as described by Eccles (2008). Normalized cDNA libraries from each individual fish were constructed by Evrogen, Moscow (http://www.evrogen.com; see Zhulidov *et al.* (2004)). Pyrosequencing was conducted on a Roche Genome Sequencer FLX at 454 Life Sciences (http://www.454.com/). A total of six plates were sequenced, three from each of the two libraries. For simplicity, we will refer to these fish (and corresponding libraries) as 'SUSI' and 'DELT' based on collection site.

Assembly and annotation

De novo assembly criteria were 50% sequence overlap with a minimum of 0.9 similarity. The 0.9 similarity criterion, also used by Sanchez *et al.* (2009) to assemble 454 reads from the duplicated rainbow trout, will exclude some PSVs while allowing assembly of sequences containing multiple SNPs. Reads with <0.9 similarity are expected to be infested with PSVs (Renaut *et al.* 2010).

Sequence assembly was carried out using CLC Genomics Workbench version 4.0 (CLC Bio). Prior to analysis, all sequences were trimmed using quality scores (0.05 Phred; Ewing & Green (1998); Ewing *et al.* (1998)). Reads that aligned to more than one location were not assembled.

Candidate SNPs were also identified using CLC Bio (Altshuler *et al.* 2000; Brockman *et al.* 2008). Parameters were as follows: window length = 11, maximum gap and mismatch count = 2, minimum average quality of surrounding bases = 15, minimum quality of central base = 20, maximum coverage = 100, minimum coverage = 8, minimum variant frequency (%) = 35.0, maximum expected variations (ploidy) = 2, (Fig. 2a). Putative SNPs that were homozygous for different alleles in SUSI and DELT were identified by comparing consensus sequences.

To gauge the number of transcripts and gene function, BLASTX was carried out on consensus sequences generated from assembly of all reads (Fig. 2b). We used NCBI's Non-Redundant and Gene Ontology (GO) databases. GO terms were analysed using CateGOrizer (Hu *et al.* 2008).

Validation

SNP selection and primer testing. To further characterize a subset of SNPs identified from sequence analysis, 202 candidate SNPs were selected based on presumed genotypes in SUSI and DELT and, where possible, annotation to gene families likely to include candidate genes (e.g., Nielsen *et al.* (2005); Kasahara *et al.* (2007)). Of these 202 putative SNPs, 93 were presumed heterozygotes in SUSI and 109 were homozygous for different alleles in SUSI and DELT. PCR primers were designed using Primer3 (Rozen & Skaletshy 2000) to amplify a template approximately 200 bp long that contained a single putative SNP in the ascertainment fish. A PCR test was carried out using 2× LightCycler480 High Resolution Melting Master (Roche Applied Science) following manufacturer's instructions. Primer pairs that produced a single, clean amplicon (Fig. 2c) were passed to the second validation step.

High-resolution melt curve analysis. For the second validation step, an HRMA (Wu *et al.* 2008) was performed in a fashion that would also screen for variability of putative SNPs among populations (Fig. 2d). Tissue samples (muscle, liver, or fin depending upon source) were obtained from archives maintained by colleagues from 10 individuals from each of 19 geographical locations spanning the species' range (Fig. 1). Genomic

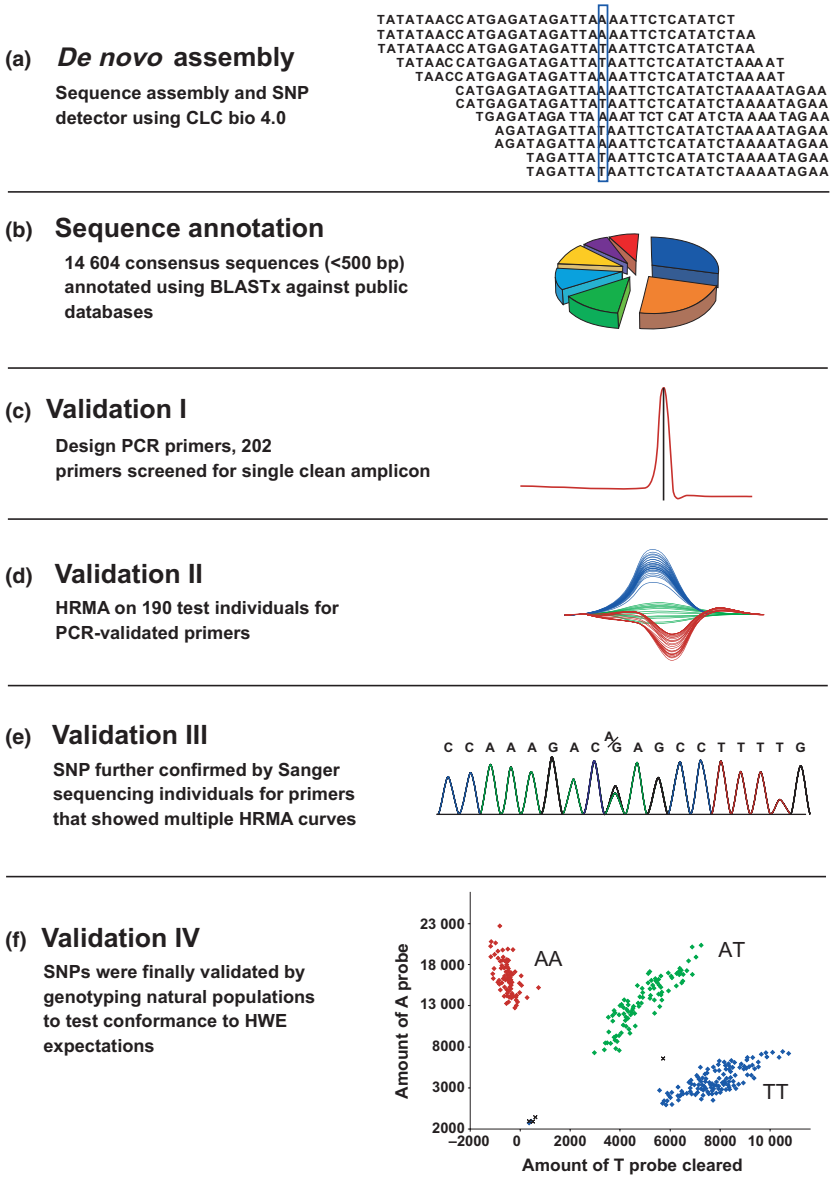

Fig. 2 Diagram of the workflow to identify and validate single nucleotide polymorphisms in chum salmon using transcriptome sequencing.

DNA was extracted using DNeasy 96 Blood and Tissue kits (Qiagen) and quantified using Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) following manufacturer's instructions. Fluorescence was measured in a 200-μL reaction on a VICTOR3 multilabel microplate reader (Perkin Elmer). DNA concentrations were then normalized for both HRMA and later Sanger sequencing. PCR was conducted in a 10-μL volume containing 10 ng of genomic DNA, 1× LightCycler 480 High Resolution Melting Master (Roche Applied Science), 3.5 mM of MgCl2 and 0.2 μM of each PCR primer. Thermal cycling was performed on a Veriti 384-Well Thermal Cycler (Applied Biosystems; AB) as follows: 95 °C hold for 10 min followed by 45 cycles of 95 °C for 15 s, 60 °C for 15 s and 72 °C for 15 s. The plates were transferred to a LightCycler 480 Real-Time PCR System (Roche Diagnostics) after PCR and heated to 95 °C for 1 min and then cooled to 40 °C for 1 min. HRMA data were collected between 62 and 95 °C at 25 acquisitions per degree celsius, using a ramp rate of 0.02°C per second. The amplicons were analysed for the presence of discrete melt curve families, signalling the presence of SNPs, using the LightCycler 480 Gene Scanning Software v. 1.5.0 SP1 (Roche Diagnostics).

Sanger sequencing. Templates that demonstrated multiple HRMA curve families were sequenced using Sanger sequencing to reconfirm the base call of the polymorphisms as observed in the 19 test populations (Fig. 2e). These sequences were also used to evaluate primer and probe sequences for the high-throughput assay design.

A total of 12 individuals were selected from the array of curve families observed for each polymorphic HRMA. These individuals were sequenced in both directions using ABI PRISM BigDye Terminator version 3.1 Cycle Sequencing Kit and analysed on a 3730 DNA Analyzer (AB) by High-Throughput Sequencing Solutions (UW, Genome Sciences). Sequence chromatograms were aligned and visually screened for polymorphisms using Sequencher 4.9 (GeneCodes Corporation). BLASTX was carried on these genomic sequences to verify the original cDNA annotations.

High-throughput genotyping and population analysis. Assay designs for the 5′-nuclease reaction (Holland *et al.* 1991) were attempted for 53 of 54 putative SNPs taken from Sanger-validated sequences (no design was attempted for the polymorphism in template Oke054 which was observed only as a single variant in a single population). Although the ascertainment fish contained only a single putative SNP in each template, several polymorphisms were identified in some templates when the 190 test fish were subjected to HRMA. Linkage disequilibrium has been shown to add power for population discrimination (Morin *et al.* 2009) even for tightly linked SNPs (Habicht *et al.* 2010). For this reason, we designed assays for five pair of linked SNPs (10 total), selected at random, from individual sequences that were found to contain multiple polymorphisms. Twenty-five sequences contained a single polymorphism, so a single assay was attempted for each. Finally, assays were developed for a single candidate SNP in the 18 remaining sequences that contained multiple polymorphisms. Uniplex reactions for primer and probe tests were carried out on a 7900HT Fast Real-Time PCR System (AB).

All successfully designed assays were used to genotype up to 95 fish from each of three populations: Reidova Bay, Kamchatka, Russia; Nitinat River, British Columbia, Canada; and Kitoi Bay, Alaska, USA. Parallel uniplex reactions for genotyping were performed on the BioMark 96-array (Fluidigm) following the methods of Seeb *et al.* (2009).

As final validation of a polymorphism as a true SNP, we used GENEPOP V4 (Rousset 2008) to perform exact tests for fit to HWE expectation. We also used GENEPOP V4 to calculate Fisher's tests for genotypic linkage disequilibrium between each pair of loci across samples.

Results

454 GS-FLX sequencing

Pyrosequencing generated over 800 000 reads in SUSI and over 1 000 000 reads in DELT (Table 1, NCBI Sequence Read Archive accession number SRP001388.1). Average read length was 245 bp, providing 4.59×10^8 bp of redundant transcriptome. *De novo* assembly of these reads resulted in 118 546 contiguous sequences with an average length of 412 bp. Average length of the reads that assembled was 246 bp. The average coverage depth was approximately five.

Table 1 Sequence results from three 454 FLX runs on testes cDNA from each of two sexually mature chum salmon: one from the Susitna River (SUSI) and one from the Delta River (DELT), Alaska (NCBI Sequence Read Archive accession number SRP001388.1)

Metric	SUSI	DELT	Combined
Total bases	1.96×10^8	2.63×10^8	4.59×10^8
Reads			
Total	810 125	1 063 174	1 873 299
Average length	242	247	245
Assembled	675 906	851 994	1 539 224
Singletons	134 219	211 180	334 075
Contigs			
Total	52 678	86 824	118 546
Average depth	5.7	4.8	5.9
Average length	426	409	412
Average read number	12.8	9.8	13.0
SNPs			
Total	8646	11 744	25 995
Fixed differences			16 376

Putative single nucleotide polymorphisms (SNPs) were scored in contig assemblies using a filter that required the minimum depth of coverage = 8 and the minimum variant frequency = 35%. Putative SNPs scored in the individuals do not sum to the combined number for several reasons: (i) some SNPs were common to both individuals, (ii) some SNPs had a depth of coverage <8 in the individual assembly but >8 in the combined assembly and (iii) some new SNPs were fixed for different bases in SUSI and DELT. Several thousand potential SNPs were observed in the combined assembly with depth of coverage <8.

Based on our selection criteria, we identified 8646 putative SNPs within SUSI and 11 744 putative SNPs within DELT (Table 1). The combined assembly of reads from both SUSI and DELT provided additional contigs with coverage >8 that contained additional putative SNPs for a total of 25 995. Finally, SUSI and DELT were homozygous for alternate nucleotides at 16 376 positions.

Consensus sequences were compared to public databases at NCBI using BLASTX (Table S1, Supporting Information). To get a general idea of the biological processes of this transcriptome, all consensus sequences were compared to the GO database and parent GO terms identified. A majority of consensus sequences identified in this study are involved in macromolecule, nucleobase, nucleoside, nucleotide, nucleic acid or general metabolism (32.7%) (Fig. 3). Even though these libraries were generated from testes tissue, less than 1% of the annotated

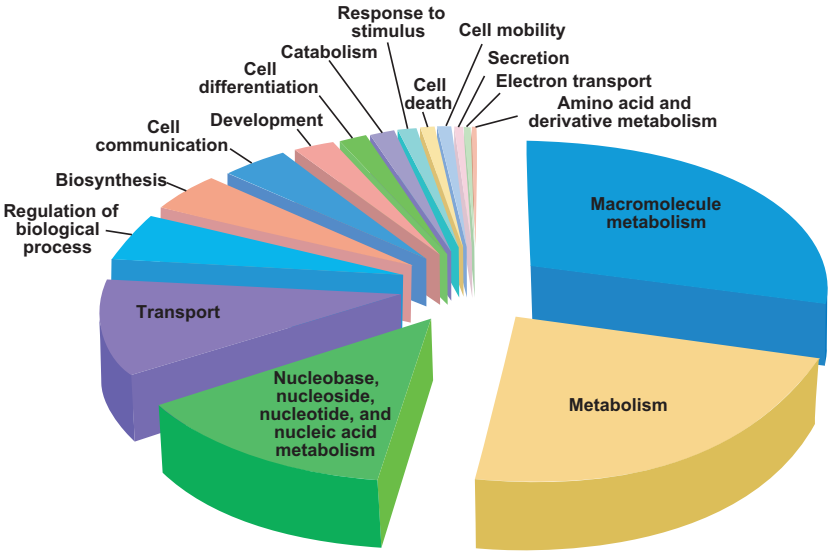

Fig. 3 Annotation. Categorization of consensus sequences into cellular processes derived from BLASTX comparisons to the Gene Ontology Database (version GO.200801). Sequences without annotation were not included in this analysis.

transcripts are directly involved in gamete production and motility.

To estimate transcriptome diversity and provide insight into the number of unique gene families, consensus sequences >500 bp were compared to public databases at NCBI using BLASTX and grouped using equivalent top BLAST. Of the 14 604 sequences subjected to annotation, there were 5841 unique gene descriptions with 1635 sequences having no significant BLAST hit (*e*-value threshold 1.0E-20). Furthermore, we set out to determine how many sequences produced in our pyrosequencing effort had not been previously sequenced in a salmonid. To address this, all reads were assembled onto all Salmonidae ESTs in NCBI's GenBank, and singletons were then subjected to *de novo* assembly. Using this approach, 7105 sequences were identified that have not been previously sequenced in a salmonid. Of these sequences, 18% were most similar to either *Danio rerio* or *Tetraodon nigroviridis* sequences.

SNP validation using PCR, HRMA and Sanger sequencing

PCR primers were designed to test a total of 202 SNP-containing templates (Table 2). Eighty-two primer pairs were successful; the success rate of the initial PCR test was 0.61 for primers based upon templates that contained putative heterozygotes in SUSI and 0.23 for primers based upon templates homozygous for different nucleotides in SUSI and DELT.

These 82 primer pairs were used for HRMA where a total of 69 templates were identified to contain putative SNPs. HRMA curve plots sometimes were clear enough to infer genotypes, especially if a single SNP was present (Fig. 4a,b). Several templates produced complicated families of curves suggesting the presence of multiple polymorphisms.

Sanger sequencing confirmed the presence of sequence polymorphisms in 49 templates (Table 2). In

Table 2 Results of primer design and validation of candidate single nucleotide polymorphisms (SNPs)

Alignment and primer pairs	Primer pairs	Successful PCR	HRMA validation	Sanger validation	5′ nuclease validation
SUSI (Oke001–Oke116)	93	57	46	31	22
SUSI vs. DELT (Oke117–Oke236)	109	25	23	18	15
Total	202	82	69	49	37

Ninety-three primer pairs were designed to test SNPs scored as putative heterozygotes in the alignment for SUSI, and 109 primer pairs were tested for SNPs scored as putative homozygotes for alternate alleles in SUSI and DELT. These putative SNPs were subjected to four validation steps. First, we tested for successful PCR to produce a single, clean amplicon. Second, the successful primer pairs were subjected to high-resolution melt analysis (HRMA) using a suite of 190 individuals from 19 populations. Third, templates that demonstrated multiple curve families during HRMA were resequenced using Sanger sequencing. Finally, putative SNPs appearing in the Sanger sequences were validated using 5′ nuclease genotyping in three test populations. SNPs were considered validated if they were observed at frequencies >0.02 and conformed to Hardy–Weinberg equilibrium expectations.

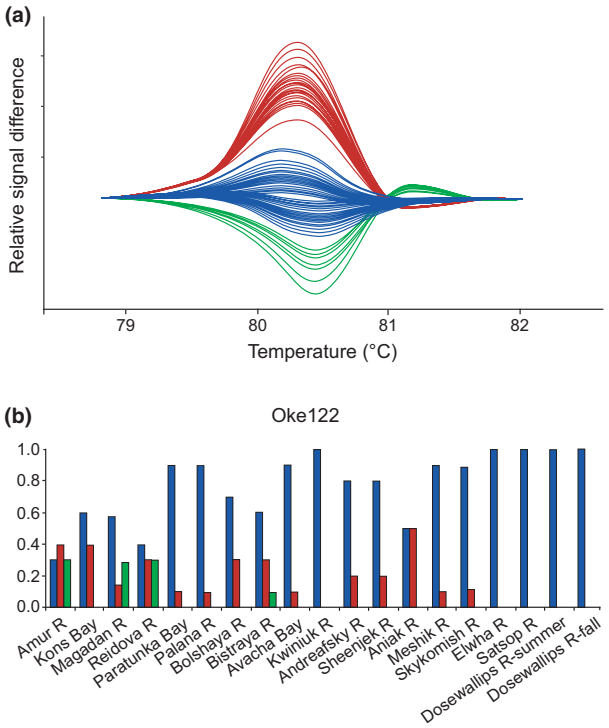

Fig. 4 High-resolution melt analysis for primer pair Oke122. (a) Three curve families suggest a single nucleotide polymorphism (SNP) and provide an indirect assessment of SNP genotypes. For this example, we interpret green curves as XX, red as XY and blue as YY. (b) Histogram showing distribution of inferred genotypes across 19 test populations in Asia and North America. The XX genotype is seen only in Russian populations, and the YY genotype predominates in the North American populations. Sanger sequencing confirmed a single SNP where green = TT, red = TC and blue = CC.

total, there were 106 putative SNPs across these 49 templates (Table 3). The number of polymorphisms per template ranged from one to eight with an average of two. All 49 genomic DNA sequences were compared to NCBI's Non-Redundant database using BLASTX (Table 3). Each SNP in sequences with *e*-values <1.0E-10 was determined to be synonymous or nonsynonymous based on visual inspection of aligned open reading frames from BLAST output. Of the 23 polymorphisms that were able to be characterized in this manner, six were nonsynonymous.

Genotyping, Hardy–Weinberg and linkage disequilibrium

We successfully developed 5′ nuclease assays for 51 candidate SNPs—primer and probe sequences could not be designed for two (sequences and GenBank accession numbers in Table S2, Supporting Information). All 51 were screened in the three test populations (Table 4).

Fourteen candidate SNPs were removed from further consideration. Six of the 14 candidates were completely monomorphic in the three populations that were genotyped. Eight candidates yielded highly significant test statistics for deviations from HWE across the majority of test populations (*Oke_U0603-101*, *Oke_U1030-82*, *Oke_U1003-335*, *Oke_ZNF-87*, *Oke_U0604-70*, *Oke_U1013-95*, *Oke_TBAT-94*, *Oke_U1026-142*). With the exception of one assay, *Oke_U0603-101*, excess of heterozygotes was observed as evidenced by negative F_{IS} values (Table 4).

Of the remaining 37 candidates, 32 always conformed to HWE when polymorphic. Five (*Oke_U1012-60*, *Oke_Estl3-95*, *Oke_rabl2-84*, *Oke_brd2-118* and *Oke_U1015-255)* conformed to HWE in two of three tests.

Tests for linkage disequilibrium were performed for four of the five pairs of candidate SNPs that originated from a within single DNA sequence. The assay for *Oke_U1010-154* was monomorphic in all test populations, so this pair could not be tested. Tests were significant ($P < 0.05$) in all possible tests for each pair with a single exception. The Russian population was in linkage equilibrium for the paired SNPs *Oke_U1002-165* and *Oke_U1002-262* ($P = 0.29$), unlike the other two test populations.

We also performed tests for linkage disequilibrium between all loci in each population. Overall, the number of significant tests was less than expected by chance alone ($P < 0.05$). We did, however, find significant tests ($P < 0.01$) in two populations between both of the paired SNPs in the Oke061 template and *Oke_U1006-137*. We also detected significance in all three populations between *Oke_U1010-251* and *Oke_UBA3-245* ($P < 0.01$).

Discussion

Here, we present a SNP discovery workflow that provides relatively rapid development of SNPs for population genetic studies while also generating a wealth of transcriptomic information. Following *de novo* assembly of 4.59×10^8 bases, approximately 4.8×10^7 bp of consensus sequences were identified. These included several thousand transcripts not present in the over 180 000 sequences from salmonids in the NCBI Expressed Sequence Tag Database. Of particular biological interest are chum salmon genes with significant sequence similarity to a variety of receptors including an IGF-I receptor subtype b, vasoactive intestinal peptide receptor, activin receptor type IIA, serotonin receptor 5A, Mullerian inhibiting substance type II receptor, RFamide-related peptide receptor and a natriuretic peptide receptor type-C. These transcripts would be expected in testes tissue given the importance of hormonal control, and along with the rest of the chum salmon testes transcriptome generated as part of this research effort, will facilitate research into

Table 3 Annotation results for 49 templates containing candidate single nucleotide polymorphisms (SNPs) that were Sanger-validated. High-resolution melt analysis and resequencing of the 190 test fish sometimes detected more that a single polymorphisms in each template

GenBank accessions	Primer ID	Gene	*e*-value	Number of variants	Position of putative SNP	S/N
GQ910745	Oke003	Ubiquinol-cytochrome c reductase iron-sulphur subunit, mitochondrial precursor	8.00E-15	1	116	S
GQ910746	Oke010	Diablo homologue, mitochondrial precursor	3.00E-11	1	244	N
GQ910747	Oke011	No hits		1	79	
GQ910748	Oke017	No hits		5	81, 140, 165, 262, 377	
GQ910749	Oke018	No hits		4	31, 57, 73, 140, 355	
GQ910750	Oke025	Exostoses-like 3	9.00E-16	1	95	S
GQ910751	Oke026	Zinc finger protein 135	1.00E-18	3	33, 72, 87	S, S, S
GQ910752	Oke027	PR domain zinc finger protein 9	3.00E-15	2	70, 79	S, S
GQ910753	Oke028	No hits		7	199, 391, 392, 413, 415, 416, 476	
GQ910754	Oke034	No hits		1	101	
GQ910755	Oke035	No hits		2	70, 137	
GQ910756	Oke037	Transmembrane protein 136	5.00E-11	1	387	S
GQ910757	Oke040	No hits		1	52	
GQ910758	Oke045	No hits		8	83, 199, 217, 226, 312, 330, 352, 445	
GQ910759	Oke054	Brain protein 16	3.00E-21	1	59	S
GQ910760	Oke055	No hits		1	65	
GQ910761	Oke056	No hits		5	127, 154, 189, 251, 304	
GQ910762	Oke057	Unnamed protein product	4.00E-13	1	245	N
GQ910763	Oke058	No hits		1	42	
GQ910764	Oke061	No hits		4	16, 60, 164, 241	
GQ910765	Oke073	No hits		2	56, 95	
GQ910766	Oke077	Mediator of RNA polymerase II transcription subunit 26	3.00E-18	3	32, 35	N, N
GQ910767	Oke079	Similar to cathepsin A	3.00E-10	1	67	S
GQ910768	Oke081	Hypothetical protein	1.00E-10	1	247	S
GQ910769	Oke082	RAB, member of RAS oncogene family-like 2A	7.00E-21	1	84	S
GQ910770	Oke083	Tubulin alpha chain, testis-specific	2.00E-25	3	58, 94, 145	S, S, S
GQ910771	Oke085	No hits		2	81, 255	
GQ910772	Oke091	Clusterin-1	1.00E-24	1	138	N
GQ910773	Oke093	Coiled-coil domain-containing protein 16	8.00E-23	1	77	S
GQ910774	Oke096	No hits		2	154, 232	
GQ910775	Oke099	No hits		1	52	
GQ910776	Oke122	No hits		1	50	
GQ910777	Oke123	No hits		1	218	
GQ910778	Oke128	No hits		3	75, 108, 149	
GQ910779	Oke129	No hits		2	102, 198	
GQ910780	Oke132	No hits		3	114, 126, 139	
GQ910781	Oke136	No hits		2	139, 147	
GQ910782	Oke140	No hits		2	37, 113	
GQ910783	Oke147	No hits		1	135	
GQ910784	Oke150	Bromodomain containing 2	2.00E-15	1	118	S
GQ910785	Oke156	No hits		6	142, 159, 228, 269, 303, 317	
GQ910786	Oke162	No hits		3	89, 119, 121	
GQ910787	Oke166	No hits		1	100	
GQ910788	Oke179	No hits		1	137	

Table 3 *Continued*

GenBank accessions	Primer ID	Gene	*e*-value	Number of variants	Position of putative SNP	S/N
GQ910789	Oke194	Ring finger and SPRY domain containing 1	2.00E-15	1	106	N
GQ910790	Oke213	No hits		1	82	
GQ910791	Oke219	No hits		3	99, 106, 132	
GQ910792	Oke222	No hits		4	101, 125, 151, 157	
GQ910793	Oke230	No hits		1	106	

GenBank accession numbers, primer ID, number of variants (putative SNPs) and position in the sequence are given. Synonymous (S) and nonsynonymous (N) changes are noted in the last column. *e*-value = expectation value: the number of different alignments with scores equivalent to or better than the score of an alignment that are expected to occur in a database search by chance. Only hits with an *e*-value <1.00E-10 are reported.

many aspects of salmonid physiology, ecology and evolution.

The consensus sequences contained nearly 26 000 putative SNPs at a depth of coverage ≥8. This SNP frequency is somewhat smaller than reports using Sanger sequencing of EST libraries. For instance, Smith *et al.* (2005b) sequenced about 89 kb of ESTs and observed a frequency of 4.30×10^{-3} SNPs per base pair; we observed a frequency of 5.4×10^{-4} putative SNPs per base pair in our 454 sequence. One reason for the difference between these estimates might be that our depth of coverage criterion excluded thousands of potential candidates observed in contigs with depth <8. Of course, the observed frequencies are dependent upon the parameters selected for assembly, SNP identification and validation as well as the regions sequenced.

Consensus sequences containing putative SNPs were selected for further validation. Selection was based on factors including presumed gene function to include potential candidate genes (see Table S1, Supporting Information where many of the putative SNPs annotate to structural proteins of interest). However, in some instances, annotation of the final Sanger sequence did not match annotation of respective 454 consensus sequences, and in some instances, a significant BLAST hit was not identified for the final sequenced product (e.g. 'no hit' Table 3). A primary reason for any discrepancy was that 454 pyrosequencing was carried out on cDNA whereas Sanger sequencing was carried out on genomic DNA. Any intron amplified in genomic DNA would differ from cDNA template, and in all instances, only protein databases were queried. In addition, in all instances, the length of the 454 consensus sequence was significantly greater than the genomic fragment that we amplified, increasing the probability of 454 consensus sequences aligning with a given sequence in public databases.

Where possible, SNPs were categorized as synonymous or nonsynonymous (Table 4). Some nonsynonymous SNPs might be expected to provide elevated differences between populations over synonymous SNPs. No clear pattern emerged, although our sample size is small. In our test populations, the largest difference in minor allele frequencies between populations was observed in the nonsynonymous SNP *Oke_U1001-79* (Table 4).

Interestingly, in the initial PCR test, we had substantially different success rates between the two categories of variant: 0.61 for heterozygotes in SUSI and 0.23 for homozygous nucleotide differences between SUSI and DELT. One would expect that some primers would not produce any products given that the primers were designed from cDNA and the templates used were genomic DNA (e.g. an intron could be present between the two primers). However, failures because of intron/exon boundaries should occur at equal frequencies in the two categories.

We often observed multiple PCR products in those classified as primer 'failures.' This result may reflect nonspecific amplification of distinct genes or more likely may be associated with amplification of paralogues of different length remaining in the pseudo-tetraploid salmonids. Modern salmonids originated from a duplication event 25–100 million years ago (Allendorf & Thorgaard 1984). Although largely diploidized, approximately 50% of the duplicated loci created by that event are still detectable by their protein products; many of the nonexpressed duplicates may be retained as pseudogenes (cf., Allendorf & Danzmann 1997). Putative SNPs in salmonids, discovered by sequencing, often fail to validate at a much higher rate than SNPs discovered in other vertebrates. This higher failure rate has been attributed to the discovery of PSVs—the occurrence of different alleles on paralogous genes rather than SNPs on homologues (Smith *et al.* 2005b). We speculate that our selection criterion for the second category (homozygous differences between SUSI and DELT) enriched the identification of

Table 4 Genotype results from three populations: Reidova Bay, Russia; Kitoi Bay, Alaska; and Nitinat River, British Columbia, Canada, including F_{IS} (Weir and Cockerham 1984) and minor allele frequency for each population

		F_{IS}			Minor allele frequency		
Primer pair	SNP name	Russia	Alaska	Canada	Russia	Alaska	Canada
Paired SNPs							
Oke017	*Oke_U1002-165*	−0.05	−0.11	0.17	0.32	0.11	0.24
Oke017	*Oke_U1002-262*	0.01	−0.06	0.08	0.16	0.45	0.43
Oke045	*Oke_U1008-83*	m	−0.01	−0.08	0.00	0.02	0.43
Oke045	*Oke_U1008-199*	0.17	0.04	0.07	0.18	0.13	0.37
Oke056	*Oke_U1010-154*	m	m	m	0.00	0.00	0.00
Oke056	*Oke_U1010-251*	0.18	0.20	0.01	0.06	0.11	0.37
Oke061	*Oke_U1012-60*	0.38***	0.08	−0.08	0.49	0.39	0.25
Oke061	*Oke_U1012-241*	−0.11	−0.05	−0.20	0.42	0.39	0.24
Oke132	*Oke_U1022-114*	−0.04	0.12	−0.08	0.05	0.27	0.08
Oke132	*Oke_U1022-139*	−0.05	0.20	0.02	0.37	0.42	0.10
Single SNP per template							
Oke003	*Oke uqcrfs-69*	−0.06	m	m	0.07	0.00	0.00
Oke010	*Oke_U0602-244*	0.00	0.88	0.13	0.39	0.05	0.37
Oke011	*Oke_U1001-79*	−0.02	−0.02	0.18	0.40	0.03	0.40
Oke025	*Oke_Estl3-95*	−0.20	0.17	−0.74***	0.17	0.41	0.44
Oke034	*Oke_U0603-101*	0.52***	m	0.36**	0.49	0.00	0.23
Oke055	*Oke_brp16-65*	−0.13	0.15	0.22	0.13	0.42	0.11
Oke057	*Oke_UBA3-245*	−0.06	0.16	−0.09	0.07	0.09	0.40
Oke058	*Oke_U1011-42*	−0.12	0.01	0.18	0.35	0.47	0.20
Oke079	*Oke_U1014-67*	−0.01	m	0.26	0.02	0.01	0.04
Oke081	*Oke_PSA4-247*	m	−0.13	−0.06	0.01	0.18	0.06
Oke082	*Oke_rabl2-84*	−0.17	0.38**	0.16	0.46	0.27	0.30
Oke091	*Oke_clu1-138*	m	m	m	0.00	0.00	0.00
Oke093	*Oke_ccd16-77*	0.06	0.10	0.12	0.46	0.11	0.27
Oke099	*Oke_U1017-52*	−0.05	0.00	0.14	0.06	0.48	0.13
Oke122	*Oke_U1018-50*	0.03	−0.02	m	0.27	0.03	0.01
Oke123	*Oke_U1019-218*	m	0.08	m	0.00	0.08	0.00
Oke147	*Oke_U1025-135*	0.18	m	m	0.41	0.00	0.00
Oke150	*Oke_brd2-118*	−0.13	0.13	0.29*	0.18	0.06	0.07
Oke166	*Oke_U1028-100*	−0.06	m	m	0.39	0.00	0.00
Oke179	*Oke_U1029-137*	m	m	m	0.00	0.00	0.00
Oke194	*Oke_RSPRY1-106*	−0.05	m	0.02	0.14	0.01	0.14
Oke213	*Oke_U1030-82*	−0.83***	−0.12	−0.76***	0.47	0.17	0.47
Oke230	*Oke_U1033-108*	m	m	m	0.00	0.00	0.00
Multiple SNPs per template							
Oke018	*Oke_U1003-355*	−0.55***	0.12	−0.57***	0.44	0.29	0.47
Oke026	*Oke_ZNF-87*	−0.79***	−0.33**	−0.61***	0.47	0.30	0.43
Oke027	*Oke_U0604-70*	−0.50***	−0.76***	−0.60***	0.43	0.49	0.41
Oke028	*Oke_U1004-476*	m	m	m	0.00	0.00	0.00
Oke035	*Oke_U1006-137*	−0.03	0.03	−0.07	0.04	0.17	0.48
Oke073	*Oke_U1013-95*	−0.32**	−0.24*	−0.35***	0.44	0.49	0.29
Oke077	*Oke_MED26-35*	m	m	m	0.00	0.00	0.00
Oke083	*Oke_TBAT-94*	−0.39***	−0.13	−0.38***	0.29	0.12	0.28
Oke085	*Oke_U1015-255*	−0.02	0.27*	−0.07	0.13	0.48	0.07
Oke096	*Oke_U1016-154*	−0.04	−0.16	−0.09	0.29	0.14	0.40
Oke128	*Oke_U1020-75*	0.09	−0.03	−0.05	0.16	0.03	0.06
Oke129	*Oke_U1021-102*	−0.11	−0.01	0.07	0.43	0.01	0.35
Oke136	*Oke_U1023-147*	−0.09	−0.05	0.02	0.35	0.13	0.30
Oke140	*Oke_U1024-113*	0.20	−0.01	0.26	0.19	0.01	0.04
Oke156	*Oke_U1026-142*	−0.86	−0.58	−0.80	0.46	0.37	0.44
Oke162	*Oke_U1027-89*	0.17	−0.05	−0.13	0.44	0.05	0.12

Table 4 *Continued*

Primer pair	SNP name	F_{IS}			Minor allele frequency		
		Russia	Alaska	Canada	Russia	Alaska	Canada
Oke219	*Oke_U1031-132*	m	−0.01	0.12	0.01	0.02	0.19
Oke222	*Oke_U1032-125*	−0.08	−0.01	0.04	0.10	0.02	0.23

Five-prime nuclease assays were successfully designed for 51 candidate single nucleotide polymorphisms (SNPs) from 46 Sanger-validated sequences. The first 10 assays were designed for pairs of linked candidate SNPs, the next 23 assays originate from templates that contained only a single polymorphism, and the last 18 assays were developed for a single candidate SNP in remaining sequences that contained multiple polymorphisms. Monomorphic loci are identified by (m).
*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.

PSVs in our transcriptome sequences; fixed differences are more likely to evolve between duplicated sets of loci because of reduced or absent crossing over between the homeologues.

We anticipated PSVs in the individual assemblies for SUSI and DELT. The stringency criterion of 90% sequence identity to allow the potential assembly of highly polymorphic loci would also assemble some paralogues. Stringent assembly criteria are required when the end product is homologous gene sequence in salmonids (see Koop *et al.* 2008); less stringent criteria are satisfactory for SNP discovery given that PSVs can be identified through other validation steps. During HRMA, we identified several templates that contained multiple variants, and genotyping confirmed that some of these contained two to four valid SNPs (Tables 3 and 4). However, our highest failure rate was in this category, and many of the templates containing more than two variants resulted from the assembly of paralogues (Table 4, Fig. 5).

Inserting the HRMA into a SNP development workflow facilitates validation using hundreds of individuals, enabling substantial test sets of populations. Studies using only Sanger sequencing at this stage routinely test only a few tens of individuals (Aitken *et al.* 2004; Smith *et al.* 2005a; Elfstrom *et al.* 2007). By screening larger numbers of individuals, this approach can reveal insights into genotype distribution. The blue curve family of primer pair Oke122, confirmed by Sanger sequencing to be

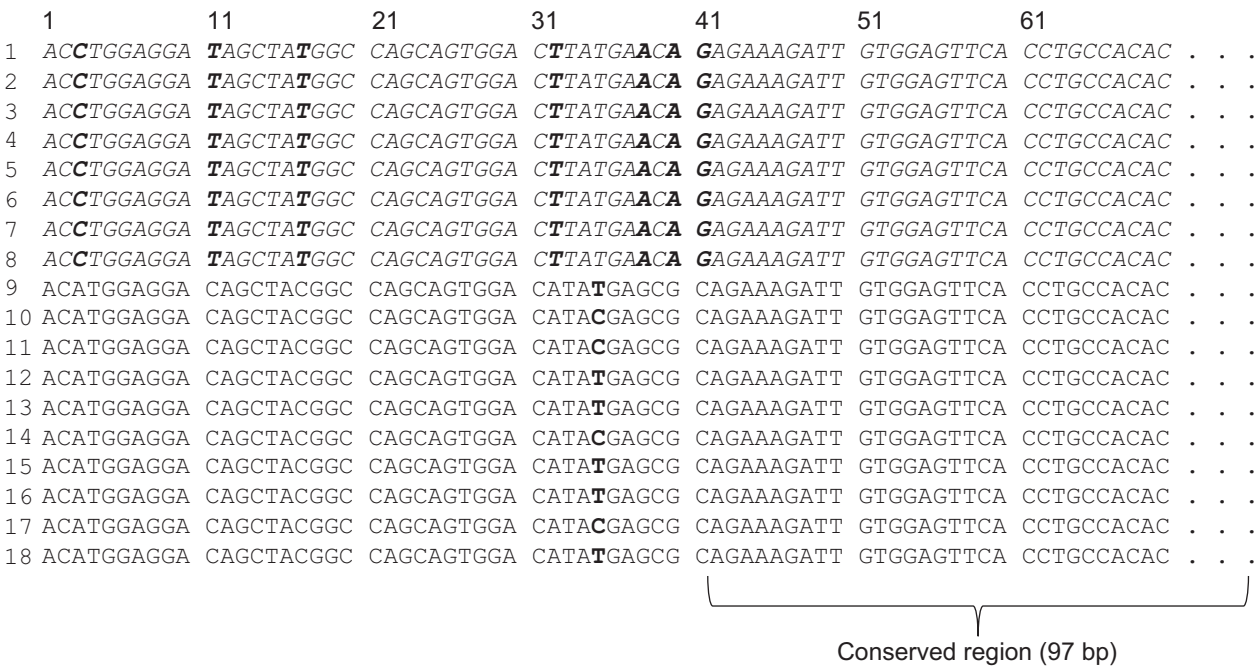

Fig. 5 The sorting of an alignment that assembled paralogues into a contig. A total of eight sequence variants occur at positions 3, 11, 17, 32, 35, 38, 40 and 41. Seven of these invariably co-occur, defining sequences 1–8 and sequences 9–18 as paralogues. A T/C single nucleotide polymorphism occurs in position 35 in the second set of sequences.

CC homozygotes, clearly dominates in North American populations, and the green curve family, TT homozygotes, is only observed in Asian populations (Fig. 4). Genotyping of three test populations for *Oke_U1018-50* confirmed the frequency of the C allele to be 0.73 in Russia, 0.97 in Alaska and 0.99 in Canada (Table 4).

With larger test sets available with HRMA, there are more opportunities to control for or engage the power of ascertainment bias. Broad and uniform coverage of test populations will reduce ascertainment bias which can be useful, especially for the studies of broad-scale questions. Alternatively, one can enrich population coverage in specific geographical regions to increase ascertainment bias and better resolve populations of local importance (Smith & Seeb 2008). While not uniformly distributed, the HRMA test populations in this study do span the entire species range. SNPs may be identified within this set of test individuals that are useful to address questions in Asia, questions in North America or assignment of unknown individuals migrating in the Gulf of Alaska and Bering Sea to continent of origin (Seeb *et al.* 2004).

It is important to note that we advocate HRMA as a polymorphism screen, but at our hands genotype calling using HRMA had two potential weak points (contrast Smith *et al.* 2010). First, detailed curve resolution was challenging and required careful optimization of primer design and concentration, amplicon length, annealing temperature and magnesium chloride concentration. Second, the difference in melt curves for some variants (A/T, C/G) is so subtle that homozygotes might be missed. As used in this project, HRMA enabled a rapid validation step but only a coarse look at genotypic distribution across many populations.

We had some candidate SNPs drop out at the final step of validation. Some of these, scored as monomorphic, may have been low-frequency SNPs that were observed in the HRMA validation but not present in the population samples that were genotyped. The seven assays that demonstrated an excess of heterozygotes were amplifying PSVs.

Interestingly, five polymorphisms conformed to HWE in two populations, but they significantly deviated in the third ($P \leq 0.001$ in two of these). These may be valid SNPs where the deviation signals a mutation in a probe or priming site in one part of the species' range. Alternatively, these may be isoloci where polymorphisms occur at the same position in both homologues and homeologues, and depending upon the populations, these may be either SNPs, PSVs or even both (cf., Allendorf & Thorgaard 1984; Waples 1988). This phenomenon is sometimes detected in telomeric regions where crossing over may take place between homeologues during meiosis in males (residual tetrasomic segregation, (Allendorf & Danzmann 1997)). These numbers are similar to those of Allendorf & Thorgaard (1984) who observed four pairs of isoloci in 33 allozyme loci examined in rainbow trout. We propose to retain these five loci, potentially isolocus pairs, for further examination in population studies.

As expected, the majority of the tests for linkage disequilibrium among the paired candidate SNPs produced highly significant test statistics. There was one interesting exception, the paired SNPs *Oke_U1002-165* and *Oke_U1002-262*, for which the Russian population was in linkage equilibrium. This may be indicative of variable rates of recombination among broadly separated lineages of chum salmon. In tests for linkage disequilibrium between all loci, we detected additional SNPs *(Oke_UBA3-245* and *Oke_U1006-137)* associated with two of our SNP pairs, enlarging putative linkage groups among these newly described SNPs. The study of linked SNPs, especially as evidenced by the *Oke_U1002* pair, increases the statistical power for population structure and conservation studies (Morin *et al.* 2009).

Finally, a limitation of our approach, when the experiment was designed, was that the cost of next-generation sequencing runs prohibited the sequencing of many fish. We considered sequencing cDNA pools from multiple fish, but such approaches often suffer from the loss of individual genotype data. We selected ascertainment fish from two populations where spawning fish were available at the time, roughly in the centre of the species range. Now, as we sequence additional transcriptomes from more individuals and individuals from other areas, we can build upon patterns of individual genotypic variation for selecting candidate SNPs.

The value of this project was not only the 37 high-throughput assays developed here, but it was also in the discovery and validation pipeline. The cost of next-generation sequencing has dropped substantially, and this trend will continue into the future (Martinez & Nelson 2010). The use of tagged libraries now enables the pooling of individuals for the more efficient use of next-generation sequencing runs (Hohenlohe *et al.* 2010). Currently, for the same cost as this project, one can obtain a similar amount of redundant sequence for *de novo* assembly from dozens of fish. Our current strategy includes sequencing many fish that originate from populations of interest across the species' range. Using the PCR/HRMA/Sanger validation will enable us to validate 100s of SNPs for the studies of population structure and mixture analyses as well as for the studies of adaptive genetic variation.

Acknowledgements

We thank the Russian Federal Research Institute of Fisheries and Oceanography and Kamchatka Fisheries and Oceanography

Research Institute for test samples from Russia, Canadian Department of Fisheries and Oceans for samples from British Columbia, and Alaska Department of Fish and Game and Washington Department of Fish and Wildlife for samples from the USA. We also thank Brad Barbazuk for his encouragement and work with assemblies during early phases of this project. This report was partially funded by the Alaska Sustainable Salmon Fund under Study #45919 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce, administered by the Alaska Department of Fish and Game. The statements, findings, conclusions and recommendations are those of the authors and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration, the U.S. Department of Commerce, or the Alaska Department of Fish and Game. Additional funding was derived from a grant from the Gordon and Betty Moore Foundation and by funds from the State of Alaska.

References

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology*, **13**, 1423–1431.

Allendorf FW, Danzmann RG (1997) Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics*, **145**, 1083–1092.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. Turner BJ), pp. 1–53. Plenum Press, New York.

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, **18**, 763–770.

Eccles AJ (2008) RNA extraction (2008, July 29). OpenWetWare. Retrieved 21:22, June 9, 2009 from http://openwetware.org/index.php?title=Eccles:RNA_extraction_AJ&oldid=225407.

Elfstrom CM, Smith CT, Seeb LW (2007) Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. *Molecular Ecology Notes*, **7**, 1211–1215.

Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629–1631.

Everett MV, Grau ED, Seeb JE (2011) Short reads and non-model species: exploring the complexities of next generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, (in press).

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Farley E, Azumaya T, Beamish R *et al.* (2009) *Climate Change, Production Trends, and Carrying Capacity of Pacific Salmon in the Bering Sea and Adjacent Waters. Bulletin Number 5*. North Pacific Anadromous Fish Commission, Vancouver.

Habicht C, Seeb LW, Myers KW, Farley EV, Seeb JE (2010) Summer-fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms (SNPs). *Transactions of the American Fisheries Society*, **139**, 1171–1191.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**, R32; doi: 10.1186/gb-2009-1110-1183-r1132.

Hauser L, Seeb JE (2008) Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries*, **9**, 473–486.

Hemmer-Hansen J, Nielsen E, Meldrup D, Mittelholzer C (2011) Identification of single nucleotide polymporhisms in candidate genes for growth and reproduction in a non-model organism, the Atlantic cod, *Gadus morhua*. *Molecular Ecology Resources*, (in press).

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics*, **6**, e1000862.

Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5′ in place of 3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 7276–7280.

Hu Z-L, Bao J, Reecy JM (2008) A web-based program to batch analyze gene ontology classification categories. *Online Journal of Bioinformatics.*, **9**, 108–112.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Karlsson S, Moen T, Lien S, Glover K, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP chip. *Molecular Ecology Resources*, (in press).

Kasahara M, Naruse K, Sasaki S *et al.* (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447**, 714–719.

Koop BF, von Schalburg KR, Leong J *et al.* (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics*, **9**, 545.

Martinez DA, Nelson MA (2010) The Next Generation Becomes the Now Generation. *Plos Genetics*, **6**, e1000906.

McGlauflin MT, Smith MJ, Wang JT *et al.* (2010) High-resolution melting analysis for the discovery of novel single-nucleotide polymorphisms in rainbow and cutthroat trout for species identification. *Transactions of the American Fisheries Society*, **139**, 676–684.

Miller KM, Withler RE (1996) Sequence analysis of a polymorphic MHC class II gene in Pacific salmon. *Immunogenetics*, **43**, 337–351.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Narum SR, Banks M, Beacham TD *et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, **17**, 3464–3477.

Nielsen R, Bustamante C, Clark AG *et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *Plos Biology*, **3**, 976–985.

Olsen M, Volny V, Berube M *et al.* (2011) A simple route to single-nucleotide polymorphisms in a non-model species: identification and characterization of SNPs in the ringed seal (*Pusa hispida hispida*). *Molecular Ecology Resources*, (in press).

Primmer CR (2009) From conservation genetics to conservation genomics. *Year in Ecology and Conservation Biology 2009*, **1162**, 357–368.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). *Molecular Ecology*, **19**, 115–131.

Rousset F (2008) GENEPOP ′ 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Rozen S, Skaletshy HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds Krawetz S & Misener S), pp. 365–386. Humana Press, Totowa, NJ.

Sagarin R, Carlsson J, Duval M *et al.* (2009) Bringing molecular tools into environmental resource management: untangling the molecules to policy pathway. *Plos Biology*, **7**, 426–430.

Sanchez CC, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Seeb LW, Crane PA, Kondzela CM *et al.* (2004) Migration of Pacific Rim chum salmon on the high seas: insights from genetic data. *Environmental Biology of Fishes*, **69**, 21–36.

Seeb LW, Antonovich A, Banks MA *et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries*, **32**, 540–552.

Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009) SNP genotyping by the 5′-nuclease reaction: advances in high throughput genotyping with non-model organisms. In: *Methods in Molecular Biology, Single Nucleotide Polymorphisms*, 2nd edn. (ed. Komar A), pp. 277–292. Humana Press, New York.

Seeb LW, Templin WD, Sato S *et al.* (2011) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources*, (in press).

Slate J, Gratten J, Beraldi D *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97–107.

Smith CT, Seeb LW (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.

Smith CT, Baker J, Park L *et al.* (2005a) Characterization of 13 single nucleotide polymorphism markers for chum salmon. *Molecular Ecology Notes*, **5**, 259–262.

Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005b) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology*, **14**, 4193–4203.

Smith BL, Lu CP, Bremer JRA (2010) High-resolution melting analysis (HRMA): a highly sensitive inexpensive genotyping alternative for population studies. *Molecular Ecology Resources*, **10**, 193–196.

Stephenson J, Campbell M, Hess J *et al.* (2009) A centralized model for creating shared, standardized, microsatellite data that simplifies interlaboratory collaboration. *Conservation Genetics*, **10**, 1145–1149.

Storfer A, Eastman JM, Spear SF (2009) Modern molecular methods for amphibian conservation. *BioScience*, **59**, 559–571.

Vasemagi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Waples RS (1988) Estimation of allele frequencies at isoloci. *Genetics*, **118**, 371–384.

Waples RS, Punt AE, Cope JM (2008) Integrating genetic data into management of marine resources: how can we do it better? *Fish and Fisheries*, **9**, 423–449.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wu SB, Wirthensohn M, Hunt P, Gibson J, Sedgley M (2008) High resolution melting analysis of almond SNPs derived from ESTs. *Theoretical and Applied Genetics*, **118**, 1–14.

Zhulidov PA, Bogdanova EA, Shcheglov AS *et al.* (2004) Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. *Nucleic Acids Research*, **32**, e37.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Forward and reverse sequence for all primer pairs used to validate putative SNPs in chum salmon.

Table S2 Five-prime nuclease assays for 51 candidate SNPs in chum salmon.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.