Canadian Science Publishing
NRC Research Press journals | Publishing Services

Canadian Journal of Fisheries and Aquatic Sciences
Journal canadien des sciences halieutiques et aquatiques

Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification

Journal:	*Canadian Journal of Fisheries and Aquatic Sciences*
Manuscript ID	cjfas-2016-0443.R1
Manuscript Type:	Rapid Communication
Date Submitted by the Author:	07-Jan-2017
Complete List of Authors:	McKinney, Garrett; University of Washington, School of Aquatic and Fishery Sciences Seeb, James; University of Washington, School of Aquatic and Fishery Sciences Seeb, Lisa; University of Washington, School of Aquatic and Fishery Sciences
Keyword:	Genetic stock identification, mixture analysis, haplotype, RADseq, Chinook salmon

Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification

Garrett J. McKinney[1*], James E. Seeb[1], Lisa W. Seeb[1]

[1] School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle WA 98195-5020, USA. (email: gjmckinn@uw.edu, jseeb@uw.edu, lseeb@uw.edu)

[*] **Corresponding author:** Garrett J. McKinney (email: gjmckinn@uw.edu, phone: 1-765-430-3272)

Abstract

A common challenge for fishery applications is resolving the relative contribution of closely related populations where accuracy of genetic assignment may be limited. An overlooked method for increasing assignment accuracy is the use of multi-SNP haplotypes rather than single-SNP genotypes. Haplotypes increase power for detecting population structure, and loci derived from next-generation sequencing methods often contain multiple SNPs. We evaluated the utility of multi-SNP haplotyping for mixture analysis in Western Alaska Chinook salmon. Multi-SNP haplotype data increased the accuracy of mixture analysis for closely related populations by up to seven percentage points relative to single-SNP genotype data for a set of 500 loci; 90% accuracy was achievable with as few as 150 loci with multi-SNP haplotypes but required at least 300 loci with single-SNP genotypes. Individual assignment to reporting groups showed an even greater increase in accuracy of up to 17 percentage points when multi-SNP haplotypes were used. Haplotyping multiple SNPs shows promise to improve the accuracy of assigning unknown fish to population of origin whenever haplotype data are available.

Introduction

Assignment methods that use genetic data to estimate the origins of unknown mixtures of fish are a powerful tool to help identify and shape the composition of the harvest in mixed-stock fisheries (Beacham et al. 2008; Dann et al. 2013; Ensing et al. 2013). Practitioners place a high priority on methodological enhancements to improve resolution, especially when populations contributing to the fishery are closely related. Vast efforts are undertaken to achieve even small improvements in the resolution of stocks of migratory species when diverse social interests collide during the management of charismatic species in decline (e.g., Larson et al. 2014a; Abadia-Cardoso et al. 2016). Substantial efforts center on screening additional spawning populations and the use of additional DNA markers scored as single nucleotide polymorphisms (SNPs) to enhance both assignment efficiency and accuracy (reviewed in Seeb et al. 2011).

Three common approaches to genetic assignment are parentage analysis, individual assignment, and mixture analysis. These methods differ in scope of assignment: parentage analysis assigns offspring to parents, individual assignment assigns individuals to populations or reporting groups, and mixture analysis estimates relative composition of mixed fisheries samples (reviewed in Manel et al. 2005). Collectively, these methods have been used to study questions such as composition of fishery catch (Dann et al. 2013; Bradbury et al. 2016), reproductive success (Ford et al. 2015), success of reintroduction programs (Evans et al. 2016), monitoring of hatchery programs (Steele et al. 2013), local adaptation (O'Malley et al. 2007), and migration patterns (Larson et al. 2013).

Each method has different advantages; however, mixture analysis is ideally suited for many practical applications in fisheries management. Mixture analysis does not require the extensive sampling of parents (required for parental assignment) and is more accurate than individual assignment when differentiation between populations is low or when individual assignments cannot be made with confidence (Manel et al. 2005).

A common challenge for fishery applications is resolving the relative contribution of closely related populations. Methods for increasing assignment accuracy include using putatively adaptive loci (see Ackerman et al. 2011; Araneda et al. 2016), increasing the number of loci (Larson et al. 2014b), and increasing the number of individuals and populations sampled (Beacham et al. 2010; Habicht et al. 2010). An overlooked enhancement for increasing assignment accuracy is the use of multi-SNP haplotypes rather than single-SNP genotypes. Haplotyping has been demonstrated to increase power for detecting population structure, where haplotypes are either blocks of contiguous genomic sequence (Fariello et al. 2013) or are constructed from multiple SNPs in linkage disequilibrium (Gattepaille and Jakobsson 2012; Leslie et al. 2015). To date, assignment studies have typically restricted analyses to a single SNP per locus (Larson et al. 2014a; Araneda et al. 2016). Rarely, multiple linked SNPs have been genotyped in two separate assays and then combined as a composite phenotype (e.g., MHC phenotypes, Habicht et al. 2010). Discarding the data contained in multi-SNP haplotypes results in a loss of information.

Next-generation sequencing technologies can genotype multiple SNPs per locus by directly sequencing the entire DNA template in contrast to many other assays that only genotype a single SNP. Haplotype data is naturally generated by next-generation sequencing platforms; however, most studies use bioinformatic filtering to examine only a single SNP per locus. Haplotype analysis involves no extra cost or effort over single-SNP genotyping but rather a simple change in bioinformatic procedures.

In this paper we evaluate the potential for increasing accuracy of genetic stock identification (GSI) by using haplotype data derived from restriction-site associated DNA sequencing (RADseq; Baird et al. 2008). RADseq has become a common method for developing thousands of SNP loci that can be screened for population assignment (Hohenlohe et al. 2011; Andrews et al. 2016). The recent development of amplicon sequencing techniques such as GTseq (Campbell et al. 2015) or RAD Capture (Rapture) (Ali et al. 2016) facilitates large scale cost-effective and rapid genotyping of loci identified with RADseq, and individual RADseq loci often contain multiple SNPs which allow haplotype analysis. RADseq data promise to provide many new loci, including adaptively important loci, for analyses of commercially important species such as Chinook salmon (*Oncorhynchus tshawytscha*) (Larson et al. 2014a; McKinney et al. 2016).

Chinook salmon is not only commercially important but also has sport, subsistence, and ceremonial importance in Pacific North America and Asia (e.g., Wolfe and Spaeder 2009). Populations of Chinook salmon show precipitous declines in many parts of their range confounding managers who must balance competing interests (Gustafson et al. 2007; Gisclair 2009; ADF&G 2013). Managing for sustainability on the population level is highly desirable, but discrete populations of Chinook salmon are often closely related, challenging the use of DNA datasets for analysis of population mixtures (Templin et al. 2011; Larson et al. 2013).

We re-examined a previous RADseq dataset by Larson et al. (2014b) to determine if haplotype analysis can increase power for mixture analysis and individual assignment over that of single-SNP genotype data. We found that haplotype data increased the accuracy of mixture analysis for closely related populations by up to seven percentage points and individual assignment up to 17 percentage points relative to single-SNP genotype data for a set of 500 loci. These results suggest that haplotyping is an efficient method to improve accuracy of genetic stock identification when multi-SNP data are available.

Methods

Raw sequence data from Larson *et al.* (2014b) were downloaded from Dryad (doi:10.5061/dryad.rs4v1) and reprocessed using *STACKS* (Catchen et al. 2013). This dataset consisted of 250 Chinook salmon from five populations from Western Alaska: Big Salmon (n=47), Tubutulik (n=56), Anvik (n=47), Kogrukluk (n=44) and Koktuli (n=56). Default settings were used for all *STACKS* modules with the following exceptions: *process_radtags* (-c –r –q –filter_illumina –t 94), *ustacks* (-m 2 –M 2, --model_type bounded --bound_high 0.05), *cstacks* (-n 2). Loci (individual RADtags) were retained if the genotype rate was at least 80% and if the minor allele frequency was at least 0.05 in one or more populations following the methods of Larson *et al.* (2014b). The *STACKS* catalog for the Chinook salmon dataset was based on the catalog from McKinney *et al*. (2016) to ensure locus names were consistent among studies.

Two datasets were generated with the loci retained from *STACKS*; a single-SNP dataset and a multi-SNP (haplotype) dataset. The single-SNP dataset was constructed by retaining only a single SNP from loci that had multiple variant sites. For these loci, the SNP with the highest minor allele frequency across all populations was retained. For the haplotype dataset, all variant sites at a locus were retained, and each unique haplotype was considered an allele.

Samples from each population were split evenly into training and holdout sets for both the single-SNP and haplotype datasets; if populations could not be split exactly in half the remaining individual was assigned to the training dataset. Locus selection was conducted using only the single-SNP training set; both the single-SNP and haplotype holdout sets were used for mixture analysis. Separate training and holdout sets prevent an upward bias in assignment accuracy (Anderson 2010).

F_{ST} was estimated for each locus in the single-SNP training set using Genepop (Rousset 2008). Loci were ranked by F_{ST} and the 500 loci with highest F_{ST} were chosen for mixture analysis. Loci were chosen using the single-SNP dataset for two reasons: 1) to reflect the method that has been historically used to choose loci for GSI studies, and 2) to evaluate the increase in power of haplotypes for GSI even when haplotype information was not considered when initially choosing loci. The accuracy of single-SNP genotype vs haplotype data for mixture analysis was then assessed using 500 loci. Preliminary testing (not shown) suggested that multi-SNP haplotypes increased mixture analysis accuracy up to five percentage points over single-SNP genotypes using the 500 highest F_{ST} loci, approximately 1/3 of which contained multiple SNPs. We speculated that a panel where all loci contained multiple SNPs would further increase the accuracy of haplotyping relative to genotyping. The panel used in this study was obtained by filtering the full locus set to retain only loci with three or more haplotype alleles and taking the 500 highest F_{ST} ranked multiple SNP loci. Even though the F_{ST} was slightly reduced in the final panel relative to the preliminary panel, there was no difference in the mixture analysis results for single-SNP genotypes.

Mixture analysis was conducted with both the single-SNP genotypes and haplotypes to estimate assignment accuracy using the selected panel and procedures of Anderson (2010). Populations were first aggregated into reporting groups following the population structure previously identified by Larson *et al*. (2014b). The Big Salmon, Tubutulik River, and Anvik rivers were placed in their own reporting groups while the Kogrukluk and Koktuli rivers were placed into a combined reporting group (Kogrukluk/Koktuli).

Mixture analysis was run in GSIsim (Anderson et al. 2008) which implements the training holdout leave-one-out (THL) simulation method of Anderson (2010); simulation allows the accuracy of the panel to be estimated for a range of mixture proportions and allows an estimate of the variance in panel accuracy. Rather than directly assigning individuals in the holdout set to the baseline, GSIsim simulates genotypes for each individual in a mixture by drawing from population allele frequencies in the holdout dataset. Mixture proportions are estimated by comparing the simulated mixture individuals against the baseline of all individuals. The leave-one-out method is implemented at this point; an individual's alleles are excluded from the baseline when calculating the probability of the individual's genotype for their population of origin. Options in GSIsim were set as follows: leave-one-out yes, 1000 mixedFisherySamples, 400 mixedFisheryIndividuals.

GSIsim was run iteratively with one reporting group at a time as the focal group and the other reporting groups as the secondary groups. The true proportion of the focal reporting group in the mixture was simulated from 0 to 100% in 5% intervals. The remaining mixture proportion was divided into 5% increments and split equally among the secondary reporting groups. Any remaining proportion was randomly assigned to one of the secondary reporting groups. When multiple populations were contained within a reporting group, the mixture proportion for that reporting group was evenly divided among the populations. Reporting groups were simulated with a range of mixture proportions to determine accuracy through a range of stock compositions. Simulation results are reported as the mean mixture estimate ± the threshold for p=0.025 and p=0.925. This encompasses the 95% distribution of estimates.

Management agencies commonly evaluate reporting groups by using 100% simulations where 90% allocation is the minimum acceptable threshold (Seeb et al. 2000; Seeb et al. 2007; Beacham et al. 2012). Further panels were developed by removing loci from the 500 locus panel to determine how few loci could be used and still obtain 90% allocation at 100% simulation for mixture analysis with single-SNP genotype or haplotype data. Loci were removed based on F_{ST} rank, with the lowest F_{ST} loci removed first. Additional panels were tested with 400, 300, 200, 150, and 100 loci.

The output of GSIsim includes the assignment probability of each simulated individual to each population. To assign an individual to a reporting group, the probability of assignment to each population within the reporting group was summed. A threshold of 90% cumulative probability was required to assign an individual to a reporting group.

Results

A total of 14,494 loci passed quality filters. The majority of the loci (72%) contained a single SNP while the second most common category was 2 SNPs (24%) (Table 1). Loci with 2 SNPs had between 2 and 4 alleles: three alleles were most common (84%) followed by two alleles (14%) and four alleles (2%). Substantially more alleles with frequencies <0.10 were found for haplotypes rather than single-SNP genotypes (8,959 vs 5,647).

Accuracy of mixture analysis in 100% simulations varied by reporting group but was approximately 90% or greater for all cases. The Big Salmon River and Tubutulik River reporting groups each had 100% assignment with both the single-SNP genotype and the haplotype datasets (Table 2). This was expected given the high degree of differentiation between these reporting groups and the others (Larson et al. 2014b).

For the Anvik River and Kogrukluk/Koktuli reporting groups, the accuracy of mixture analysis was influenced both by the number of markers and whether haplotypes or single-SNP genotypes were used. With 500 markers the Anvik River reporting group had a maximum of ~90% accuracy at 100% simulation for the single-SNP genotype data, but increased to ~97% accuracy with the haplotypes (Table 2, Fig. 1). The Kogrukluk/Koktuli reporting group had a maximum accuracy of ~94% at 100% simulation with the single-SNP genotype data and increased to ~98% accuracy with the haplotypes (Table 2, Fig. 2). Precision of mixture analysis also increased, with the 95% distribution of mixture estimates decreasing by approximately half for haplotype vs single-SNP analyses. Decreasing the number of loci for mixture analysis to only 100 loci resulted in assignment accuracy as low as ~77% for single-SNP genotypes and ~86% for haplotypes for 100% simulations (Table 3). The panel of 150 loci was the smallest number of loci that achieved 90% accuracy with haplotypes while at least 300 loci were needed to achieve 90% accuracy for the single-SNP genotype data (Table 3).

Individual assignment to reporting groups varied considerably across reporting groups (Table 4). The Anvik reporting group had the lowest assignment rate at 75% followed by the Kogrukluk/Koktuli reporting group (81%) with the 500 locus genotype data. With the 500 locus haplotype data both of these reporting groups had 92% accuracy. The Big Salmon and Tubutulik reporting groups had 100% assignment for both datasets. When fewer than 500 markers were used, neither the single-SNP genotype, or haplotype datasets were able to achieve 90% accuracy for individual assignment to the Anvik or Kogrukluk/Koktuli reporting groups.

Discussion

Haplotype data showed greater accuracy and precision for mixture analysis and individual assignment than did single-SNP genotyping for the same set of loci where comparisons were possible. This increase in accuracy was obtained through a simple change in analysis of the same sequence data. No comparisons were possible for the Big Salmon River and Tubutulik River reporting groups as they showed 100% accuracy in mixture analysis as well as individual assignment with both the single-SNP genotype and haplotype data. The high differentiation in these populations allowed near perfect accuracy, even when tested with only 100 loci (data not shown).

The power of haplotyping was demonstrated in discriminating populations that were less differentiated, in this case the Anvik and Kogrukluk/Koktuli rivers. The Anvik and Kogrukluk/Koktuli rivers showed increased accuracy and precision with the haplotype data relative to the single-SNP genotype data for both mixture analysis and individual assignment. Importantly, for individual assignment, the proportion of unassigned individuals decreased by one-half to two-thirds when haplotypes were used rather than a single SNP per locus (Table 4). This improvement is consistent with previous simulation and microsatellite studies, where loci with more alleles had greater assignment accuracy than loci with fewer alleles (Kalinowski 2004; Beacham et al. 2012).

The increased information in haplotype data can improve GSI analyses either by allowing greater resolution for reporting groups (finer-scale divisions) or by allowing for smaller locus sets in management applications. We were not able to further divide the reporting groups in our dataset but haplotypes were able to achieve the same accuracy as single-SNP genotypes with fewer loci. A 90% threshold is often used as the standard for accuracy at 100% simulation for mixture analysis (Seeb et al. 2000). Using this threshold, only 150 loci would be necessary to achieve sufficient accuracy with the haplotype data while at least 300 loci would be necessary with the single-SNP genotype data. For an amplicon sequencing panel, this would translate into a reduction in sequencing costs by a factor of two.

While we were not able to test accuracy of parentage assignment with our dataset, haplotypes are likely to have an even greater positive impact. This is particularly true for distinguishing half-sibs, aunts/uncles, or other types of complicated relationships that are likely to be more frequent in small populations or species with promiscuous/polygamous mating systems. As parentage assignment relies on comparing genotypes of individuals to the genotypes of parents, the addition of more alleles per locus through haplotype analysis should reduce uncertainty in parentage assignment.

Haplotyping clearly outperformed genotyping for mixture analysis and individual assignment; however, there are considerations for implementation in management. RADseq data are often used in the preliminary stages of mixture analysis to survey loci for informative SNPs, but haplotyping using RADseq alone is not suitable for continued monitoring because of cost and work flow. On the other hand, amplicon sequencing is likely to gain popularity for stock identification studies because of the increased flexibility in choosing the number of SNPs as well substantial cost efficiencies (Campbell et al. 2015). For amplicon sequencing there is a trade-off between the number of loci and number of individuals that can be genotyped for a given batch of sequencing; decreasing the number of loci would allow more individuals to be genotyped for the same sequencing cost. Alternatively, additional loci could be added that discriminate other populations, allowing expanded utility of a panel.

The evolution of haplotypes

Multiple-SNP DNA templates generally arise from sequential occurrence of single nucleotide substitutions. The first substitution in a monomorphic template results in two SNP alleles. When a new substitution occurs proximal to the existing SNP, the new substitution is associated with only one of the alleles from the original SNP, resulting in three possible alleles (Fig. 3). DNA templates with two polymorphic sites can have up to four alleles (a vast majority of SNPs are binary); the fourth possible allele can only arise from recombination between the original and new SNP or, very rarely, through recurrent substitution (Hudson and Kaplan 1985).

This model of haplotype evolution is supported by observations from our data. Loci with two SNPs most commonly had three alleles (84%) with two alleles being the second most common situation (14%) (Table 1). Loci with two SNPs and two alleles occur when the allele on which the second substitution occurred is lost in the population (ie. TC haplotype in Fig. 3). The relatively high proportion of loci with two SNPs and two alleles suggests that loss of alleles is relatively common. Two SNP loci with four haplotype alleles were relatively uncommon (2%) suggesting recombination occurring between two SNPs in a locus has been exceedingly rare in these short templates.

The added power of haplotypes

Multiple SNPs within a locus are expected to have arisen at different times and the geographic distribution of the resulting haplotype alleles will reflect gene flow among populations following the genesis of each allele (Mathieson and McVean 2014). Alleles resulting from recent SNPs should manifest in haplotype data as rare allelic variants. Rare allelic variants tend to be more geographically restricted than common variants and have been demonstrated to increase power to detect fine-scale population structure (O'Connor et al. 2015). Individuals with the rare allele may be assigned to population of origin with some confidence, although most individuals will be homozygous and uninformative. The haplotyping of three or more alleles enables the increased resolution of rare alleles while still providing information from the more common alleles.

We recognize that choosing loci for GSI based on haplotype information rather than ranking loci based on only one of the SNPs present will likely lead to further increases in GSI accuracy, particularly if locus selection and GSI methods can be developed that take into account the evolutionary relationship between haplotypes. In this manuscript we focused on demonstrating that haplotypes provide a benefit over using a single SNP per locus; this is particularly relevant as amplicon sequencing applications for stock identification are becoming increasingly common (e.g., many agencies that manage salmonid fisheries are adopting SNP panels similar to the methods of Campbell et al. 2015). Sequencing data from these panels can be examined to determine if loci contain multiple SNPs; if so, haplotype analysis can be conducted to evaluate relative improvements in accuracy and precision. The fact that we saw improved accuracy in both mixture analysis and individual assignment, even when markers were chosen based on single-SNP F_{ST}, supports the argument that existing panels can be improved by changing to haplotype analysis.

Analysis of haplotypes rather than single-SNP genotypes for RADseq data increased accuracy for mixture analysis in Alaskan Chinook salmon. Haplotype data are readily available from RADseq analyses, allowing existing datasets to be re-examined for informative loci. Haplotype data are also readily generated from amplicon sequencing methods, allowing informative loci to be incorporated into population monitoring. We believe that haplotyping multiple SNPs shows promise to improve the accuracy of assigning unknown fish to population or group of origin or parentage analysis whenever haplotype data are available. The most effective panels will likely include a combination of the highest F_{ST} single SNPs along with multi-SNP haplotypes.

Acknowledgements

We thank members of the Seeb Laboratory for helpful comments and two anonymous reviewers who improved the manuscript.   This research was partially funded by Alaska Sustainable Salmon Fund awards 44812 and 44913 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce, administered by the Alaska Department of Fish and Game. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Commerce, National Oceanic and Atmospheric Administration, or the Alaska Department of Fish and Game.

Literature Cited:

Abadia-Cardoso, A., Pearse, D.E., Jacobson, S., Marshall, J., Dalrymple, D., Kawasaki, F., Ruiz-Campos, G., and Garza, J.C. 2016. Population genetic structure and ancestry of steelhead/rainbow trout (*Oncorhynchus mykiss*) at the extreme southern edge of their range in North America. Conservation Genetics **17**(3): 675-689.

Ackerman, M.W., Habicht, C., and Seeb, L.W. 2011. Single-nucleotide polymorphisms (SNPs) under diversifying selection provide increased accuracy and precision in mixed-stock analyses of sockeye salmon from the Copper River, Alaska. Transactions of the American Fisheries Society **140**(3): 865-881.

ADF&G. 2013. Chinook salmon stock assessment and research plan Special Publication No. 13-01. Alaska Department of Fish and Game, Special Publication No. 13-01. Available from http://www.adfg.alaska.gov/static/home/news/hottopics/pdfs/chinook_research_plan.pdf.

Ali, O.A., O'Rourke, S.M., Amish, S.J., Meek, M.H., Luikart, G., Jeffres, C., and Miller, M.R. 2016. RAD capture (Rapture): flexible and efficient sequence-based genotyping. Genetics **202**(2): 389-400.

Anderson, E.C. 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Molecular Ecology Resources **10**(4): 701-710.

Anderson, E.C., Waples, R.S., and Kalinowski, S.T. 2008. An improved method for predicting the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences **65**(7): 1475-1486.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics **17**(2): 81-92.

Araneda, C., Larrain, M.A., Hecht, B., and Narum, S. 2016. Adaptive genetic variation distinguishes Chilean blue mussels (*Mytilus chilensis*) from different marine environments. Ecol Evol.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS ONE **3**(10): e3376.

Beacham, T.D., Jonsen, K., and Wallace, C. 2012. A comparison of stock and individual identification for Chinook salmon in British Columbia provided by microsatellites and single-nucleotide polymorphisms. Marine and Coastal Fisheries **4**(1): 1-22.

Beacham, T.D., McIntosh, B., and Wallace, C. 2010. A comparison of stock and individual identification for sockeye salmon (*Oncorhynchus nerka*) in British Columbia provided by microsatellites and single nucleotide polymorphisms. Canadian Journal of Fisheries and Aquatic Sciences **67**(8): 1274-1290.

Beacham, T.D., Winter, I., Jonsen, K.L., Wetklo, M., Deng, L.T., and Candy, J.R. 2008. The application of rapid microsatellite-based stock identification to management of a Chinook salmon troll fishery off the Queen Charlotte Islands, British Columbia. North American Journal of Fisheries Management **28**(3): 849-855.

Bradbury, I.R., Hamilton, L.C., Chaput, G., Robertson, M.J., Goraguer, H., Walsh, A., Morris, V., Reddin, D., Dempson, J.B., Sheehan, T.F., King, T., and Bernatchez, L. 2016. Genetic mixed stock analysis of an interceptory Atlantic salmon fishery in the Northwest Atlantic. Fisheries Research **174**: 234-244.

Campbell, N.R., Harmon, S.A., and Narum, S.R. 2015. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources **15**(4): 855-867.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. **22**(11): 3124-3140.

Dann, T.H., Habicht, C., Baker, T.T., and Seeb, J.E. 2013. Exploiting genetic diversity to balance conservation and harvest of migratory salmon. Canadian Journal of Fisheries and Aquatic Sciences **70**(5): 785-793.

Ensing, D., Crozier, W.W., Boylan, P., O'Maoileidigh, N., and McGinnity, P. 2013. An analysis of genetic stock identification on a small geographical scale using microsatellite markers, and its application in the management of a mixed-stock fishery for Atlantic salmon *Salmo salar* in Ireland. J Fish Biol **82**(6): 2080-2094.

Evans, M.L., Johnson, M.A., Jacobson, D., Wang, J.L., Hogansen, M., and O'Malley, K.G. 2016. Evaluating a multi-generational reintroduction program for threatened salmon using genetic parentage analysis. Canadian Journal of Fisheries and Aquatic Sciences **73**(5): 844-852.

Fariello, M.I., Boitard, S., Naya, H., San Cristobal, M., and Servin, B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics **193**(3): 929-941.

Ford, M., Pearsons, T.N., and Murdoch, A. 2015. The spawning success of early maturing resident hatchery Chinook salmon in a natural river system. Transactions of the American Fisheries Society **144**(3): 539-548.

Gattepaille, L.M., and Jakobsson, M. 2012. Combining markers into haplotypes can improve population structure inference. Genetics **190**(1): 159-174.

Gisclair, B.R. 2009. Salmon bycatch management in the Bering Sea walleye pollock fishery: threats and opportunities for Western Alaska. *In* Pacific Salmon: Ecology and Management of Western Alaska's Populations. *Edited by* C.C. Krueger and C.E. Zimmerman. American Fisheries Society Symposium 70, Bethesda, Maryland. pp. 799-816.

Gustafson, R.G., Waples, R.S., Myers, J.M., Weitkamp, L.A., Bryant, G.J., Johnson, O.W., and Hard, J.J. 2007. Pacific salmon extinctions: Quantifying lost and remaining diversity. Conservation Biology **21**(4): 1009-1020.

Habicht, C., Seeb, L.W., Myers, K.W., Farley, E.V., and Seeb, J.E. 2010. Summer-fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms (SNPs). Transactions of the American Fisheries Society **139**: 1171-1191.

Hohenlohe, P.A., Amish, S.J., Catchen, J.M., Allendorf, F.W., and Luikart, G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources **11 Suppl 1**: 117-122.

Hudson, R.R., and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**: 147-164.

Kalinowski, S.T. 2004. Genetic polymorphism and mixed-stock fisheries analysis. Canadian Journal of Fisheries and Aquatic Sciences **61**(7): 1075-1082.

Larson, W.A., Seeb, J.E., Pascal, C.E., Templin, W.D., and Seeb, L.W. 2014a. Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. Canadian Journal of Fisheries and Aquatic Sciences **71**(5): 698-708.

Larson, W.A., Seeb, L.W., Everett, M.V., Waples, R.K., Templin, W.D., and Seeb, J.E. 2014b. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evolutionary Applications **7**(3): 355-369.

Larson, W.A., Utter, F.M., Myers, K.W., Templin, W.D., Seeb, J.E., Guthrie III, C.M., Bugaev, A.V., Seeb, L.W., and Moran, P. 2013. Single-nucleotide polymorphisms reveal distribution and migration of Chinook salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean. Canadian Journal of Fisheries and Aquatic Sciences **70**(1): 128-141.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Wellcome Trust Case Control Consortium, International Multiple Sclerosis Genetics Consortium, Lawson, D.J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., and Bodmer, W. 2015. The fine-scale genetic structure of the British population. Nature **519**(7543): 309-314.

Manel, S., Gaggiotti, O.E., and Waples, R.S. 2005. Assignment methods: matching biological questions techniques with appropriate techniques. Trends in Ecology & Evolution **20**(3): 136-142.

Mathieson, I., and McVean, G. 2014. Demography and the age of rare variants. PLoS genetics **10**(8): e1004528.

McKinney, G.J., Seeb, L.W., Larson, W.A., Gomez-Uchida, D., Limborg, M.T., Brieuc, M.S., Everett, M.V., Naish, K.A., Waples, R.K., and Seeb, J.E. 2016. An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (*Oncorhynchus tshawytscha*). Molecular Ecology Resources **16**(3): 769-783.

O'Connor, T.D., Fu, W., Project, N.G.E.S., ESP Population Genetics and Statistical Analysis Working Group, Turner, E., Mychaleckyj, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal, S.M., Smith, J.D., Rieder, M.J., Bamshad, M.J., Nickerson, D.A., and Akey, J.M. 2015. Rare variation facilitates inferences of fine-scale population structure in humans. Mol Biol Evol **32**(3): 653-660.

O'Malley, K.G., Camara, M.D., and Banks, M.A. 2007. Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). Molecular Ecology **16**(23): 4930-4941.

Rousset, F. 2008. Genepop'007: A complete re-implementation of the Genepop software for Windows and Linux. Molecular Ecology Resources **8**(1): 103-106.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L.W. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources **11**: 1-8.

Seeb, L.W., Antonovich, A., Banks, M., Beacham, T., Bellinger, R., Blankenship, S., Campbell, M., Decovich, N., Garza, J.C., Guthrie, C., Lundrigan, T., Moran, P., Narum, S., Stephenson, J., Supernault, J., Teel, D., Templin, W.D., Wenburg, J.K., Young, S., and Smith, C.T. 2007. Development of a standardized DNA database for Chinook salmon. Fisheries **32**(11): 540-552.

Seeb, L.W., Habicht, C., Templin, W.D., Tarbox, K.E., Davis, R.Z., Brannian, L.K., and Seeb, J.E. 2000. Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the *Exxon Valdez* oil spill. Transactions of the American Fisheries Society **129**: 1223-1249.

Steele, C.A., Anderson, E.C., Ackerman, M.W., Hess, M.A., Campbell, N.R., Narum, S.R., Campbell, M.R., and Grant, J. 2013. A validation of parentage-based tagging using hatchery steelhead in the Snake River basin. Canadian Journal of Fisheries and Aquatic Sciences **70**(7): 1046-1054.

Templin, W.D., Seeb, J.E., Jasper, J.R., Barclay, A.W., and Seeb, L.W. 2011. Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. Molecular Ecology Resources **11 Suppl 1**: 226-246.

Wolfe, R.J., and Spaeder, J. 2009. People and salmon of the Yukon and Kuskokwim Drainages and Noron Sound in Alaska: fishery harvests, culture change, and local knowledge systems. *In* Pacific Salmon Ecology and Management of Western Alaska's Populations. *Edited by* C.C. Krueger and C.E. Zimmerman. American Fisheries Society, Bethesda, Maryland. pp. 349-379.

Table 1. Distribution of number of SNPs and number of alleles per locus for all loci.

Number of SNPs	Number of Loci	Number of Alleles		
		2	3	4
1	10,448 (72%)	10,448 (100%)		
2	3,441 (24%)	476 (14%)	2,878 (84%)	87 (2%)
3	556 (4%)		205 (37%)	351 (63%)
4	46 (<<1%)		4 (9%)	42 (91%)
5	3 (<<1%)			3 (100%)

Table 2. Estimated mixture proportion at 100% simulation for each reporting group using the 500 highest F_{ST} ranked multiple SNP loci. Results are reported as mean ± threshold for the 95% distribution of mixture estimates.

Reporting Group	Single-SNP Genotype	Haplotype
Big Salmon River	100%±0	100%±0
Tubutulik River	100%±0	100%±0
Anvik River	89.6%±3.6	97.4%±1.8
Kogrukluk/Koktuli Rivers	93.7%±3.1	97.8%±1.7

Table 3. Estimated mixture proportion at 100% simulation for each reporting group reported as mean ± threshold for the 95% distribution of mixture estimates. The number of loci in each panel were decreased in intervals of 100 to test the accuracy of different panel sizes. The 150 locus panel is the exception and was the smallest panel to achieve 90% assignment with the haplotype data. Results are not shown for the Big Salmon River and Tubutulik River reporting groups; mixture estimates were ~100% for Big Salmon River and Tubutulik River reporting groups for all panels of loci.

Reporting Group	Number of Loci	Single-SNP Genotype	Haplotype
Anvik River	500	89.6%±3.6	97.4%±1.8
	400	90.7%±3.5	97.2%±1.9
	300	87.3%±4.3	94.3%±2.7
	200	87.1%±4.4	93.6%±3.1
	150	80.9%±5.5	92.7%±3.3
	100	61.7%±7.6	77.7%±5.2
Kogrukluk/Koktuli Rivers	500	93.7%±3.1	97.8%±1.7
	400	92.4%±3.2	97.1%±2.0
	300	89.4%±3.9	95.4%±2.5
	200	83.2%±5.0	91.2%±3.6
	150	80.1%±5.5	90.1%±3.8
	100	77.0%±6.8	86.1%±4.8

Table 4. Individual assignment rates to reporting groups for simulated individuals with A) 500 locus genotype dataset and B) 500 locus haplotype dataset. True reporting groups are rows while assigned reporting groups are columns.

A. Genotype

	Big Salmon	Tubutulik	Anvik	Kogrukluk/Koktuli	Unassigned
Big Salmon	100%	0%	0%	0%	0%
Tubutulik	0%	100%	0%	0%	0%
Anvik	0%	0%	75%	8%	17%
Kogrukluk/Koktuli	0%	0%	5%	81%	14%

B. Haplotype

	Big Salmon	Tubutulik	Anvik	Kogrukluk/Koktuli	Unassigned
Big Salmon	100%	0%	0%	0%	0%
Tubutulik	0%	100%	0%	0%	0%
Anvik	0%	0%	92%	3%	5%
Kogrukluk/Koktuli	0%	0%	3%	92%	6%

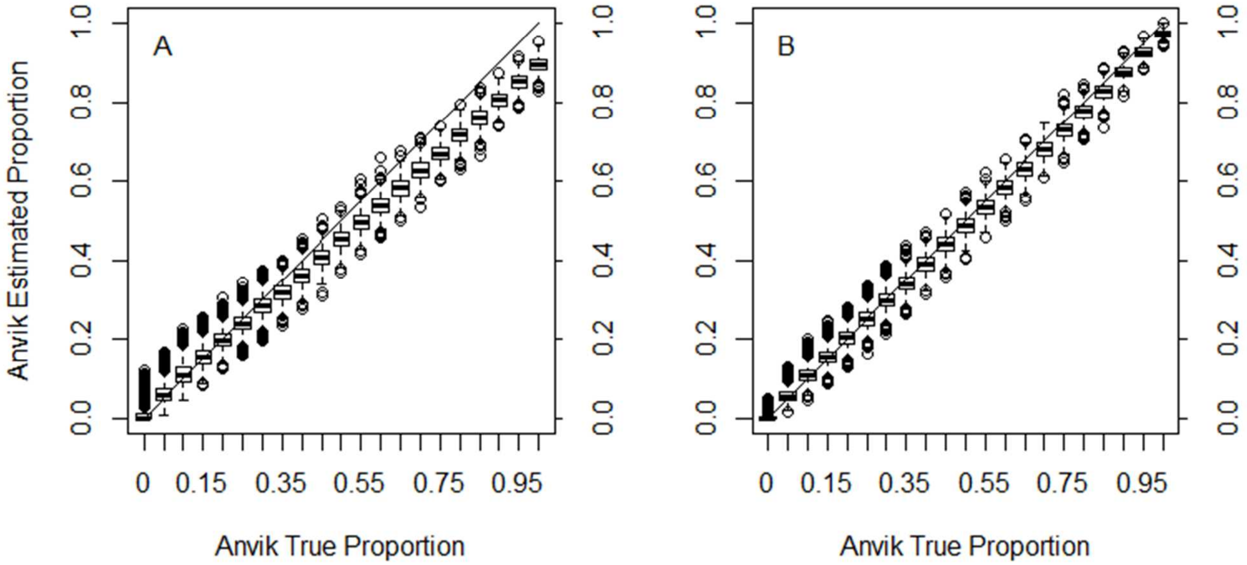

Figure 1. Boxplots showing estimated mixture proportion for: A) the Anvik River using single-SNP genotypes for the 500 highest F_{ST} loci which have 3 or more haplotype alleles. B) Anvik River using haplotypes for the same loci as A. The diagonal line shows the expected relationship for perfect assignment. The true proportion simulated in the mixture is given on the x-axis while the boxplots show the estimated proportions on the y-axis.

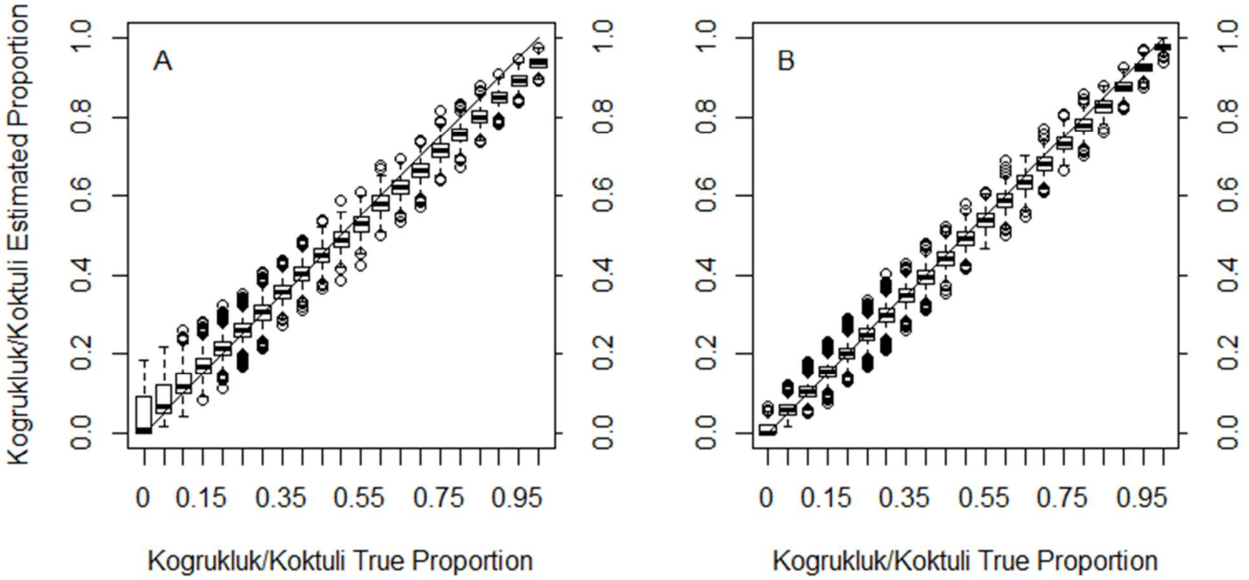

Figure 2. Boxplots showing estimated mixture proportion for A) the Kogrukluk/Koktuli reporting group using single-SNP genotypes for the 500 highest F_{ST} loci which have 3 or more alleles. B) Kogrukluk/Koktuli reporting group using haplotypes for the same loci as A. The diagonal line shows the expected relationship for perfect assignment. The true proportion simulated in the mixture is given on the x-axis while the boxplots show the estimated proportions on the y-axis.

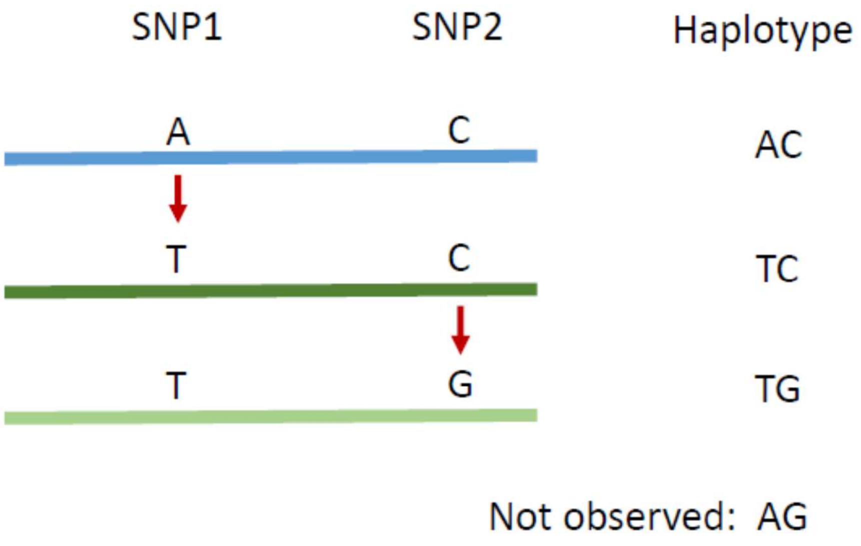

Figure 3. Haplotypes resulting from two hypothetical SNPs in the absence of recombination. The ancestral DNA sequence is in blue; the first SNP is an A/T variant that gives rise to the dark green DNA sequence. The second SNP (C/G variant) will arise on one of these two alleles, in this case the T allele. The two SNPs result in three haplotypes: AC, TC, and TG (light green DNA sequence). Assuming the probability of a second mutation at the same site is negligible, the fourth possible haplotype (AG, unobserved) can only arise through recombination between the first and second SNP.