Expanded Bilateral Chum Salmon SNP Genetic Baseline for Genetic Stock Identification

Final Report for the Pacific Salmon Commission – Southern Endowment Fund

Ben J. G. Sutherland[1], John Candy[1], Carrie Gummer[1], Kim Jonsen[1], Terry Beacham[1]

[1] Fisheries and Oceans Canada, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, B.C., Canada, V9T 6N7

Final Report, June 29 2020

A project funded by the Southern Boundary Restoration and Enhancement Fund (2018-2020)

Table of Contents

Abstract

An amplicon sequencing panel was generated to genotype single nucleotide polymorphisms (SNPs) in chum salmon (*Oncorhynchus keta*) in order to improve the resolution of the current genetic baseline, as well as to improve both automation and genotyping result transferability between laboratories in the United States and Canada. In short, this allowed for the development of a high-resolution, bilateral chum salmon SNP panel that can be applied in the Southern Boundary Region. This panel was generated from a series of markers shared between laboratories, and thus common baselines can be used by both countries, making results more comparable. Specifically, this project aimed to expand the initial SNP baseline containing 38 populations to include better representation of Canadian chum salmon populations in the Fraser River and along the Strait of Georgia, as well as Puget Sound populations in Hood Canal. The SNP markers that have been implemented by the US genetics laboratory are now being genotyped using the DFO sequencing platform and genotyping pipeline. Efforts between the laboratories to calibrate and synchronize output have been largely successful. The goal to use this SNP platform to add an additional 25 Canadian populations that were not yet in the US SNP baseline has been met and surpassed with supplemental genotyping. The chum baseline is now fully operational within DFO, initial steps of calibrating between the US and Canada have been taken to facilitate baseline sharing, and mixed-stock analyses are ready to be undertaken with this new platform. Here we report the details on the SNP baseline, with a specific focus on the improvements of resolution for specific goals proposed in the study.

Introduction

Along the Southern Boundary of Canada and the United States, a moderate level of concern exists for populations of chum salmon *Oncorhynchus keta*, which has led to a reduction or suspension of commercial exploitation in some areas and years. Improvements in the ability to distinguish stocks of concern from healthy stocks are needed, creating an impetus for the development of advanced genetic methods to improve resolution within genetic baselines for use in mixed-stock analysis. Furthermore, given that these issues impact both sides of the US-Canada southern border, it is also a major goal to have a panel that is largely common between the countries so that results can be easily compared. This will facilitate the ability to share baseline genotypes to conduct mixed-stock analysis with a common reference baseline. The fishing areas in the US and Canada expected to be particularly informed by such results are shown in Figures 1A and 1B, respectively.

Populations that were highlighted as needing expansion at the start of this project included regions in southern British Columbia, the Juan de Fuca strait, Puget Sound (Hood Canal), and Coastal Washington. Specifically, additions were needed to the SNP baseline including the Canadian stocks in the Fraser River

(e.g., Chilliwack River summer and fall runs; Chilliwack being the index stock for the abundance of chum salmon in the Fraser River), on the east and west coast of Vancouver Island (ECVI & WCVI), and on the coastal mainland (i.e., Johnstone Strait and inlet stocks). Additions were also needed from US regions including southern Puget Sound (SPS; summer and fall runs), and the Washington coast (fall runs). Stocks that were present at the beginning of this project are shown in Table 1, and stocks that were proposed to be added as a result of this project are shown in Table 2. By including key stocks, such as the threatened Hood Canal summer run chum salmon, fishery management can make efforts to avoid pressures on these stocks depending on when and where they arise in mixed-stock fisheries or bycatch.

In addition to generally increasing the comprehensiveness of the SNP baseline, there are specific goals that were to be addressed with the improved genetic baseline, in particular with expected improvements in resolution afforded by a larger number of SNP markers relative to previous microsatellite baselines. These challenges included (A) distinguishing lower Fraser (LFR) stocks from those in Northern Puget Sound (NPS); (B) distinguishing the Puntledge River and Qualicum River populations; (C) distinguishing between the Chilliwack River and related populations from other Fraser River populations for analyzing results from the Fraser River Test Fishery; and more broadly (D) improving the overall resolution among stocks in the Southern Boundary Region between Canada and the US relative to previous baselines, including the chum salmon microsatellite baseline at DFO (Beacham et al. 2009).

The chum salmon SNP amplicon panel should enable the Chum Technical Committee to "…*better address its obligations under the Pacific Salmon Treaty (Annex IV Chapter 6) to estimate and document the stock composition and exploitation rates in fisheries of concern to the treaty, evaluate stock composition information for fisheries using bilaterally agreed upon methods, and manage for catch composition in mixed-stock fisheries.*…", all of which were deemed as high impact items within this proposal. The Chum Technical Committee considers this project to be an important step forwards in managing chum salmon on both sides of the border.

The Canadian efforts of this proposal are put towards two parts: 1) implement amplicon (SNP) technology for chum salmon using the Ion Torrent (Thermo Fisher) sequencer and genotyping platform as previously implemented for coho salmon (Beacham et al. 2017) and Chinook salmon (Beacham et al. 2018); and 2) use the developed panel to run an additional 25 Canadian populations not yet in the US baseline. We also have put efforts towards improving the transferability of results between countries by ensuring comparability of marker genotypes. In addition to the populations proposed, there are a series of other populations that were added to the baseline, which we include within this report for completeness.

Materials and Methods

Amplicon panel design

Markers were obtained from several sources, including the University of Washington (UW) GTseq panel (Oke_GTseq350; n = 198 markers), the Washington Department of Fish and Wildlife (WDFW) marker panel (n = 163 markers), two collections of additional markers provided by UW (Jim Seeb, *pers. comm*; n = 200 and n = 113 markers), as well as markers from the DFO Molecular Genetics Lab including species identification markers (MGL; n = 11 markers). These datasets were all collected using the code repository *fasta_SNP_extraction* (see *Data Availability*). Collectively, this resulted in a total of 685 targets for development. Two markers from UW were screened out as they were insertion/deletions rather than SNPs, leaving a total of 683 markers in this amplicon panel.

For the 683 input markers, the available sequence data typically derived from RADseq data (often less than 200 bp) were collected into a fasta file and aligned against a contig-level chum salmon reference genome produced by the University of Victoria (B. Koop, *pers. comm.*), using BLAST (Altschul et al. 1990) with an e-value cutoff of $1e^{-30}$. To avoid primer designs in repetitive regions of the genome, only those markers that aligned four or fewer times were retained for further analysis, which resulted in dropping 86 markers. Using the *fasta_SNP_extraction* pipeline, a total of 400 bp of flanking sequence was taken from the reference genome to optimally obtain 200 bp on either side of the targeted variant. The two alleles for the variant were inserted into the 400 bp fragment, and this was submitted to Thermo Fisher for design of an AmpliSeq panel, as previously conducted (Sutherland et al. 2020). The Thermo Fisher design team was able to design all primers for the panel, as per manufacturers' methods, with the exception of six markers that were dropped from the panel due to design issues. For full details on this method, see Sutherland et al. (2020) and the *fasta_SNP_extraction* pipeline (see *Data Availability*).

Sample selection and population inclusion/annotation

Populations as listed in Table 2 were obtained from either the tissue archive at DFO or from US collaborators on the Chum Technical Committee (Maureen Small, *pers. comm.*). Collections were preferentially selected from more recent years rather than older collections when the option was possible. For each population, 100 individuals were targeted. Further, for some collections, quality control resulted in screening out of individuals (see below) and in some cases this necessitated adding additional samples from other sampling years. In some cases, additional samples were genotyped to improve the resolution in stocks that were similar to other populations in the area and that were identified as important stocks where advanced resolving power would be beneficial, such as those on ECVI (Pieter van Will, Lee Kearey, *pers. comm.*) or in the Fraser River (Joe Tadey, *pers. comm.*).

Sequencing and variant calling

DNA was extracted from tissue using a variety of methods (see Sutherland et al. 2020 for more details). The preferred method applied when possible was the BioSprint (QIAGEN) extraction approach due to improved yields in the SNP panel genotyping after quality control. Samples were prepared in 96-well plates with a negative control on each plate. DNA from each fish was normalized to 40 ng/µl, barcoded and amplified according to the AgriSeq panel protocol using the chum v.1.0 primers, as per manufacturers' instructions, using the available 768 barcodes (Thermo Fisher), as previously described (Beacham et al. 2017). Pools of barcoded 768 individuals were collected and sequenced on a Ion Torrent PI chip using the Ion Chef to prepare sets of two chips for sequencing on a single run of the Ion Torrent. Sequenced samples were de-multiplexed, the total number of reads per chip assessed for quality of the overall chip, and variants called per sample using a hotspots file and the Torrent Suite software (TS v.5.10.1; *variantCaller* v.5.6.0.4; Thermo Fisher). Variants per sample were then imported into the MGL genotype database management system for downstream filtering.

Data filtering

Data filtering was conducted sequentially, as per Sutherland et al. (2020). In brief, samples were removed if there were too many missing genotypes per sample (i.e., retain individuals that were missing no more than 270 of the 567 amplicons; genotyped at least at 52% of markers). Second, populations with fewer than 20 individuals were removed. Third, amplicons with observed heterozygosity of greater than 50% were removed in order to remove paralogs or elsewise non-functioning loci. Fourth, amplicons were removed if the amplicon was not genotyped in more than 50% of the genotyped individuals across the filtered dataset, as these indicate poorly performing amplicons. The filtered baseline, including populations from the current project as well as other surrounding populations and regions, was then exported with a date-stamp and converted to genepop format for downstream data analysis.

Population differentiation analysis and simulated mixed-stock samples (100% simulations)

Population analyses were conducted using the *simple_pop_stats* repository (see Data Availability), a pipeline that uses many population genetic tools applied at MGL. Baseline genotypes were read into R (R Core Team 2020) in genepop format using adegenet (Jombart 2008). Pairwise F_{ST} (Weir and Cockerham 1984) was calculated using all populations of interest using hierfstat (Goudet 2005). A neighbour-joining tree using the *edwards.dist* distance metric (Cavalli-Sforza and Edwards 1967) was generated for the populations of interest to this project using the aboot function of *poppr* (Kamvar et al. 2014) with 10,000 bootstraps, then exported as tree format for input to FigTree v1.4.4 for data visualization (Rambaut 2019).

In order to evaluate the ability to resolve populations at the population level or at the level of the repunit (reporting unit; here Conservation Units), an analysis using the function *assess_reference_loo* from the rubias analysis package (Moran and Anderson 2018), as implemented in *simple_pop_stats*. In short, each population was used from the baseline to simulate a mixture of 200 fish using the allele frequencies in the baseline, and then these simulated fish were assigned back to the baseline to determine how well they assigned to the population of origin, as well as the region of origin. This was conducted a total of 100 simulations in order to get summary statistics such as standard deviation on assignment success. This was conducted iteratively until each population in the baseline had a result for correct assignment to population, and correct assignment to reporting unit. This was visualized into bar graphs. This 100% simulation assignment test was conducted first using the entire baseline with reporting units at the CU level, and second using a restricted baseline of only the lower Fraser River populations with reporting units as Chilliwack populations and other Fraser River populations, which are groupings used for genotyping of juveniles with the microsatellite baseline at the Fraser River Test Fishery (Joe Tadey, *pers. comm.*).

Results and Discussion

SNP panels with around 500 polymorphic markers are expected to show improved resolution relative to microsatellite panels with under 15 markers, as demonstrated recently in the marine fish eulachon *Thaleichthys pacificus* that has low population structure and is thus difficult to resolve (Sutherland et al. 2020). Further, SNP panels can exceed the power of the microsatellite panels with fewer samples in the baseline, as the SNP panel is only binary SNPs whereas microsatellites can have over 100 alleles (Beacham et al. 2011; Beacham et al. 2012).

Inclusion of additional populations to the DFO baseline

In total, 25 populations from the lower Fraser River (LFR), Straight of Georgia (SoG), Johnstone Strait (JS), and west coast Vancouver Island (WCVI) were proposed to be added to the genotyped populations in the SNP baseline. Although the target was 100 samples from each population, in some cases samples did not amplify and after quality control, a critical step to develop a strong, reliable baseline, the number of samples genotyped sufficiently were often lower than the 100 samples (mean sample size = 67.3 samples per population). However, to augment the total number of samples, neighbouring populations were added to the baseline (see below). The samples that did genotype were amplified at over half the amplicons in the panel and therefore the retained samples are strong representations of the populations. Samples were typically taken from a single year, although six of the 25 target populations were sourced from collections occurring over two years, and two were from collections occurring over three years.

In addition to the 25 proposed population additions, we also added a large number of other southern populations to improve the resolution of the baseline (Table 3). In total, the populations from the Columbia River (US) to Smith Inlet (British Columbia) included 136 populations with an average of 72 individuals per population (minimum = 20; maximum = 193). Of the 136 populations, 100 populations were within Canada and 36 within the US. This dataset is comprised of a total of 9,823 individuals, and represents a strong foundation for genetic stock identification within the southern boundary region (e.g., Figure 1). This full set of southern baseline populations included populations of which collections spanned four years in four populations, three years in 11 populations, two years in 44 populations, and a single year in 77 populations.

Baseline overview

The genetic structure overall fits well with the Conservation Unit structure within Canada, and the general groupings designated in the US. Here we will go into some detail regarding the structure of the southern populations.

The US populations all grouped consistently outside of the Canadian populations with the exception of the Northern Puget Sound (NPS) grouping (see below; Figure 2). Three Hood Canal (HOOD) populations were outgroups to the entire southern baseline: Big Quilcene River (Summer run), Salmon Creek, and Jimmycomelately Creek. These populations are summer run populations, which would explain their distinctiveness to the other fall run populations. Other Hood Canal and Tulalip (TUL) populations were grouped together without any discerning structure between the two regional groupings. Therefore it is likely that these two groupings, HOOD and TUL, should be made into a single grouping, whereas the Hood Canal summer run populations as the outgroup should be a separate reporting unit. The Central and Southern Puget Sound (CPS, SPS) populations were grouped together, with the exception of South Prairie Creek, which was not grouped with the other SPS populations. Three distinct groupings exist within the SPS/CPS branch, the SPS populations Puyallup, Nisqually, and Nisqually River winter run all form a distinct grouping with strong bootstrap support (100%). The other two groupings are closer together, but also form substructure, where the CPS populations Chico Grovers Hatchery and Green River Keta Hatchery form a strongly supported grouping, and the SPS populations including Skookum Creek (fall run) and others form a strongly supported grouping (94% bootstrap support). The Coastal Washington (CWA) stocks and the Columbia River population (only one available currently) form a grouping as an outgroup to the more northern populations, and both these groupings should be identifiable within mixed-stock analysis due to high bootstrap support (100% bootstrap support of the CWA sub grouping). Although the Northern Puget Sound (NPS) is grouped closer to the Canadian stocks, there is still a very strong grouping of these separate from the Lower Fraser (LFR) stocks, where the majority of the NPS stocks are contained within a bootstrap

support of 98%, leaving only Nooksack River as an outgroup to this strongly supported branch. With such elevated bootstrap support for these populations, the separation of Canadian and US populations should be very effective with this current SNP panel. As evidence for this, both the LFR and the NPS stocks show an average assignment to the reporting unit of greater than 90% (Figure 3).

The Canadian populations are the majority of the collections in this project. The Fraser River has a strongly supported grouping, with some sub-structure, but in general all of the stocks are fairly similar. Some populations such as Silverhope Creek and Hunter Creek show strong bootstrap support of being distinct from other populations, but in general the distinctiveness within the Fraser is not substantial (see discussion below regarding separation of Chilliwack River and other Fraser stocks). The groupings of populations within the CUs Georgia Strait Southern Fjords (GStr-SFj) are spread out into several different groupings, however, overall populations within this grouping show strong assignment to reporting unit (greater than 90% on average; Figure 3). The Bute Inlet (BUTE) stocks are very well resolved, with high bootstrap support. Although the Howe Sound (HOWE) and Loughborough (LOUGH) stocks are mainly all grouped together in the dendrogram (Figure 2), they do have lower bootstrap support. Results of 100% simulations suggest that LOUGH and HOWE have poor resolution to the repunit on average (< 70% and ~70% to reporting unit, respectively). Further, one LOUGH stock, Glendale Creek, is showing a signature more similar to the northern populations, and so this population should be investigated as to exactly what it is comprised of by checking all available metadata. Whether LOUGH and HOWE remain as separate CUs resolvable with genetics will require further investigation, and may require either additional samples within existing populations or additional populations. The southwest Vancouver Island (SWVI) stocks are strongly grouped, with the exception of the outgroup Salmon Creek. Northwest Vancouver Island (NWVI) has a grouping that is contained within the highly supported west Vancouver Island group (SWVI and NWVI; 100% bootstrap support), and given that NWVI had high bootstrap support within a subgroup of this grouping (99%), the ability to distinguish NWVI and SWVI should be strong (mean assignment to repunit = 88% and >95%, respectively). Similar to some of the other groupings, the southern coastal streams (SCS) CU is more spread out, but does still largely form a collected grouping; simulations suggest close to 90% average assignment to repunit for these stocks. The Rivers Inlet (RIVERS) and Upper Knight Inlet (UKNIGHT) form a strongly supported grouping together (95%), but simulations indicate poor resolvability for RIVERS (mean = 49%) and not high resolvability for UKNIGHT (on average to repunit less than 75%). Additional populations or samples added may bring these repunits up to a level that they can be separated; this will also have to be explored further. Finally, the Smith Inlet grouping shows very strong clustering with high bootstrap support (95%), indicating that this CU will be easy to distinguish using this panel (simulations indicate > 95% average correct assignment to repunit). In sum, many of the groupings are strong, with several that show poorer resolution, possibly due to few populations (e.g.,

Columbia and JDF; n = 1 population each), as well as due to regions needing to be grouped (e.g., HOOD and TUL). These improvements will come with further development of the baseline for use in mixed-stock analysis.

Case Studies and Specific Challenges

Challenges that were posed as important questions as part of this proposal included (A) distinguishing lower Fraser (LFR) stocks from those in Northern Puget Sound (NPS); (B) distinguishing the Puntledge River and Qualicum River populations; (C) distinguishing between the Chilliwack River and related populations from other Fraser River populations for analyzing results from the Fraser River Test Fishery; and more broadly (D) improving the overall resolution among stocks in the Southern Boundary Region between Canada and the US.

Case Study 1: Distinguish Lower Fraser River stocks from Northern Puget Sound

As described above, the Northern Puget Sound (NPS) has good resolvability in the genetic dendrogram. The NPS is outside of the grouping containing other US stocks, but has highly supported bootstrap value (98%) for the majority of the grouping (with the exception of Nooksack River, the most outer population in the grouping). This suggests powerful ability to identify NPS in mixed-stock analysis. In terms of the Lower Fraser grouping, this grouping is also highly supported, where the grouping of all of the Lower Fraser stocks are supported by a 100% bootstrap support. From 100% simulations, the LFR has > 95% assignment to region on average, and the NPS shows 93% assignment to region on average (Figure 3). Therefore separating NPS and LFR will be very powerful with the SNP panel and the current baseline.

Case Study 2: Distinguish Puntledge River and Qualicum River populations

The Puntledge River and Qualicum River populations are very close together in the dendrogram (Figure 2), and are both grouped within the GStr-SFj sub-group of east coast Vancouver Island (including Little Qualicum, Qualicum, Puntledge, Campbell River, Cowichan, Nanaimo, Chemainus, and Goldstream Rivers. All of these populations are very similar genetically, and differ from other subgroupings of GStr-SFj including the populations on the mainland (Snake Bay Creek, Tzoonie River, Okeover Creek, Theodosia, Lang Creek, Sliammon Creek, and surprisingly, Englishman River. The Englishman River collections were from 2010 and 2011 (n = 56 total); it is not clear why Englishman River stocks are grouping with the mainland stocks, whereas all of the other East Coast Vancouver Island stocks are grouping together. The clustering of Englishman River is currently being investigated and may require new collections to occur in the future in case there is an issue with the previous collection. In any case, the separation of Little Qualicum, Qualicum River, and Puntledge River does not seem possible, even using this highly resolving

SNP panel. As evidence of this, when using a coastwide baseline, 100% simulations indicate that Puntledge River assigns to Puntledge River with 44% accuracy, second to Qualicum River with 20%, and third to Little Qualicum with 14% (Table 5). Alternately, these simulations indicate that Qualicum River assigns with 36% accuracy to Qualicum River, 24% to Little Qualicum River, and 18% to Puntledge River. These are very low percentages of correct assignments, and therefore without an alternative approach, such as that used to separate hatchery and wild stocks such as parentage-based tagging (PBT; Beacham et al. 2017), it will not be possible to separate Puntledge River and Qualicum River populations. PBT would require the continuous genotyping of broodstock individuals for the hatchery of interest, and therefore would need to have a good reason to justify such an approach for this species and population, but would be possible.

Case Study 3: Separating Chilliwack River from the other components of the Fraser River in the Fraser River Test Fishery

The Fraser Test Fishery aims to separate populations within a grouping containing Chilliwack River and related populations from the rest of the Lower Fraser River groupings (Joe Tadey, *pers. comm.*). The Chilliwack grouping contains six populations: Chilliwack River, Hopedale Slough, Peach Creek, Street Creek, Sweltzer River, and Vedder River. There are 25 stocks within the Fraser grouping (Table 4). Notably, all six Chilliwack stocks grouped together in the dendrogram (Figure 2), although the bootstrap support for this cluster is low. Looking to the 100% simulations, reasonable assignment to reporting unit is observed for populations Hopedale Creek, Vedder River, Chilliwack River (76-69% assignment success, respectively). However, lower assignment to repunit is observed for Sweltzer River, Peach Creek and Street Creek (mean = 52%). Fortunately, given that these six populations still have a small number of samples present in each population (i.e., mean = 61 individuals per populations; total = 367 individuals), this differentiation may be improved with additional samples (e.g., moving towards the 200 sample level). The other populations in the Fraser River (i.e., Fraser grouping; Table 4) have higher average assignment to reporting unit, where 19 populations assign to the correct reporting unit at greater than 90%, five populations between 84-89%, and one population, Kawkawa Creek, that is poorly performing. The Fraser grouping has an average of 70 individuals per populations, a minimum of 32 and a maximum of 123 individuals per population. Therefore, although the Chilliwack grouping could use additional samples, for the most part, the Fraser grouping is performing very well. Notably, this does not mean that assignments to the population are possible here, as the average percent assignment to the correct population for the 25 Fraser populations is only 28%. Results should be analyzed appropriately therefore, keeping the assignment to reporting unit as the main target for this type of analysis.

Case Study 4: Improving the overall resolution among stocks in the Southern Boundary Region between Canada and the US

The most challenging aspect of separating Canadian and US chum salmon along the Southern Boundary Region will be the LFR and the NPS populations, as these are the most similar to each other as shown in Figure 2. However, as described in *Case Study 1* above, we expect to have strong resolving power between these two populations. All of the other US stocks are more dissimilar to Canadian stocks. In terms of resolution within the US, there is fairly good resolving power already, and this is expected to improve with additional populations being added in the future. The weakest repunits in the US for resolving are Hood Canal and Tulalip, which did not separate in the dendrogram (Figure 2) and so not surprisingly are not returning good assignment results (60% - 70% to repunit). Arguably, as discussed above, these two should be contained within the same repunit and not grouped separated as they are currently in the baseline, and the summer run HOOD stocks should be a separate repunit. These revisions are currently being explored and will be concluded prior to formal publication of the baseline. There is a second grouping of HOOD stocks, including Big Quilcene River (summer run), Jimmycomelately Creek, and Salmon Creek, that has very good separation from the other HOOD and TUL populations, and possibly should be made its own reporting group. The Juan de Fuca (JDF) repunit had poor assignment to repunit, but only contains a single population thus far. In contrast, US stocks in repunits Northern, Central, and Southern Puget Sound all also have high assignment to repunit (>90%). Coastal Washington has very good resolving power, even to population, showing > 80% on average to population and close to 100% assignment to repunit.

Conclusions and Future Work

In conclusion, the current genetic baseline using the SNP panel is performing very well. Future work will compare these results with microsatellite baseline results to determine exactly how much performance improvement is observed. For some target stocks, such as the Chilliwack River proportion of the Lower Fraser River, we will be looking to add additional samples into the baseline to continue to improve resolving power. Some items remain to be considered in the baseline, such as the grouping of the main section of Hood Canal populations with the Tulalip populations, and creating a new grouping for the second grouping of the Hood Canal populations. As we add more northern populations, it will be important to determine if we can indeed keep to a CU or even sub-CU level grouping, while remaining within the requirements for sufficient mixed-stock resolving power (e.g., >80-90% assignment to reporting unit). With the coordination that has been occurring between the US and Canadian labs, we expect that this panel will be highly useful for ensuring that efforts are concordant on both sides of the border, bringing strong benefits through collaboration.

Data Availability

Pipeline *simple_pop_stats*: https://github.com/bensutherland/simple_pop_stats
Pipeline *fasta_SNP_extraction*: https://github.com/bensutherland/fasta_SNP_extraction

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. Journal of Molecular Biology **215**(3): 403-410.

Beacham, T.D., McIntosh, B., and Wallace, C.G. 2011. A comparison of polymorphism of genetic markers and population sample sizes required for mixed-stock analysis of sockeye salmon (*Oncorhynchus nerka*) in British Columbia. Canadian Journal of Fisheries and Aquatic Sciences **68**(3): 550-562. doi:10.1139/F10-167.

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. doi:10.1080/19425120.2011.649391.

Beacham, T.D., Candy, J.R., Le, K.D., and Wetklo, M. 2009. Population structure of chum salmon (*Oncorhynchus keta*) across the Pacific Rim, determined from microsatellite analysis. Fishery Bulletin **107**(2): 244-260.

Beacham, T.D., Wallace, C., MacConnachie, C., Jonsen, K., McIntosh, B., Candy, J.R., and Withler, R.E. 2018. Population and individual identification of Chinook salmon in British Columbia through parentage-based tagging and genetic stock identification with single nucleotide polymorphisms. Canadian Journal of Fisheries and Aquatic Sciences **75**(7): 1096-1105. doi:10.1139/cjfas-2017-0168.

Beacham, T.D., Wallace, C., MacConnachie, C., Jonsen, K., McIntosh, B., Candy, J.R., Devlin, R.H., and Withler, R.E. 2017. Population and individual identification of coho salmon in British Columbia through parentage-based tagging and genetic stock identification: an alternative to coded-wire tags. Canadian Journal of Fisheries and Aquatic Sciences **74**(9): 1391-1410. doi:10.1139/cjfas-2016-0452.

Cavalli-Sforza, L.L., and Edwards, A.W. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet **19**(3): 233-257.

Goudet, J. 2005. hierfstat, a package for R to compute and test hierarchical F-statistics [10.1111/j.1471-8278]. Molecular Ecology Notes **5**: 184-186. doi:papers3://publication/doi/10.1111/j.1471-8278.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers [10.1093/bioinformatics/btn129]. Bioinformatics **24**(11): 1403-1405. doi:papers3://publication/doi/10.1093/bioinformatics/btn129.

Kamvar, Z.N., Tabima, J.F., and Grunwald, N.J. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ **2**: e281. doi:10.7717/peerj.281.

Moran, B.M., and Anderson, E.C. 2018. Bayesian inference from the conditional genetic stock identification model. Canadian Journal of Fisheries and Aquatic Sciences **76**(4): 551-560. doi:10.1139/cjfas-2018-0016.

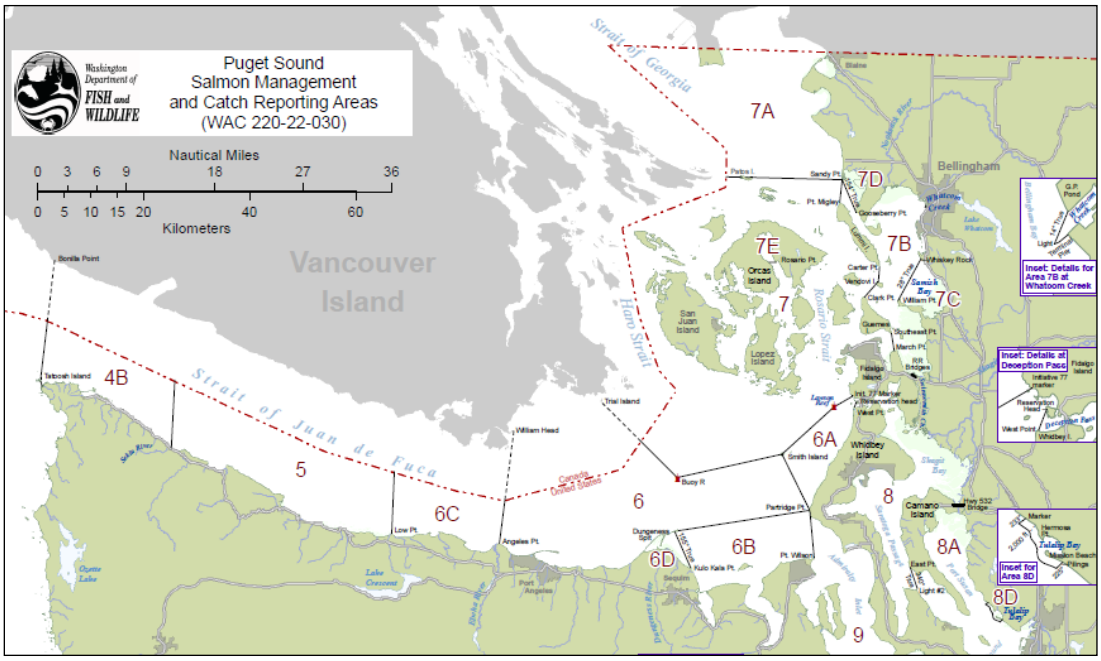
R Core Team. 2020. R: A language and environment for statistical computing. *In* R Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. 2019. FigTree.

Sutherland, B.J.G., Candy, J., Mohns, K., Cornies, O., Jonsen, K., Le, K., Gustafson, R.G., Nichols, K.M., and Beacham, T.D. 2020. Population structure of eulachon *Thaleichthys pacificus* from Northern

California to Alaska using single nucleotide polymorphisms from direct amplicon sequencing. bioRxiv: 2020.2005.2031.126268. doi:10.1101/2020.05.31.126268.
Weir, B.S., and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. Evolution **38**(6): 1358-1370. doi:10.1111/j.1558-5646.1984.tb05657.x.

Figures



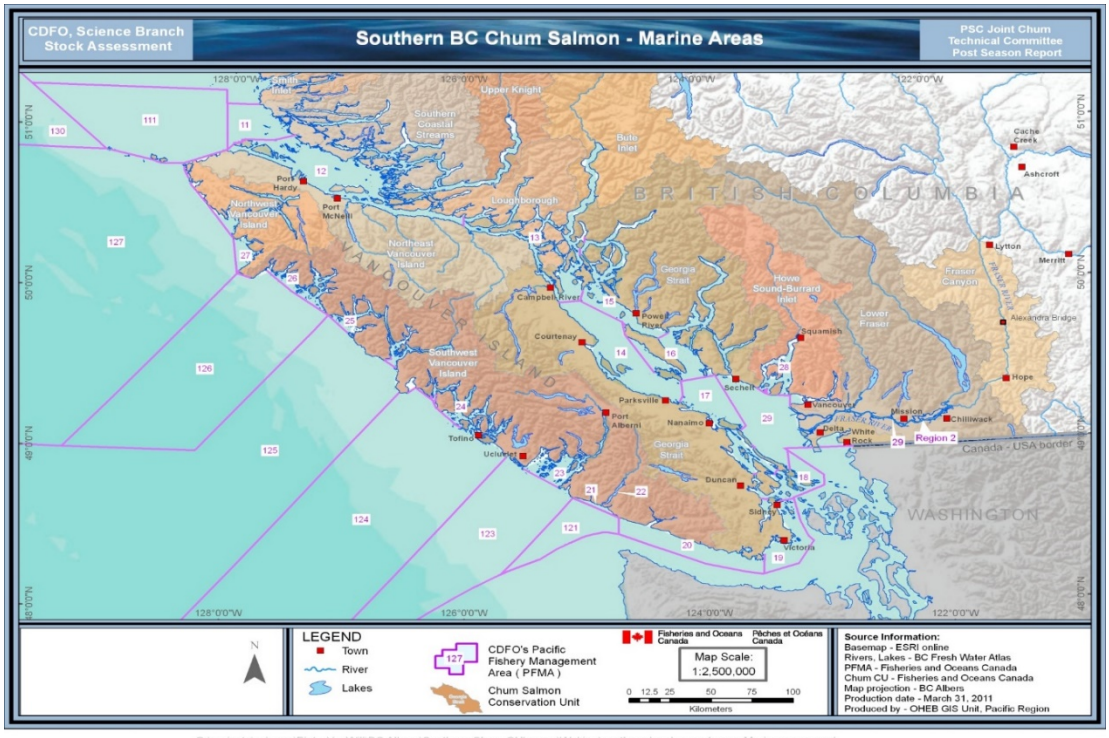


Figure 1. (A) United States commercial salmon catch areas for the Strait of Juan de Fuca and Northern Puget Sound (from 2011 CTC Annual Report). (B) Canadian Pacific Fishery Management Areas (PFMA) for Southern BC Marine areas (from 2011 CTC Annual Report).

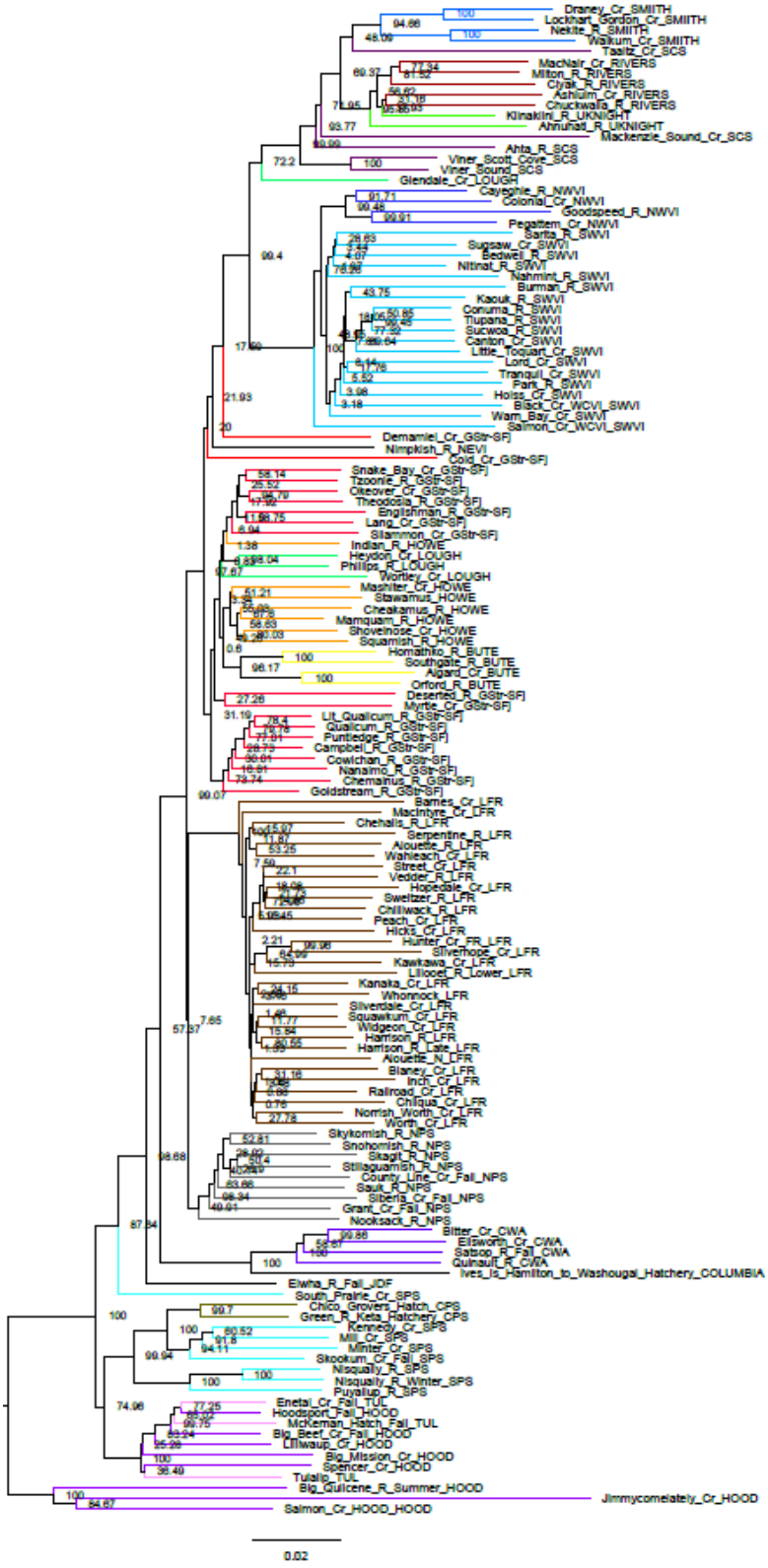


Figure 2. Dendrogram of genetic similarity among populations from Smith Inlet (Canada) to the Columbia River (US) genotyped using the newly-developed SNP panel.

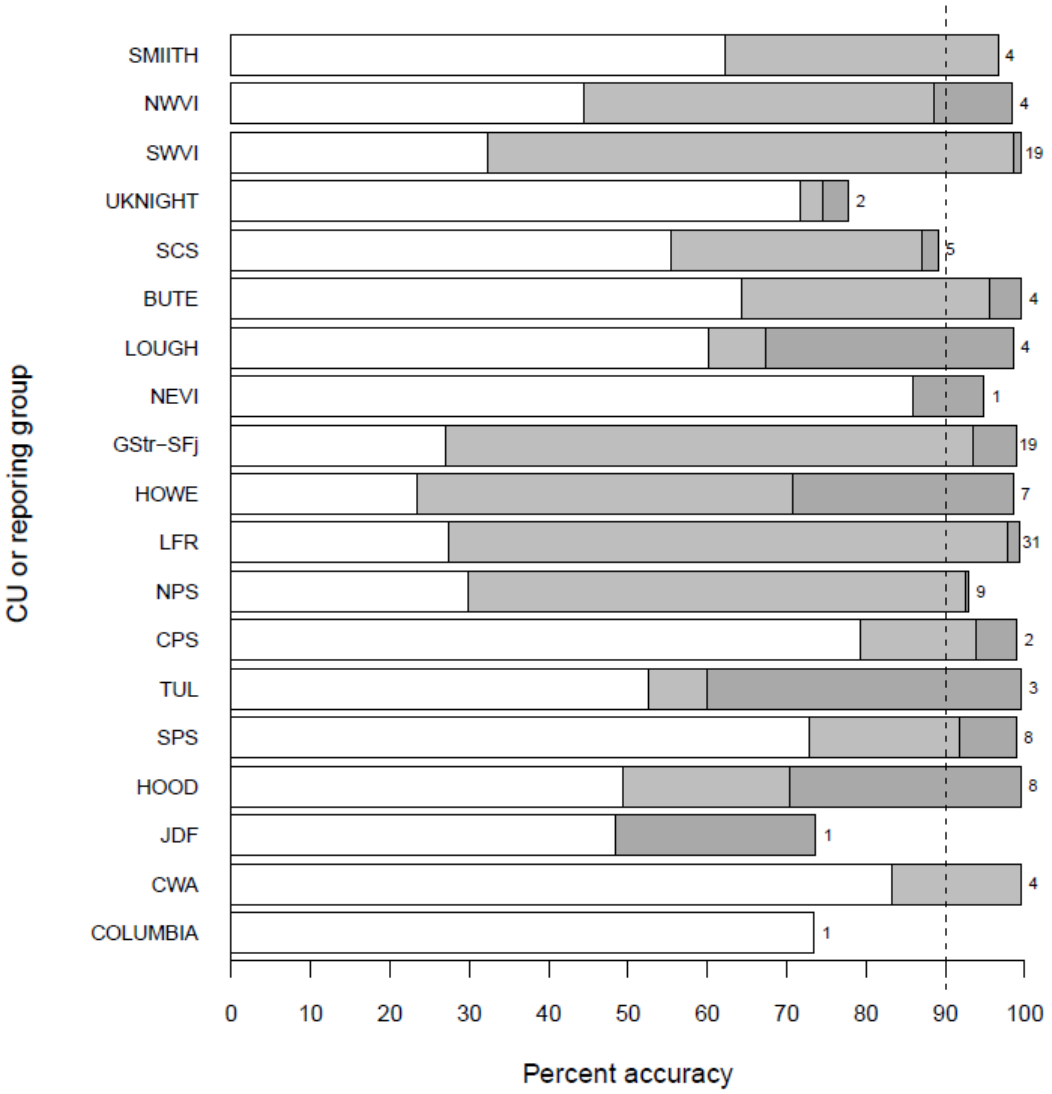


Figure 3. Summarized percent accuracy of the repunits in the genetic baseline showing average accuracy to population (white bar), to repunit (light grey), or broader region (dark grey). The repunits are shown in the y axis labels, as short names for CUs, and broader regions include: North and Central BC, Southern BC, Washington, and Columbia. Results that show accuracy at the 90% level are considered strong for genetic stock ID, although 80-90% is also suggested to provide reasonably strong results. The goal is to ensure all repunits or groupings are greater than 80% to repunit (light grey).

Tables

Table 1. List of Southern British Columbia, Puget Sound and Washington Coast stocks in the Chum SNP baseline in the US laboratories as of the start of the project.

Count	Region	run	Tributary	Year	N
CAN	Johnstone Strait	fall	Nimpkish River	2010	95
CAN	Southern mainland	fall	Southgate River	2003	95
CAN	Southern mainland	fall	Phillips River	2011	95
CAN	East Vancouver Island	fall	Puntledge River	2010	95
CAN	Southern mainland	fall	Lang Creek	2011	95
CAN	East Vancouver Island	fall	Big Qualicum River	2010	95
CAN	Lower Fraser	fall	Squawkum River	2010	95
CAN	Lower Fraser	fall	Hopedale Slough	2011	95
CAN	West Vancouver	fall	Nitinat River	2010	95
CAN	Mainland British Columbia	fall	Skeena River (Kitwanga R.)	2006	83
CAN	Mainland British Columbia	fall	Skeena River (Kitimat R.)	2006	94
CAN	Vancouver Island East	fall	Cowichan River	2010	95
USA	Hood Canal	fall	Big Beef Creek	2010, 2011	78, 17
USA	Hood Canal	fall	Lilliwaup River	2011	91
USA	Hood Canal	summ	Hamma Hamma River	2010	58
USA	Hood Canal	summ	Jimmycomelately Creek	2001	95
USA	Central Puget Sound	fall	Grovers Creek/Chico Creek	2010	14, 81
USA	North Puget Sound	fall	North Fork Nooksack River	2010	95
USA	North Puget Sound	fall	Skagit Mainstem River	2010	50
USA	North Puget Sound	fall	North Fork Stillaguamish	2010	95
USA	North Puget Sound	fall	Snohomish River	2010	95
USA	South Puget Sound	fall	Skookum/Mill Creek	2010, 2011	69, 26
USA	South Puget Sound	fall	Kennedy Creek	2010, 2011	44, 51
USA	South Puget Sound	fall	Sherwood River	1994	95
USA	South Puget Sound	summ	Sherwood River	1994	95
USA	South Puget Sound	winter	Nisqually River Hatchery	2004	95
USA	South Puget Sound	winter	Puyallup Hatchery	2011	95
USA	Strait of Juan de Fuca	fall	Elwha River	2004	95
USA	Washington Coast	fall	Quinault National Fish Hatchery	2001, 2008, 2010	285 (95, 95, 95)
USA	Washington Coast	fall	Satsop River	1998	95
USA	Columbia River	fall	Hamilton Creek	2004	95

Table 2. List of additional Southern British Columbia-Puget Sound stocks to be analyzed by Canada and US for the Chum SNP baseline under this proposal. If any populations were not able to be added (e.g., due to sample quality or availability issues; see lines with 'NA' in the top section of Table 2), additional nearby populations were added in replacement, as shown in the full list of genotyped populations in Table 3.

Country	run group	region	Collections	Target N	Post-QC Year(s)	Post-QC
CAN	fall	Fraser	Whonnock	100	2013	66
CAN	fall	Fraser	Silverdale Cr	100	2004,2009	99
CAN	fall	Fraser	Chilliwack R	100	2004	73
CAN	fall	Fraser	Alouette	100	1991,2004(N)	91
CAN	fall	Fraser	Harrison	100	2002,2012	75
CAN	fall	Fraser	Widgeon Slough	100	2004	89
CAN	fall	Fraser	NorrishWorth	100	2004	79
CAN	fall	Fraser	Kawkawa Cr	100	2012,2014	55
CAN	fall	Fraser	Serpentine	100	2004	32
CAN	fall	Fraser	Kanaka Cr	100	2004,2005,20	83
CAN	fall	Fraser	Harrison late	100	2006(Late)	97
CAN	fall	Fraser	Street Cr	100	2012,2014	53
CAN	fall	Fraser	Silverhope Cr	100	2013	60
CAN	fall	GeorgiaSt	Tzoonie	100	2008	85
CAN	fall	GeorgiaSt	Orford Su (summer)	100	2003	88
CAN	fall	GeorgiaSt	Stawamus Su	100	2004	43
CAN	fall	GeorgiaSt	Homathko	100	2004	94
CAN	fall	GeorgiaSt	Englishman R	100	2010,2011	56
CAN	fall	JohnstoneSt	Ahta (summer)	100	2002,2003,20	21
CAN	fall	JohnstoneSt	Kakweiken (summer)	100	NA	NA
CAN	fall	JohnstoneSt	Ahnuhati (summer)	100	2005,2006	68
CAN	fall	WCVI	Wanokana	100	NA	NA
CAN	fall	WCVI	Colonial	100	2002	36
CAN	fall	WCVI	Deserted	100	2008	37
CAN	fall	WCVI	Tranquil Cr	100	2010	69
US	fall	Strait of Juan de Fuca	Dungeness River	95	NA	NA
US	summe	Hood Canal	Union	95	NA	NA
US	summe	Hood Canal	Lilliwaup	95	2002,2010	101
US	summe	Hood Canal	Quilcene	95	1997,2000	73
US	summe	Hood Canal	Duckabush	48	NA	NA
US	summe	Hood Canal	Dosewallips	95	NA	NA
US	summe	Strait of Juan de Fuca	Salmon	95	2000	95
US	fall	Hood Canal	North Fork Skokomish	95	NA	NA
US	fall	Hood Canal	Hamma Hamma	95	NA	NA
US	fall	Hood Canal	Mission	95	2010	26
US	fall	Hood Canal	Hoodsport Hatchery	95	2003	94
US	fall	Hood Canal	Dewatto	95	NA	NA
US	fall	North Puget Sound	Lower Sauk	95	2010,2013	39
US	summe	South Puget Sound	Blackjack	60	NA	NA
US	summe	South Puget Sound	Johns	55	NA	NA
US	fall	Columbia River	Grays	95	NA	NA

Table 3. Regions, conservation unit (CU) number, populations, years, and sample sizes for all populations in the chum salmon SNP baseline south of Smith Inlet, inclusive. Note: this table contains only those samples that passed quality control. This table contains many populations not specifically proposed in this proposal, but are included for completeness of this documentation. These are all the same populations that are found in the dendrogram above.

Region/Conservation Unit	CU Number	Population	Years	N
Smith Inlet	CU-12	Draney_Cr	2005, 2009	72
		Lockhart_Gordon_Cr	2005, 2006, 2009	81
		Nekite_R	1989, 2000, 2008	117
		Walkum_Cr	2004, 2005	57
Northwest Vancouver Island	CU-11	Cayeghle_R	2003, 2004	86
		Colonial_Cr	2002	36
		Goodspeed_R	2002	21
		Pegattem_Cr	2002	53
Southwest Vancouver Island	CU-10	Bedwell_R	2010	51
		Black_Cr_WCVI	2010	25
		Burman_R	2008	29
		Canton_Cr	2010, 2011, 2012	97
		Conuma_R	2011, 2012	108
		Hoiss_Cr	2010	27
		Kaouk_R	2010	47
		Little_Toquart_Cr	2010	69
		Lord_Cr	2010	29
		Nahmint_R	2003	35
		Nitinat_R	2004, 2010	119
		Park_R	2010	25
		Salmon_Cr_WCVI	2010	50
		Sarita_R	2017	22
		Sucwoa_R	2011, 2012	106
		Sugsaw_Cr	2004	92
		Tlupana_R	2012	104
		Tranquil_Cr	2010	69
		Warn_Bay_Cr	2010	39
Upper Knight	CU-09	Ahnuhati_R	2005, 2006	68
		Klinaklini_R	2002	95
Southern Coastal Streams	CU-08	Ahta_R	2002, 2003, 2015	21
		Mackenzie_Sound_Cr	2004, 2016	20
		Taaltz_Cr	2016	24
		Viner_Scott_Cove	2006, 2007	107

		Viner_Sound	2006, 2008	129
Bute Inlet	CU-07	Algard_Cr	2003	63
		Homathko_R	2004	94
		Orford_R	2003	88
		Southgate_R	2003, 2004	59
Loughborough	CU-06	Glendale_Cr	2003, 2004	75
		Heydon_Cr	2001, 2011, 2018	98
		Phillips_R	2004, 2006, 2011, 2012	98
		Wortley_Cr	2002	48
Northeast Vancouver Island	CU-05	Nimpkish_R	2010, 2011	103
Georgia Strait	CU-04	Campbell_R	2011, 2018	184
		Chemainus_R	1997, 2018	59
		Cold_Cr	2002	20
		Cowichan_R	2000, 2010, 2011, 2018	108
		Demamiel_Cr	1992	50
		Deserted_R	2008	37
		Englishman_R	2010, 2011	56
		Goldstream_R	2011, 2018	184
		Lang_Cr	2008, 2009, 2011	65
		Lit_Qualicum_R	1991, 2009, 2010, 2018	190
		Myrtle_Cr	2010, 2011	27
		Nanaimo_R	2010, 2018	85
		Okeover_Cr	2013	82
		Puntledge_R	1991, 2010	193
		Qualicum_R	2010	179
		Sliammon_Cr	2008	59
		Snake_Bay_Cr	2010, 2011	89
		Theodosia_R	2013	90
		Tzoonie_R	2008	85
Howe Sound-Burrand Inlet	CU-03	Cheakamus_R	2012	75
		Indian_R	2012	89
		Mamquam_R	1991, 2004, 2008	98
		Mashiter_Cr	2004, 2007	45
		Shovelnose_Cr	2012	80
		Squamish_R	2002, 2003	69
		Stawamus	2004	43
Lower Fraser	CU-02	Alouette_N	2004	54
		Alouette_R	1991	37

		Barnes_Cr	2011, 2012, 2014	32
		Blaney_Cr	2011, 2017	59
		Chehalis_R	1991, 1992	97
		Chilliwack_R	2004	73
		Chilqua_Cr	2004, 2017	103
		Harrison_R	2002, 2012	75
		Harrison_R_Late	2006	97
		Hicks_Cr	2005, 2013, 2014, 2017	69
		Hopedale_Cr	2013, 2014	40
		Hunter_Cr_FR	2012, 2014	73
		Inch_Cr	2017	92
		Kanaka_Cr	2004, 2005, 2011	83
		Kawkawa_Cr	2012, 2014	55
		Lillooet_R_Lower	2002	55
		MacIntyre_Cr	2011	51
		Norrish_Worth_Cr	2004	79
		Peach_Cr	2009, 2011	72
		Railroad_Cr	2012, 2014	65
		Serpentine_R	2004	32
		Silverdale_Cr	2004, 2009	99
		Silverhope_Cr	2013	60
		Squawkum_Cr	2009, 2010	123
		Street_Cr	2012, 2014	53
		Sweltzer_R	2008	62
		Vedder_R	2002, 2003	67
		Wahleach_Cr	1991	50
		Whonnock	2013	66
		Widgeon_Cr	2004	89
		Worth_Cr	2013, 2014	54
North_Puget_Sound	WA-02	County_Line_Cr_Fall	1994	60
		Grant_Cr_Fall	2003	50
		Nooksack_R	1998, 2011	60
		Sauk_R	2010, 2013	39
		Siberia_Cr_Fall	1993	36
		Skagit_R	2013	46
		Skykomish_R	2007	85
		Snohomish_R	2012	76
		Stillaguamish_R	2012	76
Central_Sound	WA-07	Chico_Grovers_Hatch	2010, 2011, 2012	140
		Green_R_Keta_Hatchery	2007	92
Tulalip	WA-04	Enetai_Cr_Fall	2014	90

		McKernan_Hatch_Fall	2014	69
		Tulalip	2003	80
South_Puget_Sound	WA-01	Kennedy_Cr	2003, 2010, 2011	95
		Mill_Cr	2010, 2011	56
		Minter_Cr	2003	93
		Nisqually_R	2011	96
		Nisqually_R_Winter	2012	89
		Puyallup_R	2012	82
		Skookum_Cr_Fall	2010	56
		South_Prairie_Cr	2011, 2014	34
Hood Canal	WA-03	Big_Beef_Cr_Fall	2010, 2011	75
		Big_Mission_Cr	2010	26
		Big_Quilcene_R_Summer	1997, 2000	73
		Hoodsport_Fall	2003	94
		Jimmycomelately_Cr	2002	23
		Lilliwaup_Cr	2002, 2010	101
		Salmon_Cr_HOOD	2000	95
		Spencer_Cr	2010	30
Juan de Fuca	WA-06	Elwha_R_Fall	1995	61
Coastal_Washington	WA-05	Bitter_Cr	2000	87
		Ellsworth_Cr	2000	56
		Quinault_R	1998	81
		Satsop_R_Fall	1998	84
Columbia River	WA-08	Ives_Is_Hamilton_to_Washougal_Hatchery	2002	24

Table 4. Results of 100% simulations showing percent accuracy to collection and repunit for the baseline that will be used to separate Chilliwack River and similar populations from the other populations in the Fraser River. Improvements may be gained for the Chilliwack repunit by adding additional samples in upcoming years' sampling.

collection	sample size	repunit	avg. correct to collection	avg. correct to repunit
Hopedale_Cr	40	Chilliwack	4.2%	76.1%
Vedder_R	67	Chilliwack	40.5%	73.3%
Chilliwack_R	73	Chilliwack	27.1%	66.8%
Sweltzer_R	62	Chilliwack	31.9%	52.8%
Peach_Cr	72	Chilliwack	27.0%	52.6%
Street_Cr	53	Chilliwack	12.7%	50.7%
Alouette_R	37	Fraser	4.0%	99.8%
Kanaka_Cr	83	Fraser	13.0%	99.6%
Alouette_N	54	Fraser	14.1%	99.4%
Inch_Cr	92	Fraser	68.7%	99.1%
MacIntyre_Cr	51	Fraser	15.2%	99.0%
Chilqua_Cr	103	Fraser	77.2%	98.5%
Harrison_R_Late	97	Fraser	31.0%	98.4%
Widgeon_Cr	89	Fraser	17.2%	98.2%
Railroad_Cr	65	Fraser	13.9%	97.7%
Whonnock	66	Fraser	24.7%	97.4%
Wahleach_Cr	50	Fraser	19.2%	97.3%
Silverhope_Cr	60	Fraser	51.0%	96.9%
Harrison_R	75	Fraser	21.7%	96.0%
Silverdale_Cr	99	Fraser	22.0%	95.2%
Lillooet_R_Lower	55	Fraser	56.1%	93.1%
Barnes_Cr	32	Fraser	20.6%	92.1%
Blaney_Cr	59	Fraser	32.3%	92.0%
Hunter_Cr_FR	73	Fraser	50.8%	91.3%
Squawkum_Cr	123	Fraser	39.5%	91.0%
Serpentine_R	32	Fraser	0.6%	89.2%
Worth_Cr	54	Fraser	19.7%	88.9%
Hicks_Cr	69	Fraser	26.4%	86.5%
Norrish_Worth_Cr	79	Fraser	7.5%	86.4%
Chehalis_R	97	Fraser	39.3%	83.8%
Kawkawa_Cr	55	Fraser	14.5%	59.7%

Table 5. Results of 100% simulations assigning correctly back to the population (Self-assignment) or showing the population that received the most assignment for the simulation (e.g., simulated Little Qualicum River had a majority of the assignments go to Qualicum River).

Population	Self-assignment	Top assign pop	Top assign	Grouping
Puntledge_R	44.4%	Puntledge_R	44.4%	GStr-SFj
Qualicum_R	36.0%	Qualicum_R	36.0%	GStr-SFj
Lit_Qualicum_R	26.4%	Qualicum_R	32.4%	GStr-SFj