



Testing advances in molecular discrimination among Chinook salmon life histories: evidence from a blind test

Michael A. Banks*, David P. Jacobson*, Isabelle Meusnier[†], Carolyn A. Greig[‡],
Vanessa K. Rashbrook[§], William R. Ardren[¶], Christian T. Smith^{**}, Jeremiah Bernier-Latmani^{††},
John Van Sickle^{‡‡} and Kathleen G. O'Malley*

*Coastal Oregon Marine Experiment Station, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, Newport, OR, 97365, USA. [†]Center for Biology and Management of Populations, Montpellier, 34000, France. [‡]Department of Biosciences, College of Science, Swansea University, Swansea, Wales, SA1 1LG, UK. [§]UC Davis Genome Center – DNA Technologies Core, UC Davis, One Shields Ave, Davis, CA, 95616, USA. [¶]USFWS, 11 Lincoln St., Essex Junction, VT, 05452, USA. ^{**}Abernathy Fish Technology Center, USFWS, Longview, WA, 98632, USA. ^{††}Department of Oncology, CHUV and University of Lausanne, Lausanne, 1015, Switzerland. ^{‡‡}Department of Fisheries and Wildlife, Oregon State University, Corvallis, OR, 97331, USA.

Summary

The application of DNA-based markers toward the task of discriminating among alternate salmon runs has evolved in accordance with ongoing genomic developments and increasingly has enabled resolution of which genetic markers associate with important life-history differences. Accurate and efficient identification of the most likely origin for salmon encountered during ocean fisheries, or at salvage from fresh water diversion and monitoring facilities, has far-reaching consequences for improving measures for management, restoration and conservation. Near-real-time provision of high-resolution identity information enables prompt response to changes in encounter rates. We thus continue to develop new tools to provide the greatest statistical power for run identification. As a proof of concept for genetic identification improvements, we conducted simulation and blind tests for 623 known-origin Chinook salmon (*Oncorhynchus tshawytscha*) to compare and contrast the accuracy of different population sampling baselines and microsatellite loci panels. This test included 35 microsatellite loci (1266 alleles), some known to be associated with specific coding regions of functional significance, such as the circadian rhythm cryptochrome genes, and others not known to be associated with any functional importance. The identification of fall run with unprecedented accuracy was demonstrated. Overall, the top performing panel and baseline (HMSC21) were predicted to have a success rate of 98%, but the blind-test success rate was 84%. Findings for bias or non-bias are discussed to target primary areas for further research and resolution.

Keywords individual-identification, microsatellites, *Oncorhynchus tshawytscha*

Introduction

Salmon are prized globally as a source of high-quality food. Chinook or King salmon (*Oncorhynchus tshawytscha*) traditionally has ranked as the most favored salmon species owing to its firm quality and high-nutrient flesh. Indeed, Chinook salmon was ranked among the top five of 60 wildlife species in an economic valuation of biodiversity

(along with elk, moose, humpback whale and bald eagle; Martin-Lopez *et al.* 2008). The natural distribution of Chinook extends from Hokkaido Island (Northern Japan) up northerly through Kamchatka, Russia, the Bering Sea, Alaska, to ocean territories west of Canada, Washington, Oregon and California. Today, this species also is spawned and reared in a substantial number of hatcheries distributed across this range and in aquaculture enterprises of Chile, Brazil, Korea and New Zealand, where some naturalized populations have become established.

At the southeastern extreme of Chinook's natural distribution, California's Central Valley drainage surfaces as a unique context for this species. Broad availability of extensive habitat combined with consistent cold watering from Sierra snowmelt here has supported development of

Address for correspondence

M. A. Banks, Coastal Oregon Marine Experiment Station, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, 2030 SE Marine Science Drive, Newport, OR, 97365, USA.
E-mail: michael.banks@oregonstate.edu

Accepted for publication 15 January 2014

the most diverse range in life-history types found anywhere. Thus, there are four primary runs, named fall, late-fall, winter and spring, after seasonal peaks in numbers of freshwater returns from the ocean (Fisher 1994). Although there is overlap across seasons and essentially gravid Chinook may be found in the river year round, historically the runs occupied spatially segregated spawning habitats. Winter run utilized spring-fed headwaters, spring run utilized higher elevation streams, late-fall run utilized mainstem rivers and fall run utilized lower elevation rivers and tributaries (Yoshiyama *et al.* 2001). Today, however, approximately 70% of previously available habitats are now impounded by reservoirs or for other uses, raising questions as to how effectively these runs may be able to maintain reproductively isolated breeding groups.

These four runs also often occur together during other phases of the Chinook's life cycle, for example as juvenile out-migrants through the Sacramento/San Joaquin Delta and San Francisco estuary or during ocean-feeding migration. As migrants through the Delta, juvenile Chinook are exposed to large water export facilities operated by the State of California (State Water Project) and the U.S. Government (Central Valley Project). Some of these salmon subpopulations are listed as endangered (winter run) or threatened (spring run), thus there has been active interest to develop reliable methods for identification of run among sampled fish. This motivated early development of molecular and statistical tools for individual assignment, and Central Valley Chinook salmon were among the first salmonids to be individually assigned to run using molecular genetics (Banks *et al.* 1999, 2000). It now has been over a decade since that baseline was published, and a central goal of our effort has been to develop and upgrade methodologies in order to provide the highest resolution for individual (not population)-based discrimination among these four runs of Central Valley Chinook salmon. Two primary approaches were addressed: (i) We sought markers directly linked to life-history traits differing among the runs (such as run timing; O'Malley *et al.* 2007) and (ii) we employed statistical approaches to assess the relative power of alternate markers for run discrimination (Banks *et al.* 2003). Research presented here focused on the improvements of molecular genetics to discriminate among Chinook salmon of California's Central Valley. Three different microsatellite loci panels were contrasted between two different baseline collections of Chinook salmon.

Methods

Baselines, subpopulation assemblages, sample collection and DNA extraction

This study compared and contrasted two baseline population genetic characterizations of Chinook salmon sampled from California's Central Valley drainage (Fig. 1), hereafter called baselines, and three different microsatellite loci panels. The

first baseline collection, the Hatfield Marine Science Center (HMSC) baseline, founded on Banks *et al.* (2000), included samples that were divided among five reporting groups. Three of the reporting groups corresponded to primary runs (winter, fall and late-fall), and the other two corresponded to genetically distinct assemblages of spring run: (i) spring run from Butte Creek and (ii) spring run from Deer and Mill Creeks. These samples were assembled among ten 96-well trays (two for each primary run or reporting group) and included a total of 936 samples: comprising between six and 86 samples for each of nine years and 24 run collections taken from 1991 to 1998 by the California Department of Fish and Game (CDFG) and the U.S. Fish and Wildlife Service (Table 1). The second baseline collection, the Genetic Analysis of Pacific Salmon (GAPS) Consortium baseline, was developed and standardized among 12 fisheries genetics laboratories in the Pacific Northwest (Seeb *et al.* 2007; Moran *et al.* 2013) and included a total of nine discrete population samples from California's Central Valley drainage among a total of 166 population samples distributed from California to Alaska. These baseline collections were divided among four reporting groups (the five described in Banks *et al.* 2000 and depicted in Table 1, except late-fall). To compare assignment accuracy of these baselines, it was necessary to use common reporting groups. Because the GAPS baseline did not characterize any late-fall collections from California, fall and late-fall results derived using the HMSC baseline in the present study were pooled into a single fall-late-fall reporting group. This pooled fall-late-fall reporting group derived from GAPS and HMSC baselines also included assignments to both spring and fall individuals from the Feather River Hatchery owing to known hybridization between these stocks and difficulty in resolving population identity between them (Banks *et al.* 2000; Hedgecock *et al.* 2001).

Although 100%, jackknife and leave-one-out simulations available in population assignment applications may be useful for predicting the accuracy and precision provided by various genetic baselines, they also may provide biased or overly optimistic indications. It is thus ideal to include samples of known origin or 'blind samples' when evaluating assignment power. For this purpose, a total of 750 tissue samples from Chinook salmon of known life history stored in the CDFG tissue archive were coded (to mask their identity) and enabled a blind test of assignment accuracy of three alternate microsatellite panels. DNA extraction of blind-test samples followed a silica-based method utilizing multichannel pipettes; PALL glass fiber filtration plates; and buffer, centrifuge and transfer protocols described in Ivanova *et al.* (2006).

Microsatellite loci characterization

Baseline and blind-test samples were characterized utilizing three microsatellite panels, and following amplification protocols detailed in references cited:

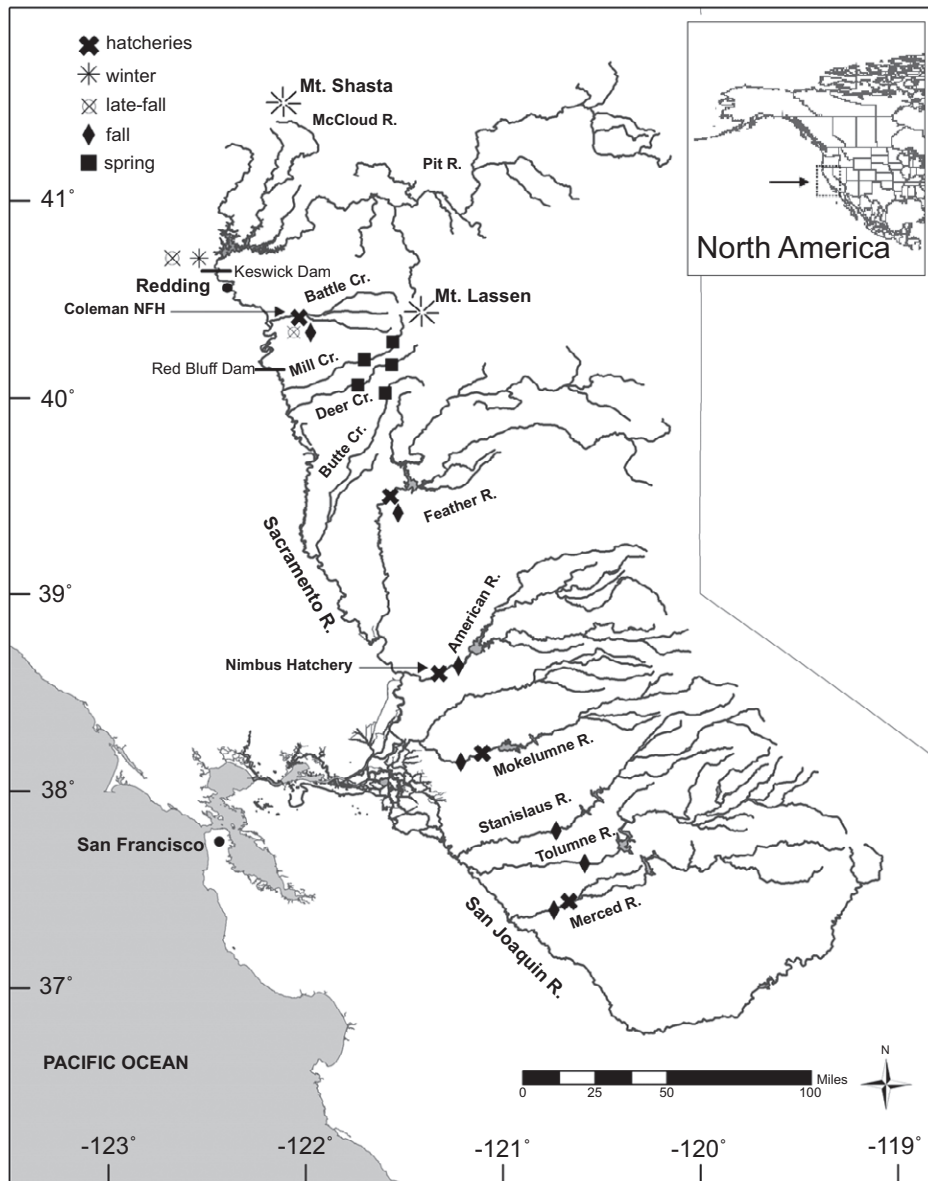


Figure 1 Rivers and tributaries of California's Central Valley indicating Chinook salmon sampling sites per run and hatcheries.

- 1 GAPS13 (from Seeb *et al.* 2007) included: *Ogo-2*, *-4* (Olsen *et al.* 1998); *Oki100* (Canadian Department of Fisheries and Oceans, unpublished); *Omm1080* (Rexroad *et al.* 2001); *Ots-3M* (Greig & Banks 1999); *Ots-9* (Banks *et al.* 1999); *Ots-201b*, *-208b*, *-211*, *-212*, *-213* (Greig *et al.* 2003); *OtsG474* Williamson *et al.* (2002); and *Ssa408* Cairney *et al.* 2000
- 2 HMSC16 (from Banks & Jacobson 2004) included: *Ots-104*, *-107* (Nelson & Beacham 1999); *Ots-201b*, *-208b*, *-209*, *-211*, *-212*, *-215* (Greig *et al.* 2003); *Ots-G78b*, *-G83b*, *-G249*, *-G253*, *-G311*, *-G422*, *-G409* Williamson *et al.* (2002); and *Ost515* (Naish & Park 2002).
- 3 HMSC21 included: the above 16 loci as well as an additional five microsatellites derived from research

characterizing alternate copies of the circadian rhythm transcription factor cryptochrome: *Cry2b.1*, *Cry2b.2*, *Cry3* (O'Malley *et al.* 2010), *Ots-701* (GenBank Accession no. KF163438) and *Ots-702* (GenBank Accession no. KF163440).

Alternate alleles were resolved through electrophoresis utilizing an Applied Biosystems (ABI) 3730xl DNA analyzer and scored using ABI GENEMAPPER software (Version 4).

Standardization of the HMSC baseline with the Abernathy Fish Technology Center

The same standardization methods developed by the GAPS group (Seeb *et al.* 2007) were employed to standardize

Table 1 Collection data for California's Central Valley Chinook baseline populations from breeding stocks separated by run timing and location. Hatfield Marine Science Center (HMSC) baselines are characterized at 16 and 21 microsatellite loci respectively; GAPS13 (from Genetic Analysis of Pacific Salmon Consortium) is a different baseline collection characterized at 13 microsatellite loci.

Run	HMSC16 and HMSC21 baselines				GAPS13 baseline			
	Year	Sampling location	Life stage	<i>n</i>	Year	Sampling location	Life stage	<i>n</i>
Winter	1991	Keswick & Red Bluff Dams	Adult	17	1992–5	Keswick & Red Bluff Dams	Adult	56
	1992	Keswick Dam	Adult	29	1997	Keswick Dam	Adult	3
	1993	Keswick & Red Bluff Dams	Adult	9	1998	Keswick Dam	Adult	17
	1994	Keswick Dam	Adult	24	2001	Keswick Dam	Adult	35
	1995	Keswick Dam	Adult	25	2003	Keswick Dam	Adult	10
	1998	Keswick Dam	Adult	87	2004	Keswick Dam	Adult	15
	Total			191				136
Spring	1994	Butte Creek	Spawned carcass	50	2002	Butte Creek	Adult	61
Butte Creek	1996	Butte Creek	Spawned carcass	12	2003	Butte Creek	Adult	83
	1997	Butte Creek	Spawned carcass	60				
	1998	Butte Creek	Spawned carcass	62				
Total			184				144	
Spring	1994	Deer Creek	Juvenile	12	2002	Deer Creek	Adult	53
Deer & Mill Creek	1995	Deer Creek	Spawned carcass	13	2002	Mill Creek	Adult	71
	1995	Mill Creek	Spawned carcass	10	2003	Mill Creek	Adult	20
Creek	1996	Deer Creek	Juvenile	68				
	1996	Mill Creek	Juvenile	12				
	1997	Deer Creek	Spawned carcass	38				
	1998	Deer Creek	Spawned carcass	26				
	1998	Mill Creek	Spawned carcass	6				
	Total			185				144
	Fall	1995	Nimbus Hatchery	Adult	75	2002	Battle Creek	Adult
	1995	Mokelumne Hatchery	Adult	67	2003	Battle Creek	Adult	77
	1995	Merced Hatchery	Adult	48	2003	Feather Hatchery	Adult	144
					2002	Stanislaus River	Adult	76
					2002	Tuolumne River	Adult	68
Total				190				432
Late-fall	1993	Keswick Dam & Battle Creek	Adult	72		Not sampled		
	1995	Coleman National Fish Hatchery	Adult	90				
	1995	Keswick Dam	Adult	24				
	Total			186				

amplification, electrophoresis, allele nomenclature and scoring methods achieved between HMSC and the Abernathy Fish Technology Center (AFTC) laboratories. Briefly, this exercise involved sharing and evaluating three independent and coded 96-well plates containing Chinook salmon DNA samples:

- 1 Bin-definition plate 1 was passed from HMSC to AFTC along with genotype data. AFTC amplified and analyzed these samples in their laboratory using an ABI 3130 DNA Sequencer to enable AFTC allele bin calibration and scoring with HMSC allele nomenclature.
- 2 Test plate 1/bin-definition plate 2 was passed from HMSC to AFTC but without any genotype data. AFTC analyzed these samples and reported results back HMSC to assess standardization.
- 3 Test plate 2/bin-definition plate 3 was passed from HMSC to AFSC without genotype data. AFTC analyzed these samples and reported results to HMSC for final assessment of standardization among laboratories.

Assignment and statistical analysis

Given that numbers of fall and late-fall migrants substantially exceed those from winter and spring runs in most scenarios in the lower reaches of the Sacramento River or the NW Pacific Ocean, simulations performed to test for precision and accuracy were designed to approximate these relative abundance differences. This was achieved through utilizing the 'realistic fishery' option within the statistical package ONCOR (Kalinowski 2008; www.montana.edu/kalinowski/Software/ONCOR.htm). Note that this technique utilizes a cross-validation over a gene copies method demonstrated to be less prone to providing over-optimistic estimates of assignment power than earlier methods (Anderson *et al.* 2008; Anderson 2010). For HMSC baselines, parameters were set to construct 1000 hypothetical mixtures of size 100 individuals each, using a 0.97 fraction for fall–late-fall reporting group and a 0.01 fraction each for the winter and spring from Butte Creek and the spring from Deer and Mill Creeks reporting groups. For the GAPS13

baseline, parameters were set to construct 1000 hypothetical mixtures of size 100 individuals each, using a 0.2475 fraction for Battle Creek fall, 0.2375 for Butte Creek fall, 0.2375 for Feather River Hatchery fall and 0.2375 for Stanislaus River fall. The GAPS13 simulation therefore had the same total 0.97 fraction for the fall-run reporting group, 0.01 for the Butte Creek spring, 0.01 for the Deer Creek spring, 0.00 for the Feather River Hatchery spring and 0.01 for the winter reporting groups. Complete multilocus data for blind-test samples were required with the exception of up to a maximum of three missing loci for all three microsatellite panels. Run identities were assessed utilizing *ONCOR*'s 'assign individual to baseline population' option, and each individual was assigned to the reporting group for which it had the greatest probability (no probability cutoff was applied). Lower and upper 95% confidence intervals for realistic results from simulation studies were calculated using standard methods ($P \pm 1.96 * \text{standard error}$; Sokal & Rohlf 1995). We cross-tabulated the counts of the 750 blind-test samples correctly (true) versus incorrectly (false) identified by each possible pair of panels, separately for each run. Because both panels of each pair were identifying the same set of samples, their correct identification proportions were not independent. Thus, we used an exact version of McNemar's test (Agresti 2002; Zar 2010) for each pair of panels to test for the equality of those proportions.

Results

Standardization results indicate the AFTC and the HMSC allele scores averaged 97% identical for test plate one and 98% correct for test plate two (Table 2). One locus, *Ots-208b*, consistently scored less than the 90% identity

Table 2 Percentage agreement in allele scoring between Abernathy Fish Technology Center and Hatfield Marine Science Center (HMSC) for microsatellite panel HMSC16.

Locus	Test plate 1	Test plate 2
<i>Ots-104</i>	95.9	99.4
<i>Ots-107</i>	100	98.8
<i>Ots-201b</i>	98.8	99.4
<i>Ots-208b</i>	88.3	87.7
<i>Ots-209</i>	97.7	97.1
<i>Ost-211</i>	96	100
<i>Ots-212</i>	99.4	98.9
<i>Ots-215</i>	100	100
<i>Ots-249</i>	99.4	97.8
<i>Ots-253b</i>	92.5	98.9
<i>Ots-515</i>	92.3	94.8
<i>Ots-G311</i>	99.2	99.3
<i>Ots-G409</i>	94.9	99.4
<i>Ost-G422</i>	100	100
<i>Ost-G78B</i>	94.4	100
<i>Ots-G83B</i>	100	99.4
Average	96.8	98.2

threshold identified by the GAPS Consortium (Seeb *et al.* 2007). Concordance between laboratories for the remaining loci was at least 90%, indicating that these loci had been successfully standardized.

Realistic fishery simulation results indicated strong correct identity assignment potential (largely in the 90th percentiles) for each of the three microsatellite panels (Table 3 and Fig. 2). Consistent ranking among the three panels also was apparent from simulation results with correct assignment parameters ranging from 70 through 100% (GAPS13), 90% through 100% (HMSC16) and 96 through 100% (HMSC21). Non-overlapping 95% confidence intervals reinforce findings that (i) spring from Butte Creek correct assignments was higher for HMSC16 and HMSC21 compared with GAPS13; (ii) spring from Deer and Mill Creeks assignments increased according to ranking for GAPS13, HMSC16 and HMSC21;

Table 3 Summary percentage correct results of realistic fishery simulations assessed at each of the three baselines for populations: W, winter; SB, spring from Butte Creek; SDM, spring from Deer and Mill Creeks; F-LF, fall and late-fall.

	GAPS	HMSC16	HMSC21
W	100	100	100
SB	87.2 (83.6, 90.9)	98.4 (97.1, 99.8)	99.1 (98.1, 100.1)
SMD	69.7 (66.3, 73.2)	89.9 (86.6, 93.3)	95.8 (93.5, 98.0)
F-LF	99.2 (99.1, 99.3)	97.9 (97.8, 98.1)	99.2 (99.1, 99.3)
Ave	89	96.6	98.5

GAPS, Genetic Analysis of Pacific Salmon Consortium; HMSC, Hatfield Marine Science Center.

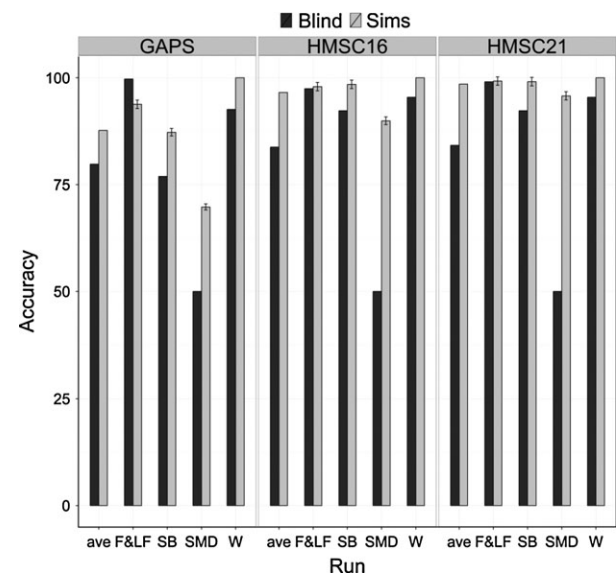


Figure 2 Blind-test ($n = 623$) and simulation correct assignment results ($n = 1000$ for winter and spring reporting groups) among California Central Valley Chinook salmon calculated using *ONCOR* (Kalinowski 2008) and assessed using three different microsatellite panels. Bars on simulations indicate 95% confidence intervals. Chinook salmon runs are indicated as follows: F&LF, pooled fall and late-fall runs; SB, spring from Butte Creek; SMD spring from Mill and Deer Creeks; W, winter.

Table 4 Summary results of percentage correct assignment for each baseline from blind-test samples (Blind) and simulations (Sims) for populations: W, winter; SB, spring from Butte Creek; SDM, spring from Deer and Mill Creeks; F-LF, fall and late-fall.

	GAPS		HMSC16		HMSC21	
	Blind	Sims	Blind	Sims	Blind	Sims
W	92.61	100.0	95.45	100.00	95.45	100.00
SB	76.92	87.24	92.31	98.46	92.31	99.09
SMD	50.00	69.75	50.00	89.92	50.00	95.76
F-LF	99.72	93.80	97.45	97.94	99.07	99.24
Ave	79.81	87.70	83.80	96.58	84.21	98.52

GAPS, Genetic Analysis of Pacific Salmon Consortium; HMSC, Hatfield Marine Science Center.

and (iii) HMSC16 ranked lower than did GAPS13 and HMSC21 for pooled fall and late-fall assignments. Finally, all run assignment averages for both HMSC16 and HMSC21 were higher than for GAPS13.

Blind test of actual power (inferred from 623 known ID samples) indicated that simulation results generally were upwardly biased but affirmed parallel relative rankings across runs and microsatellite panels (Fig. 2). Fewer of winter run, spring from Butte Creek and spring from Deer and Mill Creeks assignments were correct than predicted. Fall-run blind-test assignments matched simulation estimates most closely.

Average realistic fishery simulation rankings of microsatellite panels, HMSC21 best score of 98.5%, HMSC16 next best score of 96.6% and GAPS13 lowest score of 87.7%, were supported by blind-test assignment accuracy of 84.2% (HMSC21), 83.8% (HMSC16) and 79.8% (GAPS13) (Table 4). There is some evidence that HMSC16 and HMSC21 winter blind-test assignments were more often correct than were those of GAPS13 (McNemar's test,

$P = 0.0625$; Table 5). However, we found no differences in the classification success rates of the three panels for any of the other runs (spring from Butte Creek, fall and spring from Deer and Mill Creeks). In particular, HMSC16 and HMSC21 had identical classification success for all blind-test fish except those in the fall run (Table 5). Allele frequency data utilized in this study are available at OSU Scholars Archive (doi: 10.7267/N9KW5CXX).

Discussion

Noting that this study focused on discrimination among closely related Chinook salmon runs from the same primary watershed (that have lost 70% of their historic habitat for spatial segregation), a 98% overall correct assignment prediction from simulations and blind-test affirmation at 84% correct is astonishing. Similarly, promising overall results have been obtained for Sockeye salmon (Beacham *et al.* 2005), cod (Glover *et al.* 2010), cow (Van de Goor *et al.* 2011), sheep (Niu *et al.* 2011) and cats (Kurusima *et al.* 2012). Indeed, HMSC21 blind-test correct assignment averages of 99% (fall), 95% (winter) and 92% (spring from Butte Creek) are especially encouraging given the importance of accurate identification for endangered winter and threatened spring run life histories (NMFS 2009). These particular blind-test results were in close agreement with predictions for simulations [fall: 99% (blind) and 99% (simulations); winter: 95% (blind) and 100% (simulations); spring from Butte Creek: 92% (blind) and 99% (simulations)] (Table 6). This general agreement also is very positive because previous simulation methods have suffered from upward bias in their assessment of most likely assignment power (Anderson 2010).

The wide difference between simulation prediction (96%) and blind-test findings for spring run from Deer and Mill

Table 5 Comparisons of microsatellite panels in their classification success for three true runs. T denotes an accurately classified fish, and F denotes an error. P -values are for McNemar's test of equality in the proportions accurately classified by two panels. Spring run from Deer and Mill Creeks not shown because all three panels had identical classification success.

	True run winter ($n = 176$)			True run spring from Butte Creek ($n = 13$)			True run fall ($n = 432$)			
	H16-F	H16-T	P	H16-F	H16-T	P	H16-F	H16-T	P	
G13-F	8	5	0.0625	G13-F	1	2	G13-F	1	1	
G13-T	0	163		G13-T	0	10		0.5	G13-T	4
	H21-F	H21-T	P	H21-F	H21-T	P	H21-F	H21-T	P	
G13-F	8	5	0.0625	G13-F	1	2	G13-F	1	1	
G13-T	0	163		G13-T	0	10		0.5	G13-T	5
	H16-F	H16-T	P	H16-F	H16-T	P	H16-F	H16-T	P	
H21-F	8	0	1	H21-F	1	0	H21-F	4	1	
H21-T	0	168		H21-T	0	12		1	H21-T	2

G13, Genetic Analysis of Pacific Salmon Consortium panel; H16, Hatfield Marine Science Center 16 microsatellite panel; H21, Hatfield Marine Science Center 21 microsatellite panel.

Table 6 Blind-test result for 623 Chinook salmon. Rows indicate actual known identity; columns indicate where they were assigned by three microsatellite panels: G, GAPS (Genetic Analysis of Pacific Salmon Consortium) or H, HMSC (Hatfield Marine Science Center).

Run	Winter (W)			Spring from Butte Creek (SB)			Spring from Deer & Mill Creeks (SDM)			Fall (F)			Total Actual
	G13	H16	H21	G13	H16	H21	G13	H16	H21	G13	H16	H21	
W	163	168	168	2	1	1	0	1	1	11	6	6	176
SB	0	0	0	10	12	12	1	0	0	2	1	1	13
SDM	0	0	0	0	0	0	1	1	1	1	1	1	2
F-LF	1	1	1	1	2	1	0	2	4	430	427	426	432
													623

W, winter; SB, spring from Butte Creek; SDM, spring from Deer and Mill Creeks; F-LF, fall and late-fall.

Creeks (50%) for all three baselines, however, indicates that this upward bias for simulation methods has not been completely eradicated. There are only two samples of known spring Deer and/or Mill Creeks origin among the 623 samples considered in the blind test. This small sample size tempts one to suggest that observed upward difference between simulation and blind-test findings likely results from chance. We suggest, however, that tests with similarly small sample size scenarios are appropriate because threatened and endangered species by definition are always scarce. Identification applications commonly occur in contexts where endangered species are markedly outnumbered by their more abundant counterparts (such as large-number fall and late-fall Chinook salmon runs in the current case). Although the cross-validation methods introduced by Anderson *et al.* (2008) and 'realistic fishery' algorithms available in ONCOR (Kalinowski 2008) have begun to overcome the upward bias problem, results obtained here for spring run from Mill and Deer Creeks demonstrate that shortfalls still exist in our ability to employ simulation methods to accurately predict most likely assignment power among closely related runs. An earlier iteration of data for this blind test had a total $n = 532$. These 532 known-identity fish, however, happened to contain only one sample from Deer and Mill Creeks and 12 samples from Butte Creek spring runs, yet the three baselines correctly assigned all 13 of these spring samples to their known origin, except that GAPS(13) misassigned two of the 12 springs from Butte Creek. Thus, 100% [and 83% for Butte Creek (GAPS13)] correct blind-test results for both spring run subpopulations were in closer agreement with simulation predictions and did not show any upward bias. Given that both spring run subpopulations had few numbers of samples employed in the first blind-test 532 samples that were low, we returned to the original 750 blind-test sample to derive more data. This increased our total number (n) to 623, but did not substantially increase the numbers of spring run in the blind test. These results underscore the importance of using data that are separate from those used to train a classification process in evaluating the accuracy of that process (Anderson 2010).

No samples from any late-fall run were included in the GAPS13 baseline; however, blind-test and simulation results for late-fall run in the HMSC baselines provided further information with regard to bias. The blind sample of 623 had a total of 77 samples from late-fall run (data not shown). Simulation tests predicted a 91% success rate for late-fall, yet the blind-test score was only 44% correct. This was not unexpected considering that fall and late-fall runs are the most closely related among all Central Valley population pairs (fall–late-fall pairwise $F_{st} = 0.02$ vs. average F_{st} for all subpopulations = 0.08). Indeed, late-fall-run misassignments were largely to fall run. Note, however, that an $n = 77$ for late-fall samples is no longer small, yet this run had the highest upward bias observed between simulation and blind-test results. In contrast, this upward bias of simulation prediction was not observed for fall run. Considering fall and late-fall runs separately, the $n = 623$ blind test had 157 fall-run samples, of which 153 (97%) were correctly identified by HMSC21 in exact agreement with simulation prediction of 97%.

Comparing results attained from different microsatellite panels, the overall increasing correct assignment ranking from GAPS13, HMSC16 to HMSC21 was in parallel with increasing number of loci, as observed in other studies (Bjørnstad & Røed 2002; Bamshad *et al.* 2003; Tadano *et al.* 2008). This is supported by consistent ranking results from simulation tests for each of the runs (except GAPS13, which switched to second place for combined fall–late-fall simulation assignments) and marginal McNemar support for the same blind-test 13-16-21 loci increasing assignment ranking. However, despite consistent top performance for HMSC21, margins separating results were not sufficient to prove this statistically. Although HMSC16 and 21 panel performances are largely the same for the blind test, simulations indicate the increased value of additional loci for discrimination among fall and spring runs (Fig. 2). This and fall–late-fall discrimination remain areas of greatest challenge in addressing accuracy for individual-based population assignment among California's Central Valley Chinook salmon. However, fall-run identification across all baselines and microsatellite panels (including both blind-test and simulation results) was high (average 98% correct).

This level of success is a first and likely has strong application potential. Regionally, California's Central Valley Chinook salmon returns have been disturbingly low in recent years. Precipitously low numbers of Central Valley fall-run Chinook salmon was the primary driving force for a complete ocean fishery closure for 2008 and 2009 (NMFS 2009). This situation had significant negative economic consequences for the region and motivates continued efforts, such as the molecular and statistical methods covered here, to better quantify accuracy for individual-based population identity determination for improved management, monitoring and conservation.

Acknowledgments

We are grateful to: Pat Brandes who provided valuable comments on an earlier draft manuscript; Renee Bellinger and Nick Sard for help with figures; the California Department of Fish and Game and the US Fish and Wildlife Service for samples provided; and Environmental Services, California Department of Water Resources and California Bay-Delta Authority (now Delta Stewardship Council) for funding this research. Findings and conclusions in this article are those of the authors and do not necessarily represent the views of the universities, agencies, or departments with which they are associated.

References

- Agresti A. (2002) *Categorical Data Analysis*. Wiley & Sons, New Jersey, USA.
- Anderson E.C. (2010) Assessment of power of informative subsets of loci for population assignment; standard methods are upwardly biased. *Molecular Ecology* **10**, 701–10.
- Anderson E.C., Waples R.S. & Kalinowski S.T. (2008) An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* **65**, 1475–86.
- Bamshad M.J., Wooding S., Watkins W.S., Ostler C.T., Batzer M.A. & Jorde L.B. (2003) Human population genetic structure and inference of group membership. *American Journal of Human Genetics* **72**, 578–89.
- Banks M.A. & Jacobson D.P. (2004) Which genetic markers and GSI methods are more appropriate for defining marine distribution and migration of salmon? *North Pacific Anadromous Fish Commission Technical Note* **5**, 39–42.
- Banks M.A., Blouin M.S., Baldwin B.A., Rashbrook V.K., Fitzgerald H.A., Blankenship S.M. & Hedgecock D. (1999) Isolation and inheritance of novel microsatellites in Chinook salmon (*Oncorhynchus tshawytscha*). *Journal of Heredity* **90**, 281–8; errata *Journal of Heredity* **90**, U1–U1.
- Banks M.A., Rashbrook V.K., Calavetta M.J., Dean C.A. & Hedgecock D. (2000) Analysis of microsatellite DNA resolves genetic structure and diversity of Chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. *Canadian Journal of Fisheries and Aquatic Sciences* **57**, 915–27.
- Banks M.A., Eichert W. & Olsen J.B. (2003) Which genetic loci have greater population assignment power? *Bioinformatics* **19**, 1436–8.
- Beacham T.D., Candy J.R., McIntosh B., MacConnachie C., Tabata A., Kaukinen K., Deng L., Miller K.M., Withler R.E. & Varnavskaya N. (2005) Estimation of stock composition and individual identification of Chinook Salmon across the Pacific Rim by use of microsatellite and major histocompatibility complex variation. *Transactions of the American Fisheries Society* **134**, 1124–46.
- Bjornstad G. & Røed K.H. (2002) Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. *Animal Genetics* **33**, 264–70.
- Cairney M., Taggart J.B. & Hoyheim B. (2000) Atlantic salmon (*Salmo salar* L.) and cross-species amplification in other salmonids. *Molecular Ecology* **9**, 2175–8.
- Fisher F.W. (1994) Past and present status of Central Valley Chinook salmon. *Conservation Biology*, **8**, 870–3.
- Glover K.A., Dahle G., Westgaard J.I., Johansen T., Knutsen H. & Jørstad K.E. (2010) Genetic diversity within and among Atlantic cod (*Gadus morhua*) farmed in marine cages: a proof-of-concept study for the identification of escapees. *Animal Genetics* **41**, 515–22.
- Greig C.A. & Banks M.A. (1999) Five multiplexed microsatellite loci for rapid response run identification of California's endangered winter Chinook salmon. *Animal Genetics* **30**, 318–20.
- Greig C., Jacobson D.P. & Banks M.A. (2003) New tetranucleotide microsatellites for fine-scale discrimination among endangered Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology Notes* **3**, 376–9.
- Hedgecock D., Banks M.A., Rashbrook V.K., Dean C.A. & Blankenship S.M. (2001) Applications of population genetics to conservation of Chinook salmon diversity in the Central Valley. In: *Contributions to the Biology of Central Valley Salmonids* (Ed. by R.L. Brown), State of California Resources Agency Department of Fish and Game. *Fishery Bulletin* **179**, 45–70.
- Ivanova N.V., Dewaard J.R. & Hebert P.D.N. (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes* **6**, 998–1002.
- Kalinowski S.T. (2008) ONCOR software for genetic stock identification. <http://www.montana.edu/kalinowski/Software/ONCOR.htm>.
- Kurushima J.D., Lipinski M.J., Gandolfi B., Froenicke L., Grahn J.C., Grahn R.A. & Lyons L.A. (2012) Variation of cats under domestication: genetic assignment of domestic cats to breeds and worldwide random-bred populations. *Animal Genetics* **44**, 311–24.
- Martin-Lopez B., Montes C. & Benayas J. (2008) Economic valuation of biodiversity conservation: the meaning of numbers. *Conservation Biology* **22**, 624–35.
- Moran P., Teel D.J., Banks M.A. *et al.* (2013) Divergent life-history races do not represent Chinook salmon coast-wide: the importance of scale in Quaternary biogeography. *Canadian Journal of Fisheries and Aquatic Sciences* **70**, 415–35.
- Naish K.A. & Park L.K. (2002) Linkage relationships for 35 new microsatellite loci in Chinook salmon *Oncorhynchus tshawytscha*. *Animal Genetics* **33**, 316–8.
- Nelson R.J. & Beacham T.D. (1999) Isolation and cross species amplification of microsatellite loci useful for study of Pacific salmon. *Animal Genetics* **30**, 228–9.

- Niu L.L., Li H.B., Ma Y.H. & Du L.X. (2011) Genetic variability and individual assignment of Chinese indigenous sheep populations (*Ovis aries*) using microsatellites. *Animal Genetics* **43**, 108–11.
- NMFS (2009) Biological opinion and conference opinion on the long-term operations of the Central Valley Project and State Water Project. www.nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=2147. File ARN: 151422SWR2004SA9116.
- Olsen J.B., Bentzen P. & Seeb J.E. (1998) Characterization of seven microsatellite loci derived from pink salmon. *Molecular Ecology* **7**, 1083–90.
- O'Malley K.G., Camara M.D. & Banks M.A. (2007) Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **16**, 4930–41.
- O'Malley K., McClelland E.K. & Naish K.A. (2010) Clock genes localize to stage-specific quantitative trait loci for growth in juvenile coho salmon, *Oncorhynchus kisutch*. *Journal of Heredity* **101**, 628–32.
- Rexroad C.E. III, Coleman R.L., Martin A.M., Hershberger W.K. & Killefer J. (2001) Thirty-five polymorphic microsatellite markers for rainbow trout (*Oncorhynchus mykiss*). *Animal Genetics* **32**, 317–9.
- Seeb L.W., Antonovich A., Banks M.A. *et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries* **32**, 540–2.
- Sokal R.R. & Rohlf F.J. (1995) *Biometry*. Freeman, San Francisco.
- Tadano R., Nishibori M. & Tsudzuki M. (2008) High accuracy of genetic discrimination among chicken lines obtained through an individual assignment test. *Animal Genetics* **39**, 567–71.
- Van de Goor L.H.P., Koskinen M.T. & Van Haeringen W.A. (2011) Population studies of 16 bovine STR loci for forensic purposes. *International Journal of Legal Medicine* **125**, 111–9.
- Williamson K.S., Cordes J.F. & May B. (2002) Characterization of microsatellite loci in Chinook salmon (*Oncorhynchus tshawytscha*) and cross-species amplification in other salmonids. *Molecular Ecology Notes* **2**, 17–9.
- Yoshiyama R.M., Gerstung E.R., Fisher F.W. & Moyle P.B. (2001) Historical and present distribution of Chinook Salmon in the Central Valley drainage of California. In: *Contributions to the Biology of Central Valley Salmonids* (Ed. by R.L. Brown), State of California The Resources Agency Department of Fish and Game. *Fishery Bulletin* **179**, 71–176
- Zar J.H. (2010) *Biostatistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, USA.