2014

Evaluation of a Single Nucleotide Polymorphism Baseline for Genetic Stock Identification of Chinook Salmon (Oncorhynchus tshawytscha) in the California Current Large Marine Ecosystem

Anthony J. Clemento

Eric D. Crandall
California State University, Monterey Bay, ecrandall@csumb.edu

John Carlos Garza

Eric C. Anderson

Recommended Citation

Clemento AJ, Crandall ED, Garza JC, Anderson EC (2014) Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (Oncorhynchus tshawytscha) in the California Current Large Marine Ecosystem. Fishery Bulletin 112(2-3): 112-130. http://dx.doi.org/10.7755/FB.112.2-3.2

Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) in the California Current large marine ecosystem

Anthony J. Clemento
Eric D. Crandall
John Carlos Garza
Eric C. Anderson (contact author)

Email address for contact author: eric.anderson@noaa.gov

Fisheries Ecology Division
Southwest Fisheries Science Center
National Marine Fisheries Service, NOAA
110 Shaffer Road
Santa Cruz, California 95060

Institute of Marine Sciences
University of California, Santa Cruz
110 Shaffer Road
Santa Cruz, California 95060

Abstract—Chinook Salmon (*Oncorhynchus tshawytscha*) is an economically and ecologically important species, and populations from the west coast of North America are a major component of fisheries in the North Pacific Ocean. The anadromous life history strategy of this species generates populations (or stocks) that typically are differentiated from neighboring populations. In many cases, it is desirable to discern the stock of origin of an individual fish or the stock composition of a mixed sample to monitor the stock-specific effects of anthropogenic impacts and alter management strategies accordingly. Genetic stock identification (GSI) provides such discrimination, and we describe here a novel GSI baseline composed of genotypes from more than 8000 individual fish from 69 distinct populations at 96 single nucleotide polymorphism (SNP) loci. The populations included in this baseline represent the likely sources for more than 99% of the salmon encountered in ocean fisheries of California and Oregon. This new genetic baseline permits GSI with the use of rapid and cost-effective SNP genotyping, and power analyses indicate that it provides very accurate identification of important stocks of Chinook Salmon. In an ocean fishery sample, GSI assignments of more than 1000 fish, with our baseline, were highly concordant (98.95%) at the reporting unit level with information from the physical tags recovered from the same fish. This SNP baseline represents an important advance in the technologies available to managers and researchers of this species.

Manuscript submitted 13 February 2013.
Manuscript accepted 10 February 2014.
Fish. Bull. 112:112–130 (2014).
doi:10.7755/FB.112.2-3.2

The views and opinions expressed or implied in this article are those of the author (or authors) and do not necessarily reflect the position of the National Marine Fisheries Service, NOAA.

Chinook Salmon (*Oncorhynchus tshawytscha*) are found in rivers from central California around the North Pacific Rim and the Bering Sea to Russia and are the target of valuable commercial and recreational fisheries. A key aspect of the life history of Chinook Salmon is natal homing, whereby each fish of this anadromous species typically returns to spawn in the same river in which it was born. This homing generates populations (or stocks) that may be genetically differentiated from neighboring populations and can exhibit local adaption (Utter et al., 1989; Taylor, 1991). Recent population declines, particularly at the southern end of the native range of this species, have resulted in the listing of many stocks under the U.S. Endangered Species Act (ESA; Federal Register, 1990, 1999) and have highlighted the need to refine the management and conservation of Chinook Salmon. However, such refinements are challenging because the migratory life history of salmon means that the many effects from anthropogenic sources that occur in rivers or in the ocean (e.g., fisheries, water diversion, or turbine entrainment) may affect multiple, intermingled stocks. In such cases, it may be necessary to discern the stock of origin of affected fish to monitor stock-specific impacts and design management strategies accordingly.

The use of pre-existing biological markers to distinguish salmon stocks has a long history. The traits used in these efforts have included morphometric and meristic characters (Fournier et al., 1984; Claytor and MacCrimmon, 1988), scale patterns (Cook, 1982), parasite assemblages (Boyce et al., 1985), and stable isotope ratios (Barnett-Johnson et al., 2008). However, the most universally applicable methods have involved the use of genetic markers because every fish has a unique genetic makeup. The first genetic markers widely used for identification in salmon were electrophoretically detectable protein polymorphisms known as allozymes (Milner et al., 1985; Shaklee and Phelps, 1990; Tessier et al., 1995; Allendorf and Seeb, 2000). With the advent of polymerase chain reaction (PCR), many more types of genetic markers became available to

discriminate salmon populations, including mitochondrial DNA polymorphisms (Cronin et al., 1993), minisatellites (Beacham et al., 1996; Miller et al., 1996), microsatellites (Seeb et al., 2007; Moran et al., 2013), amplified-fragment length polymorphisms (Flannery et al., 2007) and, most recently, single nucleotide polymorphisms (SNPs; Smith et al., 2005a, 2005b; Aguilar and Garza, 2008; Narum et al. 2008; Abadía-Cardoso et al., 2011; Clemento et al., 2011).

Genetic stock identification (GSI) typically proceeds in 2 steps. First, samples are collected from potential source populations and genotyped with a set of genetic markers in order to estimate population allele frequencies. These genotypes are called the "baseline." Then, data from individuals sampled from a mixed-stock collection (called a "mixture") and genotyped with the same set of genetic markers are compared with the baseline to estimate the relative proportions of individuals that came from each of the represented source populations. Single individuals of unknown origin also can be assigned to specific populations. Maximum likelihood or Bayesian methods typically are used to carry out GSI inference (Smouse et al., 1990; Pella and Masuda, 2000).

For the first large-scale baseline for GSI of Chinook Salmon allozyme markers were used (Teel et al.[1]), but technical and logistical issues limited their future appeal. The allozyme database was supplanted in Canada by a microsatellite baseline developed by the Department of Fisheries and Oceans (Beacham et al., 2006) and more broadly by a microsatellite baseline database developed through a large, international collaboration (Seeb et al., 2007). This collaboration required enormous effort to standardize data across laboratories because microsatellite allele names and sizes usually are not consistent between different laboratories and genotyping equipment.

The Seeb et al. (2007) microsatellite baseline has been an effective tool for GSI but has a number of disadvantages: genotyping and scoring of microsatellites is labor-intensive; genotyping error rates can be relatively high, making the 13 microsatellites in that baseline inadequate for applications such as pedigree reconstruction (Anderson and Garza, 2006; Garza and Anderson[2], Abadía-Cardoso et al., 2013); missing data rates also can be quite high; and, finally, any new laboratory that wishes to use that baseline must undertake a costly standardization process. Additionally, it now has been demonstrated that SNPs, despite typically having only 2 alleles per locus, do have sufficient power to be employed successfully in a GSI context with a modest number of genetic markers (Smith et al., 2007; Narum et al., 2008; Templin et al., 2011; Larson et al., 2013).

Early simulation studies indicated that the bi-allelic nature of SNPs would make them less useful than highly polymorphic microsatellites for population discrimination (Bernatchez and Duchesne, 2000; Kalinowski, 2004). However, SNPs are located throughout the genome and may be discovered in genetic regions with higher than average divergence (Nosil et al., 2009), increasing their utility for GSI. Moreover, SNPs do not suffer from many of the disadvantages of using microsatellites: SNP markers are amenable to the automated, high-throughput genotyping required for large projects; SNP genotyping error rates are very low, making them suitable for pedigree reconstruction; and, importantly, SNP assays typically do not require standardization between labs and, therefore, a SNP baseline is immediately useful to any group or agency that genotypes a mixture sample with the markers used in that baseline (Seeb et al., 2011).

Here, we describe the development and evaluation of a new baseline of SNP marker data for Chinook Salmon in the southern part of their native range for use in ecological investigation in the California Current large marine ecosystem (and its tributaries) and in fisheries managed by the Pacific Fishery Management Council (PFMC). We introduce a panel of 96 SNP markers and a baseline of more than 8000 salmon from 68 populations of Chinook Salmon ranging from California to Alaska and a single collection of Coho Salmon (*Oncorhynchus kisutch*) from California. We describe the procedures used to select these SNP markers from among a larger number of candidates and document the resulting patterns of genetic differentiation between various populations. We evaluate the power of this new baseline for GSI by both self-assignment (genetic identification of the most likely population of origin) and simulated mixture analyses, focusing on stocks commonly encountered in PFMC fisheries. Finally, we analyze 2090 fish sampled in 2010 from the sport and commercial fisheries off the coast of California and compare the results of these analyses with the coded wire tag (CWT) data from these fish to demonstrate the effectiveness of this baseline for classifying individuals to specific management units.

Materials and methods

Baseline populations

Populations were selected for inclusion in the new baseline to provide broad geographic coverage across

[1] Teel, D. J., P. A. Crane, C. M. Guthrie III, A. R. Marshall, D. M. Van Doornik, W. Templin, N. V. Varnavskaya, and L. W. Seeb. 1999. Comprehensive allozyme database discriminates Chinook salmon around the Pacific Rim. NPAFC document 440, 25 p. [Available from Alaska Department of Fish and Game, Division of Commercial Fisheries, 333 Raspberry Rd., Anchorage, AK 99518.]

[2] Garza, J. C., and E. C. Anderson. 2007. Large scale parentage inference as an alternative to coded-wire tags for salmon fishery management. *In* PSC genetic stock identification workshop: Logistics Workgroup final report and recommendations; Portland, OR, 15–17 May 2007 and Vancouver, Canada, 11–13 September 2007, p. 48–55 p. [Available from Pacific Salmon Commission, 600-1155 Robson St., Vancouver, BC V6E 1B5, Canada.]

the range of Chinook Salmon in the the United States from Washington to California, while also allowing for the identification of fish from elsewhere in the geographic range of this species. Adult fish were sampled on spawning grounds, in terminal fisheries, or at hatcheries during the period of 2003–13 and were provided by numerous contributors (see the *Acknowledgments* section and Warheit et al.[3]). We included populations expected to be encountered in ocean fisheries off California and Oregon, as well as populations with special management status (e.g., ESA-listed populations). Accordingly, the major lineages of Chinook Salmon from California and Oregon were emphasized in this baseline, as were populations distinguished by life history strategy (e.g., spring-run, fall-run, and winter-run strategies), but representatives of the major lineages from farther north also were included.

DNA was extracted from samples for California populations with DNeasy Blood & Tissue Kits on a BioRobot 3000[4] platform (QIAGEN, Inc., Valencia, CA) according to the manufacturer's protocols, and DNA from populations in Oregon, Washington, Canada, and Alaska was extracted by contributors (see *Acknowledgments* section) who used various methods. Sample sizes ranged from 44 to 1409 individuals per population and averaged 116 individuals per population. The 1409 fish from the population in the Trinity River Hatchery initially were genotyped with our SNP panel for another purpose, but they were included here in total to provide a comprehensive reference sample for identification of this important group. Excluding this disproportionately large sample, the average number of individuals per population was 97. In total, the new baseline included 7984 Chinook Salmon from 68 distinct populations (Table 1).

Each population in this baseline belongs to a single reporting unit, a designation established in previous GSI research that reflects a combination of "genetic similarity, geographic features, and management applications" (Seeb et al., 2007). Reporting units generally are composed of multiple populations that share genetic similarity or are subject to similar management regimes. The 68 populations of Chinook Salmon in our baseline fall into 38 distinct reporting units (Table 1), and some reporting units in Alaska and Canada are represented by only a single population.

Coho Salmon occasionally are misidentified as Chinook Salmon in ocean fisheries and in ecological sampling. We included a collection of 47 Coho Salmon from California as the 69th population in our baseline to help us to identify Coho Salmon that have been identified incorrectly as Chinook Salmon.

Markers and genotyping

We compiled a list of 192 TaqMan (Life Technologies Corp., Carlsbad, CA), or 5'-nuclease, SNP genotyping assays from previously published discovery studies (Smith et al., 2005a, 2005b; Campbell and Narum, 2008; Narum et al., 2008; Clemento et al., 2011) to test their scorability and power for GSI. TaqMan technology combines standard PCR primers that target the genomic region around a SNP with 2 different fluorescent probes that identify the 2 nucleotide bases present at the SNP. As recommended by the manufacturer, we used a multiplex preamplification reaction to increase the copy number of targeted genomic regions. Multiplex PCR products were diluted with 15 μL of 2 mM Tris buffer and were frozen.

Samples then were genotyped on 96.96 Dynamic Arrays with an EP1 System (Fluidigm Corp., South San Francisco, CA) according to the manufacturer's protocols. Fluidigm Dynamic Arrays use integrated nanofluidic circuitry to simultaneously determine the genotype at 96 SNP loci for 96 samples (2 of which are no-DNA template controls). Genotypes were determined with the Fluidigm Genotyping Analysis software (vers. 2.1.1). The use of quantitative PCR methods for genotype determination involves discerning, on a 2-D graph, clusters of fluorescence intensity of the probes for the 2 alleles; the 2 homozygote clusters have fluorescence primarily from only 1 probe, but a heterozygote cluster has similar intensities from both probes.

Marker selection

We selected a panel of 95 SNP markers from among the 192 candidates, reserving 1 marker for a species identification assay (see final paragraph of this section). The risk of "high-grading bias" (i.e., wrongly inflating the apparent resolving power of a group of loci for GSI) is particularly great when selecting a panel of markers to distinguish between populations that are closely related, as many of the populations in our baseline are. To avoid high-grading bias, we employed the "training-holdout-leave-one-out" (THL) procedure of Anderson (2010); this procedure requires that data be split into training and holdout sets. Training-set genotypes are used to select the loci included in a baseline and can be included in the eventual baseline, but they are not used to evaluate its performance. Rather, performance of a baseline is determined with simulation and self-assignment with only the holdout set, which was not used in any way to select baseline loci. We chose a training set of 372 individuals drawn from 22 populations (14 from California, 3 from Oregon, 3 from Washington, 1 from British Columbia, and 1 from Alaska) for initial genotyping with all 192 loci.

[3] Warheit, K. I., L. W. Seeb, W. D. Templin, and J. E. Seeb. 2013. Moving GSI into the next decade: SNP coordination for Pacific Salmon Treaty fisheries. FPT 13-09, 47 p. [Available from Washington Department of Fish and Wildlife, 600 Capitol Way N., Olympia, WA 98501-1091.]

[4] Mention of trade names or commercial companies is for identification purposes only and does not imply endorsement by the National Marine Fisheries Service, NOAA.

Table 1

Populations and reporting units in the single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) from the west coast of North America. Shown are the names used on the phylogeographic tree (Fig. 1), the total number of individuals sampled (n), the number used in the training set (n_{train}), estimates of unbiased (Unb.) and observed (Obs.) heterozygosity (Hz), and the mean number of alleles (A); also shown are the proportion of individuals that self-assign (Assign.) to the population (pop.) from which they were sampled and the proportion that self-assign to the correct reporting (rep.) unit, as well as the mean F_{ST} for each population within and between reporting units. The slash (/) in reporting unit names indicates the inclusion of multiple run types (e.g. summer and fall) or geographic regions (e.g. North California and South Oregon) in the same reporting unit. Note that mean summary values shown were calculated excluding the sample of Coho Salmon (*Oncorhynchus kisutch*). The use of "spring," "summer," "fall," or "winter" in a unit or population name specifies to which run it belongs. N. = Northern; S. = Southern; R. = River; Up. = upper; Hatch. = Hatchery.

Reporting unit	Population	Tree name	n	n_{train}	Unb. Hz	Obs. Hz	A	Assign. to pop.	Assign. to rep. group	Mean F_{ST} within	Mean F_{ST} between
Central Valley spring	Butte Creek spring	CVsp_Butte	425	26	0.357	0.330	1.99	0.68	0.93	0.017	0.196
	Mill Creek spring	CVsp_Mill	145	23	0.371	0.377	1.99	0.48	0.80	0.012	0.173
	Deer Creek spring	CVsp_Deer	119	12	0.367	0.346	1.99	0.50	0.80	0.013	0.174
	Up. Sacramento R. spring	CVsp_UpSac_late	372	–	0.368	0.355	1.99	0.26	0.78	0.008	0.175
Central Valley fall	Feather R. Hatchery spring	CVfl_FeatherRHsp	470	47	0.373	0.374	1.99	0.44	0.87	0.009	0.179
	Feather River Hatchery fall	CVfl_FeatherRHfl	146	23	0.370	0.371	1.98	0.18	0.85	0.004	0.190
	Butte Creek fall	CVfl_Butte	188	–	0.369	0.355	2.00	0.13	0.91	0.003	0.187
	Mill Creek fall	CVfl_Mill	97	12	0.366	0.358	1.98	0.14	0.95	0.004	0.200
	Deer Creek fall	CVfl_Deer	70	–	0.363	0.347	1.98	0.29	0.90	0.005	0.195
	Mokelumne River fall	CVfl_Mklmne	95	27	0.370	0.373	1.98	0.26	0.94	0.005	0.198
	Battle Creek fall	CVfl_Battle	141	23	0.369	0.351	1.99	0.29	0.89	0.005	0.188
	Up. Sacramento R. late-fall	CVfl_UpSac	93	23	0.367	0.364	2.00	0.54	0.93	0.010	0.193
Central Valley winter	Sacramento River winter	Sac_win	295	19	0.297	0.289	1.97	1.00	1.00	–	0.263
California Coast	Eel River	CACoast_Eel	95	12	0.327	0.321	2.00	0.89	0.96	0.029	0.203
	Russian River	CACoast_Russian	94	–	0.368	0.372	2.00	0.84	0.98	0.029	0.156
Klamath River	Klamath Iron Gate Hatchery	Klamath_IronGH	117	12	0.326	0.345	1.97	0.97	0.99	0.053	0.232
	Trinity River Hatchery	Klamath_TrinityHsp	1409	12	0.318	0.312	2.00	0.93	0.97	0.053	0.243
North California/ South Oregon Coast	Smith River	nCal_sOR_Smith	159	–	0.377	0.381	1.99	0.77	0.87	0.014	0.138
	Chetco River	nCal_sOR_Chetco	94	11	0.372	0.367	1.99	0.73	0.86	0.014	0.137
Rogue River	Cole Rivers Hatchery	Rogue_ColeRHsp	141	11	0.367	0.362	2.00	0.62	0.86	0.006	0.155
	Applegate Creek	Rogue_Applgt	92	–	0.369	0.361	2.00	0.50	0.77	0.006	0.153
Mid Oregon Coast	Coquille River	mOR_Coquille	47	–	0.352	0.343	1.99	0.72	0.83	0.039	0.151
	Umpqua River spring	mOR_Umpqua	137	11	0.386	0.375	2.00	0.63	0.64	0.055	0.119
	Siuslaw River	mOR_Siuslaw	93	–	0.345	0.348	1.98	0.40	0.46	0.040	0.146
North Oregon Coast	Nestucca Hatchery	nOR_NestuccaH	48	–	0.338	0.328	1.96	0.71	0.83	0.029	0.160
	Alsea River	nOR_Alsea	131	–	0.335	0.309	2.00	0.47	0.76	0.042	0.159
	Nehalem River	nOR_Nehalem	93	–	0.316	0.317	1.96	0.97	0.99	0.059	0.193
	Siletz River	nOR_Siletz	93	–	0.331	0.330	1.98	0.69	0.81	0.031	0.163

Table 1 (cont.)

Populations and reporting units in the single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) from the west coast of North America. Shown are the names used on the phylogeographic tree (Fig. 1), the total number of individuals sampled (n), the number used in the training set (n_{train}), estimates of unbiased (Unb.) and observed (Obs.) heterozygosity (Hz), and the mean number of alleles (A); also shown are the proportion of individuals that self-assign (Assign.) to the population (pop.) from which they were sampled and the proportion that self-assign to the correct reporting (rep.) unit, as well as the mean F_{ST} for each population within and between reporting units. The slash (/) in reporting unit names indicates the inclusion of multiple run types (e.g. summer and fall) or geographic regions (e.g. North California and South Oregon) in the same reporting unit. Note that mean summary values shown were calculated excluding the sample of Coho Salmon (*Oncorhynchus kisutch*). The use of "spring," "summer," "fall," or "winter" in a unit or population name specifies to which run it belongs. N. = Northern; S. = Southern; R. = River; Up. = upper; Hatch. = Hatchery.

Reporting unit	Population	Tree name	n	n_{train}	Unb. Hz	Obs. Hz	A	Assign. to pop.	Assign. to rep. group	Mean F_{ST} within	Mean F_{ST} between
Willamette River	North Santiam Hatchery	Willamette_NSantiamH	93	–	0.324	0.327	1.95	0.80	0.99	0.014	0.181
	McKenzie Hatchery	Willamette_McKenzHsp	48	–	0.334	0.376	1.94	0.69	0.96	0.014	0.175
Deschutes fall	Lower Deschutes River	Deschutes_fl	94	–	0.366	0.357	2.00	0.56	0.56	–	0.145
Lower Columbia fall	Cowlitz Hatchery fall	COlow_CowHfl	141	–	0.365	0.374	1.99	0.79	0.79	–	0.142
Lower Columbia spring	Cowlitz Hatchery spring	COlow_CowHsp	44	11	0.368	0.370	1.97	0.67	0.70	0.029	0.160
	Kalama Hatchery spring	COlow_KalamaHsp	48	12	0.372	0.359	1.99	0.50	0.61	0.029	0.134
Mid Columbia Tule fall	Spring Creek Hatchery	COmid_SpringCH	142	–	0.322	0.331	1.97	0.97	0.97	–	0.206
Upper Columbia summer/fall	Hanford Reach	COup_Hanford	92	–	0.355	0.353	1.99	0.36	0.76	0.002	0.166
	Priest Rapids Hatchery	COup_PriestHsumfl	48	–	0.361	0.359	1.99	0.25	0.83	0.002	0.164
	Wells Hatchery	COup_WellsHsumfl	48	–	0.355	0.369	1.99	0.46	0.92	0.004	0.175
Mid and Upper Columbia spring	Wenatchee River	COmup_Wenatchee	48	–	0.209	0.202	1.88	0.85	0.85	0.048	0.260
	Cle Elum Hatchery	COmup_CleEHsp	48	–	0.262	0.255	1.95	0.94	0.96	0.048	0.219
Snake River fall	Lyons Ferry Hatchery	Snake_LyonsFHfl	119	12	0.359	0.360	2.00	0.45	0.45	–	0.158
Snake River spring/summer	Rapid River Hatchery	Snake_RapidRHsumsp	48	–	0.191	0.194	1.84	0.85	0.94	0.034	0.272
	McCall Hatchery	Snake_MCallHsumsp	48	–	0.199	0.196	1.84	0.75	0.96	0.034	0.278
Washington Coast	Forks Creek Hatchery	WACoast_ForksCH	93	–	0.350	0.345	1.99	0.89	0.94	0.042	0.143
	Quinalt Lake fall	WACoast_Quinalt	48	–	0.348	0.341	1.98	0.90	0.96	0.042	0.152
South Puget Sound	Soos Creek Hatchery	sPuget_SoosCH	142	–	0.360	0.358	2.00	0.91	0.91	–	0.158
North Puget Sound	Kendall Hatchery spring	nPuget_KendlHsp	48	–	0.326	0.336	1.95	0.92	0.96	0.042	0.170
	Marblemount Hatch. spring	nPuget_MrblHsp	48	–	0.343	0.337	1.99	0.92	0.94	0.042	0.156
Lower Fraser	Harrison River	Fraser_Harris	48	–	0.329	0.326	1.98	0.96	0.96	0.152	0.169
	Birkenhead Hatchery	Fraser_BirkenH	91	–	0.259	0.255	1.84	1.00	1.00	0.152	0.231
Lower Thompson River	Spius Creek Hatchery	Thompson_SpiusCH	46	11	0.271	0.275	1.89	1.00	1.00	–	0.201
East Vancouver Island	Big Qualicum Hatchery	eVancI_BigQual	48	–	0.352	0.338	2.00	0.83	0.83	–	0.145

Table 1 (cont.)

Populations and reporting units in the single nucleotide polymorphism baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) from the west coast of North America. Shown are the names used on the phylogeographic tree (Fig. 1), the total number of individuals sampled (n), the number used in the training set (n_{train}), estimates of unbiased (Unb.) and observed (Obs.) heterozygosity (Hz), and the mean number of alleles (A); also shown are the proportion of individuals that self-assign (Assign.) to the population (pop.) from which they were sampled and the proportion that self-assign to the correct reporting (rep.) unit, as well as the mean F_{ST} for each population within and between reporting units. The slash (/) in reporting unit names indicates the inclusion of multiple run types (e.g. summer and fall) or geographic regions (e.g. North California and South Oregon) in the same reporting unit. Note that mean summary values shown were calculated excluding the sample of Coho Salmon (*Oncorhynchus kisutch*). The use of "spring," "summer," "fall," or "winter" in a unit or population name specifies to which run it belongs. N. = Northern; S. = Southern; R. = River; Up. = upper; Hatch. = Hatchery.

Reporting unit	Population	Tree name	n	n_{train}	Unb. Hz	Obs. Hz	A	Assign. to pop.	Assign. to rep. group	Mean F_{ST} within	Mean F_{ST} between
West Vancouver Island	Robertson Hatchery	wVancI_RobHfl	48	–	0.341	0.364	1.98	0.96	0.96	–	0.152
Lower Skeena River	Lower Kalum River	lSkeena_Kalum	48	–	0.303	0.303	1.96	0.77	0.77	–	0.156
Mid Skeena River	Morice River	mSkeena_Morice	47	–	0.279	0.276	1.89	0.81	0.91	0.013	0.173
	Kitwanga River	mSkeena_Kitwanga	48	–	0.291	0.290	1.94	0.54	0.75	0.013	0.175
S. Southeast Alaska	Little Port Walter - Unuk	sSEAK_Unuk	48	–	0.301	0.290	1.94	0.79	0.79	–	0.165
Alsek River	Goat Creek	AlsekAK_Goat	48	–	0.243	0.245	1.69	0.96	0.96	–	0.248
Karluk River	Karluk River	KarlukAK	47	–	0.230	0.220	1.73	1.00	1.00	–	0.237
Taku River	Little Tatsamenie Lake	Taku_LilTats	48	–	0.271	0.265	1.92	0.90	0.90	–	0.188
Chilkat River	Pullen Creek Hatchery	nSEAK_PullenCH	48	–	0.260	0.276	1.77	0.98	0.98	–	0.209
Situk River	Situk River	SitukAK	48	12	0.244	0.248	1.77	0.94	0.94	–	0.210
Copper River	Sinona Creek	CopperAK_Sinona	47	–	0.229	0.226	1.63	0.98	0.98	–	0.244
Susitna River	Montana Creek	SusitnaAK_Montana	48	–	0.210	0.201	1.73	0.92	0.92	–	0.249
Western AK, Lower	George River	WestAK_George	47	–	0.234	0.229	1.78	0.43	0.98	0.004	0.239
Kuskokwim River	Kanektok River	WestAK_Kanektok	48	–	0.241	0.232	1.81	0.38	0.96	0.001	0.233
	Togiak River	WestAK_Togiak	48	–	0.241	0.229	1.79	0.40	0.94	0.005	0.233
Yukon River	Kantishna River	Yukon_Kantishna	48	–	0.208	0.204	1.67	0.94	0.94	–	0.269
Coho Salmon	California Coho	Coho	47	–	0.089	0.094	1.33	1.00	1.00	–	0.463
		total	8031	mean	0.320	0.317	1.93	0.69	0.88	0.028	0.188

For each locus, k, the observed relative frequencies, p_{ik} and q_{ik}, of the 2 SNP alleles were calculated for each population, i, in the training set. These values then were used to compute the expected probability of misassignment, $P(\mathrm{Mis}_{ijk})$, between every pair of populations i and j with only a single locus k:

$$P(\mathrm{Mis}_{ijk}) = 0.5\ [\delta(p_{ik} \leq p_{jk}) p_{ik}^2 + \delta(p_{ik} q_{ik} \leq p_{jk} q_{jk}) 2 p_{ik} q_{ik} + \delta(q_{ik} \leq q_{jk}) q_{ik}^2 + \delta(p_{ik} \geq p_{jk}) p_{jk}^2 + \delta(p_{ik} q_{ik} \geq p_{jk} q_{jk}) 2 p_{jk} q_{jk} + \delta(q_{ik} \geq q_{jk}) q_{jk}^2],$$

for all k where $\delta(x) = 1$ if the condition x is true and 0 if otherwise.

The values of $P(\mathrm{Mis}_{ijk})$ were used to rank the loci for their suitability for resolving between populations i and j in GSI; a lower $P(\mathrm{Mis}_{ijk})$ indicates better resolving power.

The rankings derived from $P(\mathrm{Mis}_{ijk})$ values were combined with other criteria in a nonautomated process to select the final panel of loci (Table 2). Each SNP assay was evaluated for scorability and evidence of Hardy-Weinberg disequilibrium and linkage disequilibrium (LD). Assays with overly dispersed clusters, more than 3 clusters, or inadequate spacing between clusters were excluded. Loci with significant deviations from equilibrium expectations also were removed. SNPs with large differences in allele frequencies between populations are particularly effective for GSI, whereas SNPs with high minor allele frequencies (MAFs) are most useful for parentage analysis (Anderson and Garza, 2006).

The remaining 168 loci were then ranked by their MAFs in hatchery populations to be included in pedigree reconstruction studies (see *Discussion* section). Previous simulations indicated that about 100 loci with an MAF >0.2 would be required to achieve the necessary statistical power to assign parentage with sufficiently low false-negative and false-positive rates (Anderson and Garza, 2006). However, the observed MAFs for many loci were in fact >0.2 (and as high as 0.5), indicating that the desired statistical power could be achieved with fewer loci. Therefore, we selected the 70 loci with the highest MAF in the Feather River population, the primary target for subsequent parentage investigations. We then used the $P(\mathrm{Mis}_{ijk})$ rankings to select 25 additional loci that were useful for distinguishing between difficult-to-resolve populations and reporting units. Finally, an assay to discriminate between Chinook and Coho salmon was included as the 96th assay for genotyping with the Fluidigm 96.96 Dynamic Arrays.

Population genetics analyses

The 7669 samples that were not in the training set for locus selection were genotyped with the final panel of 96 SNPs and used as the holdout set in subsequent power analyses (see the next section). This holdout set also was used for standard population genetics analyses. We tested each locus-population pair for deviations from Hardy-Weinberg equilibrium (HWE) with the complete enumeration method (Louis and Dempster, 1987) in GENEPOP software, vers. 4.0 (Rousset, 2008). Similarly, in each population, all pairwise locus combinations were investigated for LD. Default Markov chain parameters were used, except for the number of batches, which was increased to 500 to reduce the standard error to acceptable levels (<0.02; Rousset, 2008).

Genetic differentiation (F_{ST}) was estimated (with θ of Weir and Cockerham, 1984) between all pairs of populations with the software package GENETIX, vers. 4.05 (Belkhir[5]). The data set was permuted 1000 times to determine the significance of F_{ST} estimates. Phylogeographic trees were constructed with the chord distance (DCE) of Cavalli-Sforza and Edwards (1967) and the neighbor-joining algorithm in the software package PHYLIP, vers. 3.69 (Felsenstein[6]) and were visualized with Dendroscope software, vers. 3.2.10 (Huson et al., 2007). Majority-rule consensus values were calculated from 10,000 bootstrap samples of the data through the use of the PHYLIP component CONSENSE. The F_{ST} values and genetic distances computed are expected to provide an inflated estimate of the isolation between populations because the SNP loci used in our analyses were not a random sample from the genome; some SNP loci were chosen for their power in resolving specific population pairs in our baseline. Nonetheless, these estimates are useful for assessment of the relative genetic differentiation among the populations described here.

Power analyses

We used 3 different methods to assess the power of the SNP baseline for GSI. First, we performed a self-assignment analysis, and subsequently we generated and analyzed simulated mixtures with 2 different procedures.

In self-assignment analysis, allele frequencies for each potential source population generally are estimated from the samples. Then, for each individual, the probability that its genotype would occur in each population (assuming Hardy-Weinberg and linkage equilibria) is calculated, and the individual is assigned to the population for which its genotype probability is highest. We used the likelihood method of Rannala and Mountain (1997), implemented in the software gsi_sim[7] (Anderson et al., 2008), to compute the genotype prob-

[5] Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme. 1996–2004. GENETIX 4.05, logiciel sous WindowsTM pour le génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France. [Available from http://kimura.univ-montp2.fr/genetix.]

[6] Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package), vers. 3.6. Department of Genome Sciences, Univ.Washington, Seattle. [Available from http://evolution.genetics.washington.edu/phylip.html.]

[7] Available from http://swfsc.noaa.gov/textblock.aspx?Division=FED&ParentMenuID=54&id=12964.

Table 2

List of the 96 single nucleotide polymorphism (SNP) loci used to construct the baseline for genetic stock identification of Chinook Salmon (*Oncorhynchus tshawytscha*) from the west coast of North America, with dbSNP accession numbers (from the National Center for Biotechnology Information online repository for short genetic variations; https://www.ncbi.nlm.nih.gov/snp) and source reference (S_R) where available: 1=Clemento et al., 2011; 2=Smith et al., 2005a; 3=Campbell and Narum, 2008; 4=Smith et al., 2005b.

Locus	dbSNP	S_R	Locus	dbSNP	S_R	Locus	dbSNP	S_R
Ots_94857-232	ss275518685	1	Ots_110495-380	ss275518741	1	Ots_131906-141	ss275518787	1
Ots_96222-525	ss275518688	1	Ots_110551-64	ss275518742	1	Ots_AldB1-122	ss275518788	1
Ots_96500-180	ss275518689	1	OkiOts_120255-113	unpubl.	–	Ots_AldoB4-183	ss275518789	1
Ots_97077-179	ss275518691	1	Ots_111312-435	ss275518746	1	Ots_Myc-366	ss275518795	1
Ots_99550-204	ss275518695	1	Ots_111666-408	ss275518747	1	Ots_ALDBINT1-SNP1	ss275518796	1
Ots_100884-287	ss275518696	1	Ots_111681-657	ss275518748	1	Ots_NAML12-SNP1	ss275518798	1
Ots_101119-381	ss275518697	1	Ots_112208-722	ss275518749	1	Ots_ARNT-195	unpubl.	–
Ots_101704-143	ss275518699	1	Ots_112301-43	ss275518750	1	Ots_RAG3	unpubl.	–
Ots_102213-210	ss275518702	1	Ots_112419-131	ss275518751	1	Ots_AsnRS-60	ss48398657	2
Ots_102414-395	ss275518703	1	Ots_112820-284	ss275518752	1	Ots_aspat-196	ss65917744	3
Ots_102420-494	ss275518704	1	Ots_112876-371	ss275518753	1	Ots_CD59-2	unpubl.	–
Ots_102457-132	ss275518705	1	Ots_113242-216	ss275518754	1	Ots_CD63	unpubl.	–
Ots_102801-308	ss275518706	1	Ots_113457-40	ss275518755	1	Ots_EP-529	unpubl.	–
Ots_102867-609	ss275518707	1	Ots_117043-255	ss275518757	1	Ots_GDH-81x	ss65917741	3
Ots_103041-52	ss275518708	1	Ots_117242-136	ss275518759	1	Ots_HSP90B-385	ss65713207	2
Ots_104063-132	ss275518711	1	Ots_117432-409	ss275518762	1	Ots_MHC1	ss49851328	4
Ots_104569-86	ss275518714	1	Ots_118175-479	ss275518763	1	Ots_mybp-85	unpubl.	–
Ots_105105-613	ss275518715	1	Ots_118205-61	ss275518764	1	Ots_myoD-364	ss65917726	3
Ots_105132-200	ss275518716	1	Ots_118938-325	ss275518765	1	Ots_Ots311-101x	ss65917748	3
Ots_105401-325	ss275518718	1	Ots_122414-56	ss275518767	1	Ots_PGK-54	unpubl.	–
Ots_105407-117	ss275518719	1	Ots_123048-521	ss275518768	1	Ots_Prl2	ss49851322	4
Ots_106499-70	ss275518724	1	Ots_123921-111	ss275518770	1	Ots_RFC2-558	ss48398670	2
Ots_106747-239	ss275518725	1	Ots_124774-477	ss275518771	1	Ots_SClkF2R2-135	ss48398694	2
Ots_107074-284	ss275518726	1	Ots_127236-62	ss275518773	1	Ots_SWS1op-182	ss48398635	2
Ots_107285-93	ss275518728	1	Ots_128302-57	ss275518775	1	Ots_TAPBP	unpubl.	–
Ots_107806-821	ss275518730	1	Ots_128693-461	ss275518777	1	Ots_u07-07.161	unpubl.	–
Ots_108007-208	ss275518731	1	Ots_128757-61	ss275518778	1	Ots_u07-49.290	unpubl.	–
Ots_108390-329	ss275518732	1	Ots_129144-472	ss275518779	1	Ots_u4-92	ss48398636	2
Ots_108735-302	ss275518733	1	Ots_129170-683	ss275518780	1	Ots_BMP2-SNP1	ss275518800	1
Ots_109693-392	ss275518737	1	Ots_129458-451	ss275518782	1	Ots_TF1-SNP1	ss275518802	1
Ots_110064-383	ss275518738	1	Ots_130720-99	ss275518784	1	Ots_S71-336	unpubl.	–
Ots_110201-363	ss275518739	1	Ots_131460-584	ss275518785	1	Ots_unk_526	unpubl.	–

abilities, employing a leave-one-out procedure that excludes the gene copies of the individual being assigned and recalculates population allele frequencies before assignment. Analogous to the THL procedure of Anderson (2010), both the training and holdout sets were included for estimation of population allele frequencies. However, assignments of individuals in the training set were excluded from the results to avoid any high-grading bias of assignment accuracy (Anderson, 2010).

Analysis of simulated mixed fisheries is a common method for evaluation of the resolving power of a baseline for stock identification (Fournier et al., 1984; Wood et al., 1987; Kalinowski, 2004; Beacham et al., 2006). In many studies, samples from simulated fisheries that consist entirely of fish from one population are analyzed in so called "100% simulations." However, such simulations typically do not assess how well the baseline will perform on samples from fisheries that exploit more than one stock. Therefore, we conducted simulations with 20 different mixing proportion vectors. The population composition of these mixtures was determined by using the baseline to estimate the relative proportions of reporting units present in 20 different month-by-area strata sampled from commercial fisheries off the coast of California and Oregon in 2010 and 2011 (E. Crandall et al., unpubl. data).

These vectors reflect mixing proportions that are expected to be encountered in PFMC fisheries. For a given value of the mixing proportion vector of all populations, a replicate simulation consisted of 1) simulating the number of fish from each population in a sample size of 200 by drawing a multinomial random variable with cell probabilities equal to the mixing proportion vector; 2) simulating the genotypes of

the individuals from each population in the mixture sample with 2 different techniques ("cross-validation over gene copies" [CV-GC] and K-fold cross validation [K-fold], see next paragraph); 3) calculating the maximum likelihood estimator (MLE) of the mixture proportions for all the populations from the simulated sample through use of the baseline, which contains all training and holdout individuals; and 4) estimating the mixing proportion of each reporting unit by summing the mixing proportion estimates of its constituent populations. For each of the 20 values of the mixing proportion vectors, 20,000 replicates were conducted with CV-GC, and 1000 replicates were conducted with K-fold. For both methods, the 5% and 95% quantiles of the distribution of the MLE of reporting-unit proportions were calculated from the replicates for each mixing proportion vector.

Simulations were undertaken in 2 different ways. With CV-GC, genotypes were simulated by randomly sampling gene copies from the holdout set (to avoid high-grading bias), and those same gene copies were removed from the baseline when calculating the likelihood of population origin for the simulated individual (see Anderson et al., 2008). With K-fold, genotypes were simulated by drawing entire individuals without replacement (a technique commonly referred to as "jackknifing") from the holdout set to form the mixture sample. Those sampled individuals were not included in the baseline, but all unsampled individuals from the holdout set were included in the baseline for estimation of the mixing proportions.

Mixed fishery samples

Samples from 2090 salmon landed in fisheries in 2010 were collected by the California Department of Fish and Game (now Wildlife) at California ports. Just over half of these fish carried CWTs that identified their population of origin. All samples were genotyped with our panel of 96 loci. Individuals successfully genotyped at fewer than 60 loci were removed from further analysis. Failed genotypes were ones that either clustered with negative controls during scoring or fell outside of defined heterozygote and homozygote clusters, likely indicating sample contamination (Smith et al., 2011; Larson et al., 2013). We also used an individual heterozygosity (iHz; the proportion of heterozygous loci for each fish) criterion of iHz >0.56 to identify and exclude potentially contaminated samples. Simulations of contaminated genotypes determined by using observed allele frequencies, indicated little overlap in the distribution of iHz for contaminated and uncontaminated samples (data not shown) and that uncontaminated samples rarely had iHz >0.56.

We used the maximum likelihood framework in gsi_sim to estimate the mixing proportion of different populations among the 2090 fish, and then used that MLE as the prior for calculation of the posterior probability of population of origin for each fish. Posterior probabilities of origination from different reporting units were obtained through summation of the population-specific probabilities over all populations in a reporting unit. Individuals were then assigned to the reporting unit with the highest posterior probability.

Because all fish would be assigned to a maximum a posteriori (MAP) population regardless of true origin, we employed a simulation method similar to that in Cornuet et al. (1999), but which was modified to account for missing data, to detect fish that might have originated from a population that was not in the baseline or that had an otherwise aberrant genotype. Briefly, for each fish from the fishery assigned to a population, the allele frequencies from the MAP population were used to simulate 10,000 genotypes with an identical pattern of missing data (if any) to that of the fish that was assigned.

The log-probability of each simulated genotype was computed, given that it came from the population it was simulated from, and then the distribution of those values was compared with the log-probability, L_a, of the actual assigned fish's genotype, given the allele frequencies in the MAP population, on the basis of a z-score (L_a minus the mean of the simulated values, all divided by the standard deviation of the simulated values). The z-score calculation was done conditional on the exact pattern of missing data and was implemented in the C programming language as part of the gsi_sim software. A low-confidence assignment was defined to be one that had a z-score <3.0 and had either a reporting unit posterior probability <0.9 or had fewer than 90 loci successfully genotyped. Fish with low confidence assignments were left in an "unassigned" category.

Results

Genotyping and basic population genetics

We successfully genotyped 8031 samples from 69 populations for the baseline and submitted the data to the Dryad Digital Repository (http://doi.org/10.5061/dryad.5745sv). All individuals were retained in the baseline, regardless of missing data because we desired a realistic representation of missing data patterns for subsequent power analyses. One locus failed to amplify entirely in the Copper River population, and 3 loci failed in the Coho Salmon sample. Unbiased estimates of heterozygosity (Nei, 1978) ranged from 0.194 in the Rapid River Hatchery stock of the Snake River reporting unit to 0.381 in the Smith River population. The Coho Salmon in the baseline had very low heterozygosity (0.094). Observed heterozygosity and mean number of alleles generally were lower for populations from north of the Columbia River (Table 1), likely due to an ascertainment bias resulting from the selection of SNPs with high MAFs in California and Oregon populations.

Significant deviations from HWE (P<0.0001) were observed at various loci in 17 populations but represented <0.3% of all observations. Only the Butte Creek spring-run, Trinity River Hatchery spring-run, and Smith River populations were not in HWE at more than 2 loci, with 5, 5, and 4 significant tests, respectively. Similarly, only 3 loci deviated from HWE in more than 2 populations: Ots_u07_07.161 in 3 populations, Ots_111312-435 in 6 populations, and Ots_111666-408 in 4 populations. Only 1 population (Trinity River Hatchery spring run) displayed significant LD (P<0.001) at more than 1% of locus comparisons (1.14%), and, over all populations, the percentage of significant comparisons was 0.16%. Only 2 locus pairs were significant in more than 5 populations: Ots_AldB1-122 and Ots_AldoB4-183, known to be in the same gene complex, were in LD in 42 populations, and Ots_Myc-366 and Ots_unk-526 displayed LD in 8 populations.

A large range in the degree of differentiation between populations was observed (Table 1). Mean F_{ST} across all populations (excluding Coho Salmon) was 0.183, indicating that approximately 18% of genetic variation was partitioned between population samples. Within reporting units that contained more than one population (N=18), pairwise F_{ST} was between 0.000 and 0.152 and had a mean value of 0.018. Ten pairwise comparisons, all within reporting units, were not significantly different from zero (P<0.01). Between reporting units, F_{ST} values ranged from 0.005 to 0.411 and had a mean value of 0.188. The least differentiated populations were the fall-run populations from California's Central Valley, as has been observed with other genetic data sets (Williamson and May, 2005; Seeb et al., 2007).

Genetic structure of the Chinook Salmon populations in the baseline is displayed in an unrooted neighbor-joining dendrogram (Fig. 1). Relationships are in strong agreement with expectations that were based on geography and previous studies (Waples et al., 2004; Beacham et al., 2006; Templin et al., 2011; Moran et al., 2013); populations generally are organized north to south along the main branch, and populations from within the same drainage usually cluster together.

Populations from California's Central Valley are monophyletic in relation to the remainder of the populations but are characterized by short branch lengths, small distances between nodes, and low bootstrap support. Central Valley spring-run and fall-run populations also are monophyletic, with the exception of the Feather River Hatchery spring run, which is included in the fall-run reporting unit because of a history of substantial introgression between the runs and the consequent difficulty of genetically distinguishing this stock from fall-run fish (Garza et al.[8]). Sacramento River winter-run fish are quite distinct as a result of a well-documented recent bottleneck (Hedrick et al., 1995) and have one of the longest branches on the tree, with bootstrap support of 100%. Fish from rivers in northern California and coastal Oregon also form a monophyletic group. Columbia River populations are dispersed throughout the tree, although populations from the same reporting unit generally share a common branch, as do populations from Alaska.

Accuracy of assignment and mixture estimations

The 7669 individuals that remained after removal of training-set fish were subjected to self-assignment with gsi_sim (Table 1). Correct assignment to population ranged from 13% for the Butte Creek fall-run population to 100% for 5 different populations. The following reporting units had the lowest correct assignment rates to population: Central Valley fall run, Upper Columbia River summer/fall run, and Western Alaska, Lower Kuskokwim River, averaging 28%, 36%, and 40%, respectively. The lowest rate of correct assignment to reporting unit was for the Siuslaw River population from the Mid Oregon Coast reporting unit, with over half of the individuals assigning to populations in the North Oregon coast reporting unit. The largest change in correct assignment percentage from population to reporting unit was for the Central Valley fall run, which increased to 91%.

The results of the mixture simulations for the 9 reporting units most frequently found in California and Oregon fisheries appear in Figure 2. Results for the remaining reporting units are not shown because they are relatively uninformative as a result of the rarity with which populations from north of the Columbia River are encountered at the southern end of the California Current marine ecosystem, an observation corroborated by historical CWT data: in the 3 decades since 1983, only 0.5% of all CWTs recovered from Chinook Salmon in California ocean fisheries were from stocks outside of California or Oregon (Regional Mark Information System, Regional Mark Processing Center, http://www.rmpc.org). Accurate estimates of the mixing proportions were obtained for fishery samples simulated either by CV-GC or by K-fold. The mean maximum likelihood estimate of the proportion of each reporting unit was generally highly correlated with the true proportion, indicating that any bias was very small.

For 6 reporting units (Central Valley fall run, Sacramento River winter run, Klamath River, California Coast, Rogue River, and North Oregon Coast), the 5% and 95% quantiles for reporting-unit mixing proportions corresponded closely with the quantiles one would obtain with perfect identification of all fish (see the gray regions in Fig. 2). The somewhat wider GSI quan-

[8] Garza, J. C., S. M. Blankenship, C. Lemaire, and G. Charrier. 2008. Genetic population structure of Chinook Salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. Final report for CalFed project "Comprehensive evaluation of population structure and diversity for Central Valley Salmon," 54 p. [Available from http://swfsc.noaa.gov/publications/FED/01110.pdf.]

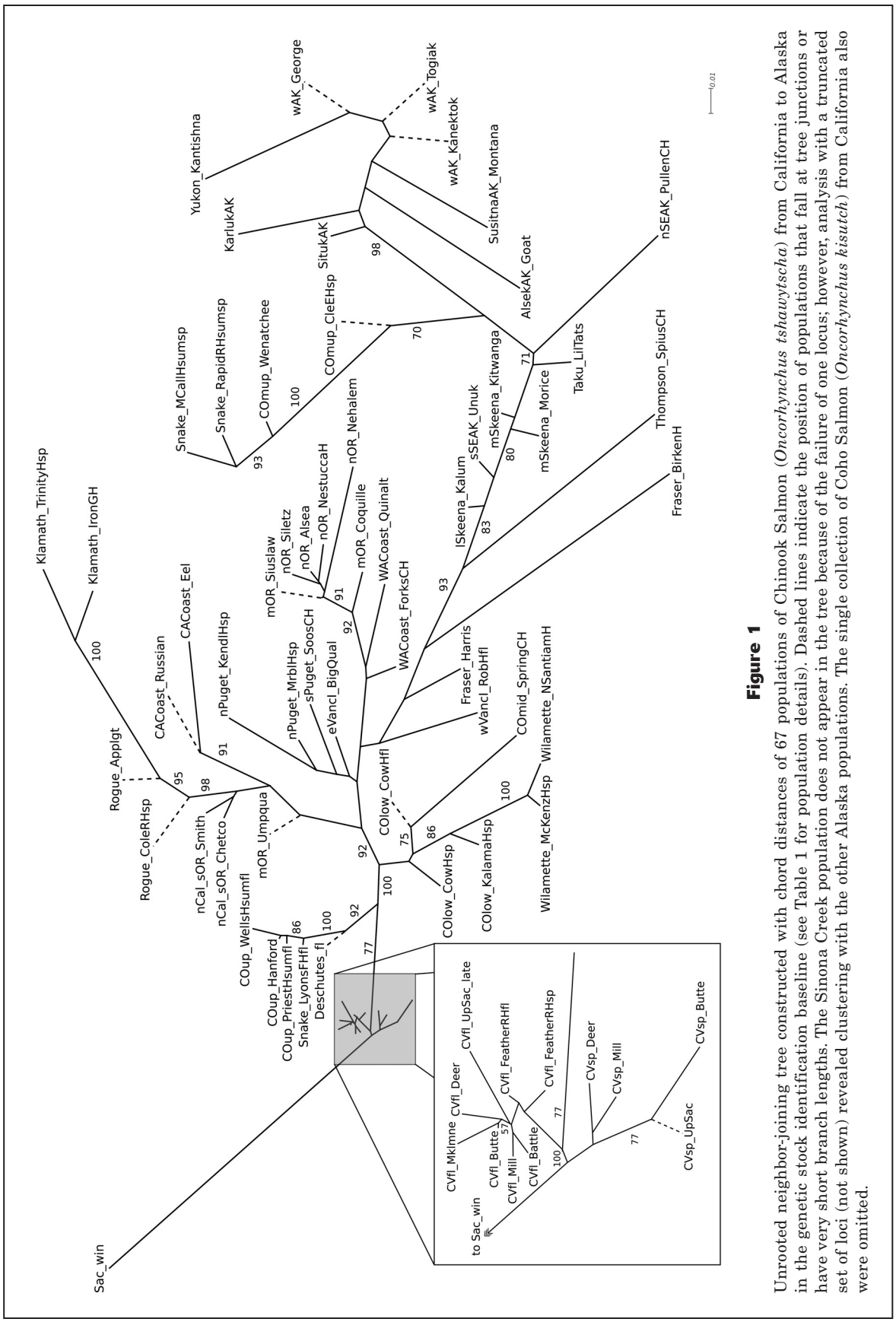

Figure 1

Unrooted neighbor-joining tree constructed with chord distances of 67 populations of Chinook Salmon (*Oncorhynchus tshawytscha*) from California to Alaska in the genetic stock identification baseline (see Table 1 for population details). Dashed lines indicate the position of populations that fall at tree junctions or have very short branch lengths. The Sinona Creek population does not appear in the tree because of the failure of one locus; however, analysis with a truncated set of loci (not shown) revealed clustering with the other Alaska populations. The single collection of Coho Salmon (*Oncorhynchus kisutch*) from California also were omitted.

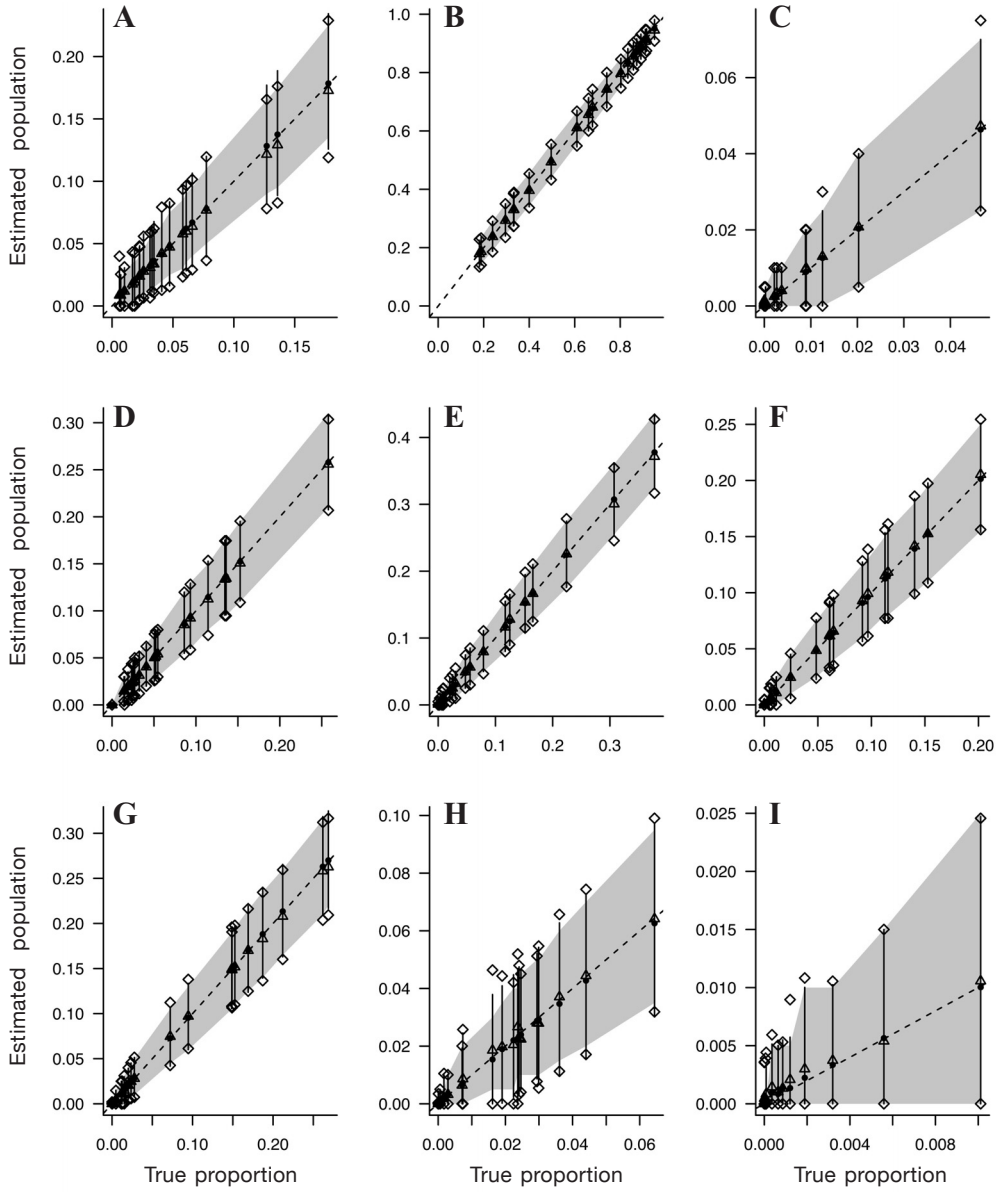

Figure 2

Estimates of mixing proportions from cross-validation over gene copies (CV-GC) and K-fold simulations for the 9 most abundant reporting units of Chinook Salmon (*Oncorhynchus tshawytscha*) encountered in California fisheries: Central Valley (**A**) spring, (**B**) fall, and (**C**) winter; (**D**) California Coast; (**E**) Klamath River; (**F**) North California/South Oregon Coast; (**G**) Rogue River; and **(H**) Mid Oregon Coast; and (**I**) North Oregon Coast. The *x*-axis gives the true proportion of fish from each reporting unit, and the *y*-axis gives the estimated proportion. The dashed line is the $y=x$ line. Gray shaded regions give the range between the 5% and 95% quantiles of estimates that would be achieved with perfect assignment of fish to a reporting unit (i.e., they represent the uncertainty due to the fact that fishery proportions are estimated with a finite sample; in our simulations, a sample of 200 fish). The 5% and 95% quantiles of the estimates derived from the CV-GC and the K-fold replicates are shown with vertical line segments and open diamonds, respectively. Reporting units for which these bars and diamonds coincide with the gray region had estimated proportions as accurate as one would expect given unambiguous identification of fish to reporting unit. Filled circles and open triangles indicate the mean over 20,000 CV-GC and 1000 K-fold replicates, respectively. These points fall along the dotted line when the estimator is unbiased.

Table 3

Genetic stock identification (GSI) results from assignment of samples of Chinook Salmon (*Oncorhynchus tshawytscha*) collected in 2010 from the California fishery to their source populations through the use of a single nucleotide polymorphism baseline, as well as concordance with recoveries of coded wire tags (CWTs). N.=North; S.=South.

Stock	Number from GSI	Number with CWT	Number of GSI~CWT matches	GSI~CWT agreement (%)
California Coast	30	1	0	0.00%
Central Valley fall	1581	958	957	99.90%
Central Valley spring	7	1	0	0.00%
Klamath River	108	50	49	98.00%
Lower Columbia spring	1	0	0	—
Mid Columbia Tule fall	7	2	2	100.00%
Mid Oregon Coast	14	1	0	0.00%
N. California/S. Oregon Coast	58	25	25	100.00%
Rogue River	154	11	5	45.45%
Snake River fall	1	1	1	100.00%
Upper Columbia summer/fall	8	2	2	100.00%
Total	1969	1052	1041	98.95%

tile intervals observed for the Central Valley spring-run reporting unit were likely due to its similarity to the Central Valley fall-run reporting unit, combined with the fact that the spring run is typically at much lower abundance than is the fall run. Likewise, the genetic similarity of fish from the Mid Oregon Coast reporting unit and the Northern California/Southern Oregon Coast reporting unit made it difficult to accurately estimate mixing proportions for these reporting units; however, the estimates were still quite good and largely unbiased. Therefore, despite the enlarged quantile intervals for Central Valley spring-run and the Mid Oregon Coast reporting units versus Northern California–Southern Oregon reporting unit, the results from both simulation methods indicated that the SNP baseline is capable of providing estimates of the true mixing proportions for most reporting units that are nearly as accurate as one would expect given perfect identification of each fish.

Fishery samples

Of the 2090 samples collected from California fisheries in 2010, 85 samples were excluded because they did not yield acceptable genotypes (<60 successfully genotyped loci) and 2 samples were excluded because they were duplicates of 2 other samples in the data set. Eight fish exceeded the iHz threshold of 0.56 and were removed because of potential contamination. Seven fish were identified as Coho Salmon through both GSI assignment and with the species-diagnostic assay. Another 18 samples did not meet assignment confidence criteria (mean z-score of –3.99 and a mean of 75 successfully genotyped loci) and were also excluded. For the remaining 1969 fish, assignment probabilities to reporting unit ranged from 36.4% to 100% (mean 98.5%) and z-scores ranged from –4.12 to 2.68 (mean –0.04). Central Valley fall-run fish dominated the stock composition, accounting for more than 80% of sampled fish, followed by the Rogue River (7.79%) and Klamath River (5.46%) reporting units and 8 other stocks with <5% (Table 3). Of the assigned fish, 1052 retained CWTs that were recovered. Genetic assignment to reporting unit disagreed with CWT origin for only 11 fish (1.05%), and, of these mismatches, 6 were fish with Klamath or Smith River tags that were assigned to the genetically similar Rogue River reporting unit.

Discussion

Here we describe one of the first large-scale SNP baselines for genetic stock identification of Chinook Salmon and the first designed for use with fisheries in the California Current large marine ecosystem off the contiguous United States. Chinook Salmon are an economically and ecologically important species and are a major component of fisheries in the North Pacific Ocean. We genotyped more than 8000 individual fish from 69 distinct populations at 96 SNP loci to construct the baseline. The reporting units included in the baseline represent the likely sources for more than 99% of the fish typically encountered in PFMC fisheries off California and Oregon.

Furthermore, results from mixture analyses and self-assignment indicate that the baseline has near maximum possible power for discrimination of Chinook Salmon stocks at the reporting unit level. Estimates of

the mixture proportions of Central Valley fall, Central Valley winter, California Coast, Klamath River, and Rogue River reporting units (Fig. 2) were no more variable than were estimates that would have been obtained if every fish had carried an unambiguous reporting-unit tag. Estimates of mixing proportions for Central Valley spring, North California/South Oregon, and Mid Oregon Coast reporting units were somewhat more variable but appeared to be nearly unbiased. In the ocean fishery sample, assignments of more than 1000 individuals to reporting unit, determined with our baseline, were highly concordant (98.95%) with the CWTs recovered from the same fish. This SNP baseline, therefore, represents an important addition to the technologies available to managers and researchers.

Methodological considerations

Management of salmon fisheries in the Pacific Ocean off North America can be roughly divided into 3 fisheries by region: California and Oregon fisheries, managed by the Pacific Fishery Management Council (PFMC); Washington, British Columbia, Canada, and southeastern Alaska fisheries, subject to the international Pacific Salmon Treaty, reporting to and regulated by the Pacific Salmon Commission; and fisheries farther north and west in Alaska that are managed by the state, with salmon bycatch under the purview of the North Pacific Fishery Management Council. The genetic baseline described here was designed primarily to identify fish caught in PFMC ocean fisheries and in ecological investigations in the southern portion of the California Current ecosystem and its associated tributary rivers and streams. We have shown that it performs well in this area but, because of an ascertainment strategy during SNP discovery that included individuals from the Columbia River and British Columbia (Clemento et al., 2011), the baseline also has sufficient statistical power to identify the source of some fish from elsewhere in the North American range of this species.

We observed high rates of self-assignment to reporting unit for all regions represented in the baseline, although some reporting units clearly were composed of populations with minimal differentiation from each other. Moreover, the utility of our baseline could be extended effectively by simply genotyping the same panel of SNPs on additional populations in those regions, despite the reduced heterozygosity and mean number of alleles (Table 1), and presumably statistical power in our baseline, for populations from Canada and Alaska.

Other SNP baselines for Chinook Salmon also have been described or are being constructed. Templin et al. (2011) described a 45 SNP locus baseline for populations in the northern and western parts of the Chinook Salmon range, designed primarily for GSI of populations from western and southcentral Alaska. This same baseline was used also to probe the seasonal distribution and migration pattern of Chinook Salmon in the Bering Sea and North Pacific Ocean (Larson et al., 2013). Despite the presence of 14 populations from California, Oregon, and Washington in that baseline, Larson et al. (2013) appropriately emphasized that resolution of those southern populations is sufficient only for broad-scale assignments. Similarly, Warheit et al.[3] described the marker selection for eventual development of a SNP baseline for application to fisheries managed by the Pacific Salmon Commission.

Although the existence of multiple regional baselines is likely to expand, it still will benefit the entire community of fishery managers and scientists to carefully design marker panels with as much overlap as possible. It is conceivable that 2 or 3 panels of 96 SNPs could provide the level of resolution needed for identification throughout the range of Chinook Salmon. Alternatively, as next-generation sequencing techniques mature, genotyping-by-sequencing (GBS) approaches might yield data for GSI at a lower cost than that with current genotyping techniques. A GBS approach could be used to simultaneously genotype all of the SNPs in each of the regional baselines, allowing mixed-stock analysis throughout the range of this species.

Inclusion of the species-diagnostic marker and Coho Salmon sample in the baseline provided insight into the prevalence of misidentification of Coho Salmon in ocean fisheries. In the 2010 fishery off California, 7 fish sampled as Chinook Salmon were found to be Coho Salmon. Without methods to identify Coho Salmon, the baseline would assign them with erroneously high confidence to a northern, low-heterozygosity Chinook Salmon population (data not shown). This problem is characteristic of most statistical methods for performing GSI: if an individual's true population of origin is *not* included in the baseline, then even if *all* the populations in the baseline are very poor candidates for that fish's origin, that fish might still be assigned with high posterior probability to one of the populations. This situation occurs when one population is much more likely to be the population of origin, than any of the other incorrect populations, even if it is not a likely origin for that individual on an absolute scale.

We introduced a simulation-based z-score method, implemented in gsi_sim, to identify fish that likely have not originated from populations in the baseline. An alternative, Bayesian nonparametric approach to dealing with fish from populations not in the baseline identifies those fish and estimates the allele frequencies in their (unrepresented) source population (Pella and Masuda, 2000). That approach is appropriate particularly when large numbers of fish are sampled from each of the populations that are not included in the baseline and when the unrepresented populations are quite divergent from all of the populations in the baseline.

We chose the *z*-score approach over the Bayesian nonparametric approach for 3 main reasons: 1) it is computationally fast and simple (there are no convergence problems that might be difficult to detect); 2) our baseline was sufficiently comprehensive for stocks that

contribute to PFMC fisheries, and therefore it was unlikely that large numbers of fish would originate from any single unrepresented population, let alone a highly divergent one; and 3) our approach is more appropriate for identification of fish whose genotypes are aberrant because of genotyping complications or sample contamination. Regardless of which method is used, all GSI estimation should include some analysis to identify fish that are either from populations not included in the baseline or that have aberrant genotypes for another reason.

GSI is highly dependent on source populations being genetically differentiated enough from one another for discrimination. In situations where hatchery broodstock transfers, supplementation, or other processes increase straying and gene flow between fish populations, genetic differentiation decreases and it can become more difficult to use GSI. Such is the case in the Central Valley of California, where average F_{ST} between populations in the fall-run reporting unit was 0.006 and in the spring-run reporting unit was 0.013. In the dendrogram (Fig. 1), this region was characterized by extremely short branch lengths, small internodal differences, and weak bootstrap support. Extensive straying of hatchery salmon due to off-site juvenile releases (California Hatchery Scientific Review Group [9]) and water operations (Fisher, 1994) has eliminated historical differentiation between populations of fall-run Chinook Salmon (Williamson and May, 2005). Introgression between fall-run and spring-run fish at the Feather River Hatchery, and likely elsewhere within the basin, has reduced differentiation between these 2 phenotypes, with mean F_{ST} of 0.025 between fall-run and naturally spawning spring-run populations.

Sampling of different stocks for baseline construction in the presence of high stray rates is not entirely straightforward, particularly when populations are largely sympatric and not visually distinguishable. For example, there is clearly a single fish from the Central Valley fall-run reporting unit that was sampled as a winter-run fish in our baseline. These types of occurrences are almost inevitable given the high degree of disturbance and hatchery supplementation over much of the range of Chinook Salmon. One approach is to move fish with discrepant genotypes from the baseline populations in which they were sampled to the ones to which they are assigned with GSI (e.g., Banks et al., 2000). However, such a procedure can introduce an upward bias in the predicted accuracy of the baseline, if, in fact, the removed fish actually do belong to the populations from which they were sampled and simply have unlikely genotypes at the genetic markers used for baseline construction. We chose to be conservative by both 1) accepting a slightly lower rate of predicted resolution obtained by not removing miscategorized fish and 2) avoiding an upward bias in predicted GSI accuracy if the fish removed are not miscategorized.

[9] California Hatchery Scientific Review Group. 2012. California Hatchery Review Report, 102 p. Prepared for the U.S. Fish and Wildlife Service and Pacific States Marine Faisheries Commission. [Available from http://swfsc.noaa.gov/publications/FED/01067.pdf and (appendices) http://cahatcheryreview.com/reports.]

Implications for management

Accurately estimating the proportion of fish from different populations in mixed-stock ocean fisheries has important applications for harvest management and conservation. Stocks that comingle in ocean fisheries can vary widely in productivity and abundance. Without precise information on their ocean distribution, as can be provided by GSI, managers have few options for protection of depressed or at-risk stocks from fishery impacts other than that of shutting down or curtailing fisheries over broad areas, as is currently done (Beacham et al., 2008). For example, in 2008 and 2009, the largest closures on record of fisheries in California and Oregon were enacted to protect the severely reduced Central Valley fall-run stock (Lindley et al., 2009). The economic effects of fishery closures are substantial, resulting in millions of dollars of lost income for fishermen, coastal communities, and retailers (Michael [10]).

Management of Chinook Salmon in California, Oregon, and Washington and in fisheries managed by the Pacific Salmon Commission depends heavily on information generated by an elaborate CWT program (Hankin et al., 2005). Tiny wire tags are mechanically implanted into the heads of juvenile fish, with each tag bearing a code that identifies the release group and source hatchery (or stock) of that fish. Tagging of naturally spawned juvenile fish has generally proven unsuccessful (Beacham et al., 1996), and, for that reason, tagged hatchery stocks are used as proxies to estimate fishery impacts for groups of natural stocks. Aside from the largely unvalidated assumption that such proxies accurately reflect fishery impacts on associated natural stocks (Hankin et al., 2005), the physical effects of tagging fish and removing their nerve-rich adipose fin (Buckland-Nicks et al., 2012) as an associated external mark can increase disease transmission (Elliott and Pascho, 2001), interfere with homing (Morrison and Zajac, 1987; Habicht et al., 1998) and swimming ability (Reimchen and Temple, 2004) and may affect size-at-return for adult salmon (Vander Haegen et al., 2005). Moreover, extremely low recovery rates mean that CWT data are often quite limited and great uncertainty is frequently associated with the estimates derived from them (Hankin et al., 2005).

GSI has been advanced as an alternative to CWTs in fishery management for several decades. Our direct

[10] Michael, J. 2010. Employment impacts of California salmon fishery closures in 2008 and 2009. Business Forecasting Center, Univ. of the Pacific, Stockton, CA. [Available from http://forecast.pacific.edu/BFC%20salmon%20jobs.pdf.]

comparison of CWT with genetic assignments demonstrates that our baseline is capable of identifying fish to reporting unit with accuracy comparable to that of CWTs. Furthermore, the use of GSI can identify considerably more fish to reporting unit, including fish from natural stocks. Confident genetic assignments were obtained for ~94% of fish from the 2010 fishery sample, but only 1052 of those fish carried CWTs and this number is inflated partially because of oversampling of fish believed to carry CWTs.

Fishery management decisions rely heavily on cohort-based ocean harvest models (cf., O'Farrell et al., 2012), which require information on both stock of origin and age of fish impacted by fisheries. Because GSI does not provide the age of individuals, it is not by itself an adequate alternative to CWTs. Nonetheless, new statistical methods capable of integrating GSI, length data, and scale- or otolith-based age data have been developed recently, allowing managers to draw important inference about PFMC fisheries that are not possible with CWTs alone (Satterthwaite et al., 2014). Moreover, pedigree-based genetic tagging *does* supply age for salmon (Anderson and Garza, 2006; Garza and Anderson[2]). This method, termed "parentage-based tagging" (PBT), can identify the actual parents of a genotyped individual through parentage analysis if they have been genotyped with the same genetic markers. If the parents' date of spawning is known, as it typically is in a hatchery, then the reconstructed pedigrees yield the offspring's precise age and any associated parental spawning information.

Importantly, both PBT and GSI can be undertaken with the same SNP genotypes, and the SNPs used in our GSI baseline are sufficiently powerful for PBT with Chinook Salmon from California to Washington (Anderson, 2012). This interoperability of genotype data enables an integrated program that uses both GSI and PBT simultaneously, providing identification for all fish in a fishery or ecological sample and yielding significantly greater inference than either method alone. For example, GSI cannot distinguish between spring-run and fall-run fish from the Feather River Hatchery in California, but PBT distinguishes them, almost without error, from any mixture. Likewise, although it is difficult to implement PBT in natural populations, the same SNP genotypes used in a PBT analysis permit accurate identification (by GSI) of fish from the naturally spawning, ESA-listed "California Coastal Chinook Salmon Evolutionarily Significant Unit."

Conclusions

The advent of high-throughput SNP genotyping already has revolutionized human genetics (Jenkins and Gibson, 2002), providing previously unattainable resolution (e.g., Novembre et al., 2008) and is poised to do the same for fisheries biology and management. As described here, we used a careful and statistically valid power analysis of SNP genotypes from a large number of Chinook Salmon populations concentrated at the southern end of the native range of this species to show that SNPs can provide a powerful baseline for genetic stock identification (see also Larson et al., 2013) in fisheries and ecological investigation in the California Current large marine ecosystem and its tributaries in California and Oregon. We predict that these advances in genetic resources and methods will foster fundamental improvements in the way salmon populations are studied, monitored, and managed.

Acknowledgments

The authors would like to thank the entire Molecular Ecology and Genetic Analysis Team in the Fisheries Ecology Division of the Southwest Fisheries Science Center (SWFSC) for their invaluable assistance with genotyping and analyses. Of critical importance to the successful completion of this project were the baseline samples provided to us by the California Department of Fish and Game (now Wildlife; S. Harris), Hoopa Valley Tribal Fisheries Department (G. Kautsky), Oregon Department of Fish and Wildlife, Oregon State University Department of Fisheries and Wildlife (M. Banks), Idaho Department of Fish and Game (M. Campbell), Columbia River Inter-Tribal Fish Commission (S. Narum), NOAA Northwest Fisheries Science Center (P. Moran), U.S. Fish and Wildlife Service (M. Brown, D. Hawkins, and C. Smith), Washington Department of Fish and Wildlife (S. Blankenship and K. Warheit), University of Washington School of Aquatic and Fishery Science (L. Seeb), Department of Fisheries and Oceans, Canada (T. Beacham), and Alaska Department of Fish and Game (W. Templin). Fishery samples were collected by the California Department of Fish and Game (now Wildlife) and provided to us by M. Heisdorf and M. Palmer-Zwahlen. We also thank T. Beacham and 2 anonymous referees for comments that improved the manuscript. This project received funding from NOAA's Cooperative Fisheries Research Program and the SWFSC. A. Clemento also received support from a California Bay Delta Science Fellowship and the University of California Coastal Environmental Quality Initiative. Many of the baseline samples were collected and DNA extracted with funds from the Pacific Salmon Commission.

Literature cited

Abadía-Cardoso, A., A. J. Clemento, and J. C. Garza.
2011. Discovery and characterization of single nucleotide polymorphisms in steelhead/rainbow trout, *Oncorhynchus mykiss*. Mol. Ecol. Resour. 11 (suppl. s1):31–49.

Abadía-Cardoso, A., E. C. Anderson, D. E. Pearse, and J. C. Garza
2013. Large-scale parentage analysis reveals reproductive patterns and heritability of spawn timing in

a hatchery population of steelhead (*Oncorhynchus mykiss*). Mol. Ecol. 22:4733–4746.

Aguilar, A., and J. C. Garza.
2008. Isolation of 15 single nucleotide polymorphisms from coastal steelhead, *Oncorhynchus mykiss* (Salmonidae). Mol. Ecol. Resour. 8:659–662.

Allendorf, F., and L. W. Seeb.
2000. Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. Evolution 54:640–51.

Anderson, E. C.
2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Mol. Ecol. Resour. 10:701–710.
2012. Large-scale parentage inference with SNPs: an efficient algorithm for statistical confidence of parent pair allocations. Stat. Appl. Genet. Mol. Biol. 11:12. doi: 10.1515/1544-6115.1833

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

Anderson, E. C., R. S. Waples, and S. T. Kalinowski.
2008. An improved method for predicting the accuracy of genetic stock identification. Can. J. Fish. Aquat. Sci. 65:1475–1486.

Banks, M. A., V. K. Rashbrook, M. Calavetta, C. Dean, and D. Hedgecock.
2000. Analysis of microsatellite DNA resolves genetic structure and diversity of Chinook salmon (*Oncorhynchus tshawytscha*) in California's Central Valley. Can. J. Fish. Aquat. Sci. 57:915–927.

Barnett-Johnson, R., T. E. Pearson, F. C. Ramos, C. Grimes, and R. B. MacFarlane.
2008. Tracking natal origins of salmon using isotopes, otoliths, and landscape geology. Limnol. Oceanogr. 53:1633–1642.

Beacham, T. D., R. E. Withler, and T. Stevens.
1996. Stock identification of chinook salmon (*Oncorhynchus tshawytscha*) using minisatellite DNA variation. Can. J. Fish. Aquat. Sci. 53:380–394.

Beacham, T. D., J. R. Candy, K. L. Jonsen, J. Supernault, M. Wetklo, L. Deng, K. M. Miller, R. E. Withler, and N. Varnavskaya.
2006. Estimation of stock composition and individual identification of Chinook salmon across the Pacific Rim by use of microsatellite variation. Trans. Am. Fish. Soc. 135:861–888.

Beacham, T. D., I. Winther, K. L. Jonsen, M. Wetklo, L. Deng, and J. R. Candy.
2008. The application of rapid microsatellite-based stock identification to management of a Chinook Salmon troll fishery off the Queen Charlotte Islands, British Columbia. N. Am. J. Fish. Manage. 28:849–855.

Bernatchez, L., and P. Duchesne.
2000. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? Can. J. Fish. Aquat. Sci. 57:1–12.

Boyce, N. P., Z. Kabata, and L. Margolis.
1985. Investigations of the distribution, detection, and biology of *Henneguya salminicola* (Protozoa, Myxozoa), a parasite of the flesh of Pacific salmon. Can. Tech. Rep. Fish. Aquat. Sci. 1405, 55 p.

Buckland-Nicks, J. A., M. Gillis, and T. E. Reimchen.
2012. Neural network detected in a presumed vestigial trait: ultrastructure of the salmonid adipose fin. Proc. R. Soc. Lond., Ser. B: Biol. Sci. 279:553–563

Campbell, N., and S. R. Narum.
2008. Identification of novel single-nucleotide polymorphisms in Chinook salmon and variation among life history types. Trans. Am. Fish. Soc. 137:96–106.

Cavalli-Sforza, L. L., and A. W. F. Edwards.
1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 19:233–257.

Claytor, R., and H. MacCrimmon.
1988. Morphometric and meristic variability among North American Atlantic salmon (*Salmo salar*). Can. J. Zool. 66:310–317.

Clemento, A. J., A. Abadía-Cardoso, H. A. Starks, and J. C. Garza.
2011. Discovery and characterization of single nucleotide polymorphisms in Chinook salmon, *Oncorhynchus tshawytscha*. Mol. Ecol. Resour. 11 (suppl. s1):50–66.

Cook, R. C.
1982. Stock identification of sockeye salmon (*Oncorhynchus nerka*) with scale pattern recognition. Can. J. Fish. Aquat. Sci. 39:611–617.

Cornuet, J., S. Piry, G. Luikart, A. Estoup, and M. Solignac.
1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153:1989–2000.

Cronin, M. A., W. J. Spearman, and R. L. Wilmot.
1993. Mitochondrial DNA variation in Chinook (*Oncorhynchus tshawytscha*) and chum salmon (*O. keta*) detected by restriction enzyme analysis of polymerase chain reaction (PCR) products. Can. J. Fish. Aquat. Sci. 50:708–715.

Elliott, D. G., and R. J. Pascho.
2001. Evidence that coded-wire-tagging procedures can enhance transmission of *Renibacterium salmoninarum* in Chinook salmon. J. Aquat. Anim. Health 13:181–193.

Federal Register.
1990. Endangered and threatened wildlife and plants; listing of the Sacramento River winter-run Chinook salmon as threatened. Vol. 55(231):4962. GPO, Washington, DC.
1999. Endangered and threatened species; threatened status for three Chinook Salmon Evolutionarily Significant Units (ESUs) in Washington and Oregon, and endangered status for one Chinook Salmon ESU in Washington. Vol. 64(56):14308–14328. GPO, Washington, DC.

Fisher, F. W.
1994. Past and present status of Central Valley Chinook salmon. Conserv. Biol. 8:870–873.

Flannery, B. G., J. K. Wenburg, and A. J. Gharrett.
2007. Variation of amplified fragment length polymorphisms in Yukon River chum salmon: population structure and application to mixed-stock analysis. Trans. Am. Fish. Soc. 136:911–925.

Fournier, D. A., T. D. Beacham, B. E. Riddell, and C. A. Busack.
1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. 41:400–408.

Habicht, C., S. Sharr, D. Evans, and J. E. Seeb.
1998. Coded wire tag placement affects homing ability of pink salmon. Trans. Am. Fish. Soc. 127:652–657.

Hankin, D. G., J. H. Clark, R. B. Deriso, J. C. Garza, G. S. Morishima, B. E. Riddell, C. Schwarz, and J. B. Scott.
2005. Report of the expert panel on the future of the coded wire tag recovery program for Pacific salmon. Pacific Salmon Commission Tech. Rep. 18, 230 p. [Available from http://www.psc.org/pubs/psctr18.pdf.]

Hedrick, P. W., D. Hedgecock, and S. Hamelberg.
1995. Effective population size in winter-run Chinook salmon. Conserv. Biol. 9:615–624.

Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp.
2007. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460. doi: 10.1186/1471-2105-8-460

Jenkins, S., and N. Gibson.
2002. High-throughput SNP genotyping. Comp. Funct. Genomics 3:57–66.

Kalinowski, S. T.
2004. Genetic polymorphism and mixed-stock fisheries analysis. Can. J. Fish. Aquat. Sci. 61:1075–1082.

Larson, W. A., F. M. Utter, K. W. Myers, W. D. Templin, J. E. Seeb, C. M. Guthrie III, A. V. Bugaev, and L. W. Seeb.
2013. Single-nucleotide polymorphisms reveal distribution and migration of Chinook salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean. Can. J. Fish. Aquat. Sci. 70:128–141.

Lindley, S. T., C. B. Grimes, M. S. Mohr, W. Peterson, J. Stein, J. T. Anderson, L. W. Botsford, D. L. Bottom, C. A. Busack, T. K. Collier, J. Ferguson, J. C. Garza, A. M. Grover, D. G. Hankin, R. G. Kope, P. W. Lawson, A. Low, R. B. MacFarlane, K. Moore, M. Palmer-Zwahlen, F. B. Schwing, J. Smith, C. Tracy, R. Webb, B. K. Wells, and T. H. Williams.
2009. What caused the Sacramento River fall Chinook stock collapse? NOAA Tech. Memo. NMFS-SWFSC-447, 121 p.

Louis, E. J., and E. R. Dempster.
1987. An exact test for Hardy-Weinberg and multiple alleles. Biometrics 43:805–811.

Miller, K. M., R. E. Withler, and T. D. Beacham.
1996. Stock identification of coho salmon (*Oncorhynchus kisutch*) using minisatellite DNA variation. Can. J. Fish. Aquat. Sci. 53:181–195.

Milner, G. B., D. J. Teel, F. M. Utter, and G. A. Winans.
1985. A genetic method of stock identification in mixed populations of Pacific salmon, *Oncorhynchus* spp. Mar. Fish. Rev. 47:1–8.

Moran, P., D. J. Teel, M. A. Banks, T. D. Beacham, M. R. Bellinger, S. M. Blankenship, J. R. Candy, J. C. Garza, J. E. Hess, S. R. Narum, L. W. Seeb, W. D. Templin, C. G. Wallace, and C. T. Smith.
2013. Divergent life-history races do not represent Chinook salmon coast-wide: the importance of scale in Quaternary biogeography. Can. J. Fish. Aquat. Sci. 70: 415–435.

Morrison, J., and D. Zajac.
1987. Histologic effect of coded wire tagging in chum salmon. N. Am. J. Fish. Manage. 7:439–441.

Narum, S. R., M. A. Banks, T. D. Beacham, M. R. Bellinger, M. R. Campbell, J. Dekoning, A. Elz, C. M. Guthrie, C. Kozfkay, K. M. Miller, P. Moran, R. Phillips, L.W. Seeb, C. T. Smith, K. Warheit, S. F. Young, and J. C. Garza.
2008. Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. Mol. Ecol. 17:3464–3477.

Nei, M.
1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590.

Nosil, P., D. Funk, and D. Ortiz-Barrientos.
2009. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18:375–402.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. Boyko, A. Auton, A. Indap, K. King, S. Bergmann, M. Nelson, M. Stephens, and C. D. Bustamante.
2008. Genes mirror geography within Europe. Nature 456:98–101.

O'Farrell, M. R., M. S. Mohr, A. M. Grover, and W. H. Satterthwaite.
2012. Sacramento River winter Chinook cohort reconstruction: analysis of ocean fishery impacts. NOAA Tech. Memo. NOAA-TM-NMFS-SWFSC-491, 74 p.

Pella, J., and M. Masuda.
2000. Bayesian methods for analysis of stock mixtures from genetic characters. Fish. Bull. 99:151–167.

Rannala, B., and J. L. Mountain.
1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94:9197–9201.

Reimchen, T. E., and N. F. Temple.
2004. Hydrodynamic and phylogenetic aspects of the adipose fin in fishes. Can. J. Zool. 82:910–916

Rousset, F.
2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. Mol. Ecol. Resour. 8:103–106.

Satterthwaite, W., M. S. Mohr, M. R. O'Farrell, E. C. Anderson, M. A. Banks, S. J. Bates, M. R. Bellinger, L. A. Borgerson, E. D. Crandall, J. C. Garza, B. J. Kormos, P. W. Lawson, and M. L. Palmer-Zwahlen.
2014. Use of genetic stock identification data for comparison of the ocean spatial distribution, size-at-age, and fishery exposure of an untagged stock and its indicator: California Coastal versus Klamath River Chinook. Trans. Am. Fish. Soc. 143:117–133.

Seeb, L. W., A. Antonovich, M. A. Banks, T. D. Beacham, M. R. Bellinger, S. M. Blankenship, M. R. Campbell, N. A. Decovich, J. C. Garza, C. M. Guthrie III, T. A. Lundrigan, P. Moran, S. R. Narum, J. J. Stephenson, K. J. Supernault, D. J. Teel, W. D. Templin, J. K. Wenburg, S. F. Young, and C. T. Smith.
2007. Development of a standardized DNA database for Chinook salmon. Fisheries 32:540–552.

Seeb, J. E., G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L. W. Seeb.
2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11 (suppl. s1):1–8.

Shaklee, J. B., and S. R. Phelps.
1990. Operation of a large-scale, multi-agency program for genetic stock identification. Am. Fish. Soc. Symp. 7:817–830.

Smith, C. T., A. Antonovich, W. D. Templin, C. M. Elfstrom, S. R. Narum, and L. W. Seeb.
2007. Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Trans. Am. Fish. Soc. 136:1674–1687.

Smith, C. T., C. M. Elfstrom, L. W. Seeb, and J. E. Seeb.
2005a. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol. Ecol. 14:4193–4203.
Smith, C. T., W. D. Templin, J. E. Seeb, and L. W. Seeb.
2005b. Use of the 5′-nuclease reaction for single nucleotide polymorphism genotyping in Chinook salmon. Trans. Am. Fish. Soc. 134:207–217.
Smith, M. J., C. E. Pascal, Z. Grauvogel, C. Habicht, J. E. Seeb, and L. W. Seeb.
2011. Multiplex preamplification PCR and microsatellite validation enables accurate single nucleotide polymorphism genotyping of historical fish scales. Mol. Ecol. Resour. 11 (suppl. s1):268–277.
Smouse, P. E., R. S. Waples, and J. A. Tworek.
1990. A genetic mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. 47:620–634.
Taylor, E. B.
1991. A review of local adaptation in Salmonidae, with particular reference to Pacific and Atlantic salmon. Aquaculture 98:185–207.
Templin, W. D., J. E. Seeb, J. R. Jasper, A. W. Barclay, and L. W. Seeb.
2011. Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. Mol. Ecol. Resour. 11 (suppl. s1):226–246.
Tessier, N., L. Bernatchez, P. Presa, and B. Angers.
1995. Gene diversity analysis of mitochondrial DNA, microsatellites and allozymes in landlocked Atlantic salmon. J. Fish Biol. 47:156–163.
Utter, F. M., G. B. Milner, G. Stahl, and D. J. Teel.
1989. Genetic population structure of Chinook salmon, *Oncorhynchus tshawytscha*, in the Pacific Northwest. Fish. Bull. 87:239–264.
Vander Haegen, G. E., H. L. Blankenship, A. Hoffmann, and D. A. Thompson.
2005. The effects of adipose fin clipping and coded wire tagging on the survival and growth of spring Chinook salmon. N. Am. J. Fish. Manage. 25:1161–1170.
Williamson, K. S., and B. May
2005. Homogenization of fall-run Chinook salmon gene pools in the Central Valley of California, USA. N. Am. J. Fish. Manage. 25:993–1009.
Waples, R. S., D. J. Teel, J. M. Myers, and A. R. Marshall
2004. Life-history divergence in Chinook salmon: historic contingency and parallel evolution. Evolution 58:386–403.
Weir, B. S., and C. C. Cockerham.
1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.
Wood, C. C., S. McKinnell, T. Mulligan, and D. Fournier.
1987. Stock Identification with the maximum-likelihood mixture model: sensitivity analysis and application to complex problems. Can. J. Fish. Aquat. Sci. 44:866–881.