Canadian Science Publishing

Canadian Journal of Fisheries and Aquatic Sciences

Dense SNP panels resolve closely related Chinook salmon populations

Journal:	*Canadian Journal of Fisheries and Aquatic Sciences*
Manuscript ID	cjfas-2019-0067.R1
Manuscript Type:	Article
Date Submitted by the Author:	23-Jul-2019
Complete List of Authors:	McKinney, Garrett; University of Washington, School of Aquatic and Fishery Sciences Pascal, Carita; University of Washington, School of Aquatic and Fishery Sciences Templin, William; Alaska Department of Fish and Game, Gilk-Baumer, Sara; Alaska Department of Fish and Game, Gene Conservation Laboratory 333 Raspberry Road Anchorage, AK, USA 99518 (907) 267-2201 Dann, Tyler; Alaska Department of Fish and Game, Gene Conservation Laboratory Seeb, Lisa; University of Washington, School of Aquatic and Fishery Sciences Seeb, James; University of Washington, School of Aquatic and Fishery Sciences
Keyword:	Chinook salmon, genetic stock identification, amplicon sequencing, RADseq, GT-seq
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

Dense SNP panels resolve closely related Chinook salmon populations

Garrett J. McKinney[1a*], Carita E. Pascal[1b], William D. Templin[2c], Sara E. Gilk-Baumer[2d], Tyler H. Dann[2e], Lisa W. Seeb[1f], James E. Seeb[1g]

[1] School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle WA 98195-5020, USA.

[2] Alaska Department of Fish and Game, 333 Raspberry Road, Anchorage, AK, USA 99518

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

.Emails: [a] gjmckinn@uw.edu; [b]cpascal@uw.edu; [c] bill.templin@alaska.gov; [d] sara.gilk@alaska.gov; [e]tyler.dann@alaska.gov; [f]lseeb@uw.edu; [g]jseeb@uw.edu

Running head: SNP markers for Chinook salmon

Key words: Chinook salmon, genetic stock identification, amplicon sequencing, RADseq, GT-seq,

Abstract

Chinook salmon are migratory fish that are highly valued for subsistence, sport, and commercial fisheries throughout their native range. Populations of Chinook salmon in Western Alaska have exhibited long-term declines, leading to restrictions on harvests. Management priorities require greater resolution for genetic stock identification (GSI) than is available with current methods. We leveraged RADseq, TaqMan, and GT-seq data originating from multiple sources, collected through time, to develop a set of GT-seq panels containing 1,092 SNPs that improved GSI resolution in Western Alaska for at-sea and in-river sampling. We generated a dense linkage map with to ensure that markers selected for panels spanned the entire genome. In addition, we identified multiple RADseq markers that were associated with sex; these aligned to a 5cM region on the sex chromosome. Finally, we developed a bioinformatic pipeline to streamline analysis of GT-seq data that is capable of genotyping microhaplotypes and paralogs, both of which can improve GSI resolution over traditional single-SNP data. Our panels and pipeline provide tools for management agencies to rapidly and easily analyze large-scale genotyping projects.

Introduction

Genomic data have become a central feature of management and conservation of fish populations (e.g., Bernatchez et al. 2017, Sylvester et al. 2017). Conservation applications enabled by genome-wide data sets include studies of adaptation, genotype-by-environment interactions, inbreeding and outbreeding depression, or loss of adaptive variation (Allendorf et al. 2010). Genomic data enable traceability of escapees from fish farms (Pritchard et al. 2016, Holman et al. 2017), important because such escapees provide a major threat to genetic variability and sustainability of wild populations (Bolstad et al. 2017, Forseth et al. 2017). Finally, effective harvest management of migratory species such as salmon requires accurate identification of unique populations as they mix along their migratory corridors (Dann et al. 2013, Meek et al. 2016), and SNP panels identified from genome-wide data sets frequently provide management and conservation solutions (McKinney et al. 2017, Beacham et al. 2018).

It is now relatively easy to generate thousands of markers to address conservation and management questions either through whole-genome or reduced representation sequencing; however, high-throughput genotyping of thousands of samples is still prohibitive in terms of cost and time. One solution is to distill these datasets to subsets of informative loci that can be genotyped using amplicon sequencing. This allows a rapid and cost-effective method for genotyping thousands of individuals for hundreds of loci.

Chinook salmon are migratory fish that are important for ceremonial, subsistence, sport, and commercial fisheries throughout their native range in Pacific Rim drainages from the Kamchatka Peninsula, in the western Pacific Ocean, to Central California, USA, in the eastern Pacific Ocean (Healey 1991). The species is culturally significant among indigenous tribes; some celebrate the first Chinook salmon caught each year with spiritual ceremonies. Their large size (up to 50 kg) and fighting ability make Chinook salmon a prized sportfish, and their size and flesh quality make them a highly valued commercial and sport fish. These factors have collided to complicate management and allocation of migrating Chinook salmon among user groups (Miller 1993, Lin et al. 1996, Gisclair 2009) and among nations (deReynier 1998, Walsh 1998).

Many populations of Chinook salmon have exhibited long-term declines throughout their native range (e.g., Schoen et al. 2017, Siegel et al. 2017). Population declines in the Eastern Bering Sea have led to restrictions on both subsistence and commercial harvest, and individuals are becoming smaller and younger at maturity (Ohlberger et al. 2018). These conservation challenges are compounded by the fact that many fisheries harvest multiple populations (i.e., mixed stock fisheries) of Chinook salmon of differing abundances. Effective management requires the ability to identify the components of individual stocks harvested in mixtures so that less productive stocks can be protected and harvest can target concentrations of productive stocks (Beacham et al. 2008). In addition to management of mixed stock fisheries, a main driver for distinguishing stocks of Western Alaska Chinook salmon is the desire to characterize the composition of bycatch in the walleye pollock fishery (Templin et al. 2011); up to 60% of the Chinook salmon bycatch can originate from Western Alaska Chinook salmon (Myers and Rogers 1988, Myers et al. 2009).

Populations of Chinook salmon throughout the coastal areas of the eastern Bering Sea show little genetic differentiation when genotyped by traditional panels of up to 96 Single Nucleotide Polymorphisms (SNPs) and subdivide into only five genetically identifiable reporting groups for mixture analysis or genetic assignment tests (Larson et al. 2014a). These reporting groups consist of Norton Sound and three reporting groups related to differentiation within the Yukon River (Upper Yukon River, Middle Yukon River, Lower Yukon River); the final reporting group combines all populations from the Kuskokwim Drainage/Bristol Bay region. Analyses that genotype 1000s of SNPs demonstrate that larger panels of SNPs show promise for further resolving these populations (restriction site associated DNA sequencing or RADseq; Larson et al. 2014b, see also Sylvester et al. 2017). At the same time, techniques have emerged that enable high-throughput genotyping of 100s of information-rich SNPs (e.g., GT-seq; Campbell et al. 2015), suitable for conservation and management applications. Our objective was to use RADseq to discover and screen 1000s of SNPs to find informative markers and then use them to construct GT-seq panels to provide further resolution of Chinook salmon populations in Western Alaska.

We report the development and testing of a set of marker panels containing a total of 1,092 existing and newly ascertained SNPs for discriminating stocks of Chinook salmon. Existing SNPs, already tested in other applications, originated from three sources: a 299-SNP GT-seq panel adapted for use with Columbia River, USA, populations (Hess et al. 2016) and developed in part from SNPs ascertained in Western Alaska (Smith et al. 2005a, Smith et al. 2005b, Smith et al. 2005c); a 96 SNP TaqMan panel developed for genetic stock identification in Western Alaska (Larson et al. 2014a); and 178 RADseq SNPs developed for population discrimination in Cook Inlet, Alaska (Dann et al. 2018). Newly ascertained SNPs originated from a dense RADseq screen of individuals from populations spanning the Kuskokwim and Nushagak river drainages in Western Alaska. The final set of 1,092 SNPs increased resolution for identifying components of potential harvest mixtures, providing a total of eight reporting groups for Chinook salmon in Western Alaska where only five existed before. In addition, simulations of sampling local fisheries supported the discrimination of three reporting groups each within the Kuskokwim and Nushagak drainages.

Finally, we developed a computational pipeline for processing and genotyping of GT-seq data that is able to genotype multiple-SNP haplotypes and markers with different levels of ploidy. Data handling is increasingly challenging with increasing numbers of SNPs; early interest suggests that the computational pipeline will expedite processing of GT-seq, RAPTURE (Ali et al. 2016), and similar data sets.

Our work demonstrates how an increasingly large number of SNPs can be collated and evaluated for high throughput analyses; we also present recommendations for further streamlining the discovery and evaluation phases that should be widely applicable to species of management and conservation interest. These new SNP panels will aid in management of Alaska Chinook salmon by improving the ability of managers to discriminate among stocks of interest during harvest. Lessons learned in this study will be widely applicable to other studies where genetic stock identification is applied to mixtures of migrating populations.

Materials and Methods

RADseq data

We conducted SNP discovery on RADseq data from 13 populations (Figure 1, Figure 2A-C; Figure 3; Table 1). Raw data for five of the populations were available from previous studies (Larson et al. 2014b (2 populations), McKinney et al. 2018 (3 populations)) and downloaded from Dryad (doi:10.5061/dryad.rs4v1) and NCBI (SRA SRP129894) (Table 1). Additional RAD sequencing was conducted on 48 fish per population from 8 new populations (Table 1). In total, nine populations from the Kuskokwim drainage and four populations from Bristol Bay (Togiak River and Nushagak drainage) were used for SNP discovery. Full materials and methods for RAD sequencing are available in supplemental file S1.

We upgraded the McKinney et al. (2016) genetic map for Chinook salmon to provide a framework to help ensure that markers chosen for our panels were distributed across the genome. Genetic maps improve the power to identify genes and gene regions important in population differentiation and adaptation. The McKinney et al. (2016) map was a consensus of maps derived from populations from Puget Sound, Washington, and likely missed variation present in northern stocks (see Templin et al. 2011). We mapped additional variation detected in five haploid families that were created from females sampled from Ship Creek (Anchorage, Alaska) to produce a new consensus map (supplemental file S2).

Sex Locus

We attempted to incorporate loci that could be used to determine the sex of immature fish into our SNP panels. Knowledge of sex of individual fish and sex ratios in migrating cohorts provides useful demographic information, but identification of reliable DNA sex markers has been elusive (Von Bargen et al. 2015). Sex data were only available from 23 fish from the Togiak River, so putative sex-associated loci were instead identified by examining population-level genotype patterns. In salmon, males are the heterogametic sex and females are the homogametic sex (Thorgaard 1977), and in Chinook salmon the sex-determining region is on chromosome 17 (Ots17, Phillips et al. 2013). Assuming equal numbers of males and females in a population, a sex locus should be revealed by the presence of high heterozygosity (~50%) but only two genotypic classes, one heterozygous and one homozygous, with a resulting minor allele frequency of 0.25. We identified putative sex-associated loci as those with heterozygosity between 45% and 55%, minor allele frequency between 0.20 and 0.30. Up to 5% of individuals with the alternate homozygous genotype were allowed to account for genotyping error. Putative sex loci were placed on the linkage map (below) to determine if they co-locate with a previously identified sex-associated locus from Chinook salmon originating from the Marblemount Hatchery, Skagit River drainage, in Washington State (University of Washington, unpublished data).

GT-seq panel construction

The populations examined in this study collapsed into a single reporting group with marker panels currently used for management. We simulated mixed stock analyses from the RADseq dataset using *GSIsim* (Anderson et al. 2008, Anderson 2010) to determine the number of markers necessary to subdivide the existing Kuskokwim/Bristol Bay reporting group into major river drainages. A training holdout leave-one-out (THL) approach was used to minimize bias in assignment accuracy (Anderson 2010). Samples within a population were randomly assigned to training and holdout datasets. The training set was used to choose high F_{ST} loci for panel testing; the holdout set was used to evaluate panel accuracy against the baseline (training plus holdout samples). Simulations were done in intervals of 250 markers, up to 1,500 markers, after which they were done in intervals of 1,000. Simulation results showed that approximately 1,000 markers were necessary to resolve reporting groups into major river drainages. GT-seq panel development targeted a final set of 1,000 SNPs for population analyses.

We chose to partition SNPs into four GT-seq panels of about 300 SNPs each. Experience has shown that up to 300 amplicons is a workable number to optimize panel performance (Beacham et al. 2018, McKinney et al. 2018). Also, compartmentalizing SNPs allowed us to potentially capture existing data genotyped on other platforms and shape modular panels for downstream applications on subsets of SNPs.

A panel of 299 existing SNPs (Panel 1) was available from Idaho Fish and Game that was originally developed by the Columbia River Inter-Tribal Fish Commission (CRITFC, Hess et al. 2016). This panel includes the majority of the 192 loci from Warheit et al. (2013) that form the basis for data collection for Pacific Salmon Commission applications from California to Southeast Alaska (e.g. Clemento et al. 2014).

Panels 2-4 were developed by this study. Panel 2 originated from other currently available SNPs, either in use by Alaska Department of Fish and Game (ADF&G) or proposed for use by ADF&G in Cook Inlet, Alaska (Dann et al. 2018) to facilitate comparison to pre-existing data (including the $F_{ST}96$ SNPs of Larson et al. (2014a) that were ascertained in western Alaska). Panels 3 and 4 originated from novel RADseq loci ascertained in this project to specifically improve resolution within and between the closely related populations in Kuskokwim and Nushagak river drainages in western Alaska.

Informative markers were identified using both outlier analysis and F_{ST} for panels 3 and 4 (Figure 2C). Outlier loci were identified using *BayeScan* with default settings (Foll and Gaggiotti 2008) as those with Q < 0.05. F_{ST} was estimated using Genepop (Rousset 2008). Initial testing showed many outliers with very low F_{ST}. These loci were likely false positives; they were generally only variable in one or a few populations, and their absence in other populations was likely due to sampling error. Loci were re-filtered to include only loci with a minimum minor allele frequency of 0.05 in three or more populations and then reanalyzed with *BayeScan*.

High F_{ST} markers from both within and between the Kuskokwim and Nushagak river drainages were considered. Excess candidate markers, to allow for dropouts, were chosen as follows (Figure 2C): 1) 500 markers with highest F_{ST} within the Kuskokwim drainage for Panel 3 (F_{ST} range: 0.051-0.005, 2) 500 markers with highest F_{ST} within the Nushagak drainage and Togiak River for Panel 4 (F_{ST} range: 0.106-0.006), and 3) 500 highest F_{ST} markers between the Kuskokwim and Nushagak drainages (F_{ST} range: 0.620-0.004) for incorporation into either Panel 3 or Panel 4. Loci were split between panels 3 or 4 based on F_{ST} within regions; panel 3 contained loci with greater F_{ST} in the Kuskokwim Bay/River while panel 4 contained loci with greater F_{ST} in Togiak Bay/Nushagak drainage. Outlier loci identified using *BayeScan* were contained within the high F_{ST} locus set.

These 1500 loci underwent additional filtering to meet design criteria for GT-seq analysis (Figure 2D). Loci with SNPs within 16 bp of the 3' end and 20 bp of the 5' end of the RAD locus were excluded to allow room for primer design. Where matches were available, paired-end contigs from Larson et al. (2014b) were used to extend the 5' end of the RAD locus for primer design. Loci with low complexity or transposable element sequence were identified using RepeatMasker (Smit et al. 2013) and removed. Primers were designed using

Primer3 (You et al. 2008) with default settings for loci that passed initial filters. Primers were then aligned to all loci using bowtie2 (Langmead and Salzberg 2012) to identify and remove cases where primers may amplify multiple loci. A total of 706 loci passed filters and were retained for panel optimization; 350 loci for panel 3 and 356 loci for panel 4. In combination with panels 1 and 2, a total of 1,343 loci were passed forward for the initial optimization (Figure 2D).

GT-seq Genotyping

For panel optimization and population genotyping we developed a bioinformatic pipeline, *GTscore* (https://github.com/gjmckinney/GTscore) to score both multiple SNP haplotypes (also referred to as microhaplotypes; Baetscher et al. 2018) and duplicated loci; both of these locus types have been shown to increase power for resolving closely related populations (Limborg et al. 2017, McKinney et al. 2017, Waples et al. 2017) and were included in the GT-seq panels. Multi-SNP haplotypes occur when multiple SNPs are in the same sequence tag, resulting in haplotypes with > two alleles. This pipeline incorporates the *polyGen* algorithm (McKinney et al. 2018) that uses a maximum likelihood method for genotyping and is capable of genotyping loci with any number of alleles and any level of ploidy.

Panel Optimization

Optimization of Panel 2, Panel 3, and Panel removed loci that did not amplify properly in the PCR reaction and perfected PCR performance for each of the retained primer pairs (Figure 2 E). Panel 1 had already undergone extensive optimization by Hess et al. (2016) but was included in panel optimization to keep PCR conditions consistent for the final sequencing run and ensure there were no interactions with loci from other panels.

Optimization was done with 100bp paired-end sequencing in two rounds on an Illumina MiSeq (Figure 2E). DNA was extracted, and sequencing libraries prepared following the methods of Campbell et al. (2015). The first round of sequencing used 48 individuals from four populations (Kogrukluk, Koktuli, Necons, and Togiak rivers). Primer performance was evaluated in this first round (see below); however, read depth for most loci was too low to allow accurate genotyping. The sample size was reduced in the second round of sequencing to 24 samples from two populations (Kogrukluk and Koktuli rivers, 12 samples each) to ensure adequate read depth to evaluate genotype concordance between RADseq and GT-seq.

After each round of sequencing, we eliminated loci that were over-amplifiers, off-target amplifiers, or cross-amplifiers. Over-amplifying loci generate excessive sequences relative to other loci in the panel and can be identified by examination of ranked number of sequence reads per locus. Off-target amplifiers generate sequences that do not match the target sequence; these sequences contain either the forward primer or the reverse primer (sometimes both) but do not contain the bioinformatic probe that identifies allelic variation within the target sequence. Total read counts for each locus were used to identify over-amplifiers, and counts of primer and probe alignments were used to identify off-target amplifiers. Paired-end sequencing generates sequence reads from both ends of a DNA sequence. With GT-seq, both reads in a pair should be from the same locus, and the R1 read should start with the forward primer while the R2 read should start with the reverse primer. Cross-amplification occurs when a product is amplified by the forward and reverse primers of two different loci. This could occur when multiple loci are physically close or when multiple regions of the genome are genetically similar. Cross-amplifiers are identified where the R1 and R2 reads of an amplicon align to different loci. We used a custom pipeline for identifying cross-amplifiers. Reference sequence for each locus was generated by trimming the RAD consensus sequence to contain only the sequence between each primer. For each individual, GT-seq sequences were aligned to the reference sequence using *GATK* (McKenna et al. 2010) and *SAMtools* (Li et al. 2009). The resulting alignments were processed with custom perl scripts to quantify cross-amplification per locus. Patterns of cross-amplification were visualized using network plots in R.

Validation of genotyping accuracy is necessary when combining results between different technologies because loci do not always genotype consistently across technologies. After the second round of sequencing, we excluded loci with discordant genotypes that could not be explained by low read depth in either the RADseq or GT-seq genotyping. Extensive cross validation between RADseq and single SNP data (assayed by 5'-nuclease reaction with TaqMan chemistry) genotypes was already done by Hess et al. (2016) for loci in Panel 1) and Larson et al. (2014a) for loci in Panel 2.

Baseline Data Set for Performance Testing and Mixture Analyses

For final testing and mixture analyses, we prepared a baseline data set that targeted a sample size of 95 in each of 17 major populations, spanning the Kuskokwim drainage, Kuskokwim Bay, and Bristol Bay (Nushagak drainage and Togiak River) (Table 1, Figure 3), for all of the GT-seq loci that passed filters. We took three genotyping steps to accomplish this: (1) we added TaqMan data to the RADseq data available for the 48 individuals in the original 13 populations used for RADseq discovery (to account for TaqMan-origin loci present in Panel 1 and Panel 2); (2) we used the four GT-seq panels to genotype up to 49 additional individuals in each of the 13 discovery populations to approach the target sample size; and (3) we used the four GT-seq panels to genotype 95 individuals in each of four new populations.

TaqMan methods were identical to those described in (Larson et al. 2014a), and GT-seq sequencing as described above was conducted on an Illumina HiSeq 4000 with 1,190 loci and 270 samples per lane. The RADseq, TaqMan, and GT-seq datasets were combined and filtered in R. Samples shared between datasets allowed further cross-validation of genotypes, ensuring genotype concordance across datasets.

A final GT-seq filtering step, examining allele ratio plots, was conducted following genotyping of all samples (Figure 2F). A histogram of allele ratios was plotted for each SNP for visual examination. Singleton loci should have up to three peaks, depending on allele frequency, that are centered at 0, 0.5, and 1 (Figure S1A). Duplicate loci should have up to five peaks centered at 0, 0.25, 0.5, 0.75, and 1 (Figure S1B). Diverged duplicate loci should have up to three peaks either centered on 0, 0.25, and 0.5 or on 0.5, 0.75, and 1 (Figure S1C). Loci that did not display distinct peaks associated with each genotype class are likely amplifying off-target sequence (Figure S1D); we attempted to recover these loci by extending the bioinformatic probe to exclude off-target sequence. Loci that could not be recovered were removed from further analysis.

A final filtering step prior to mixed-stock analysis (MSA) was necessary due to the nature of the combined data and the fact that some data types were not present or scorable in the original RADseq data from the 13 populations. Loci excluded for these reasons could be included in analysis of future GT-seq only datasets. Duplicate and diverged duplicate loci were removed prior to MSA evaluation because they could not be reliably genotyped in the RADseq data due to inadequate read depth (McKinney et al. 2018). Allele frequencies of RADseq and GT-seq data were also compared, and loci with allele frequency discrepancies were removed. Loci were also removed if their genotype rate was less than 70% or if they were monomorphic. Following locus filtering, samples with a genotype rate less than 90% were removed prior to MSA evaluation.

Modelling Mixed Stock Analysis

The potential resolution of MSA was assessed using the full RADseq dataset while the accuracy of the four GT-seq panels for MSA was assessed using the combined RADseq and GT-seq sample set. For loci that contained multiple SNPs, haplotypes were used to further improve accuracy (McKinney et al. 2017). Populations were divided into reporting groups based on genetic affinities and management objectives. The desired reporting groups included Upper Kuskokwim River, Kuskokwim Bay, Togiak Bay, and Nushagak River. These groups are highly productive, important to stakeholders, are managed separately, and have responded differently to environmental conditions responsible for recent declines throughout this region. Mixture analysis and individual assignment was performed in *GSIsim* to explore alternative reporting group configurations including region-wide and in-region MSA. Mixture analysis included 100% simulations to evaluate correct allocations of population to reporting group of origin and accuracy and precision among reporting groups. Individual assignment was conducted at the population level, with individuals assigned to a population if they met a threshold of 80% probability assignment to that population; an 80% threshold has been shown to provide a balance between a low false positive assignment rate and a successful assignment rate (Griffiths et al. 2013).

Results

Sequencing and Mapping Results

We retained 19,435 loci that were scored in 761 individuals to evaluate utility for population discrimination. Final sample size per population ranged from 32 to 56 with 46-48 in most populations (Table 1). Population pairs showed low overall F_{ST} (average 0.003) consistent with previous studies in this region (Templin et al. 2011). Results for all pairwise F_{ST} comparisons are listed in Table S1.

Three putative sex loci were identified (Table 2). Previously, we identified a sex-associated RAD locus (*RAD93920*) for Chinook salmon in the Marblemount Hatchery in Washington State within the Pacific Northwest of North America (University of Washington, unpublished data); this locus was located 8 cM from the centromere of *Ots17* which is the sex chromosome in Chinook salmon (Phillips et al. 2013). Two of the three loci identified in this study (*RAD67724, RAD29719*) co-located with the previously identified sex-locus near the centromere of *Ots17* at 4.7 cM and 9.4 cM. The other putative sex locus (*RAD27492*) could not be placed on the linkage map. The accuracy of the three putative sex loci varied from 78% to 91% when compared to observed sex for 23 Togiak River samples (Table 2). The SNP associated with sex in the Marblemount population (*RAD93920*) was invariant in all Alaska populations.

A total of 15,930 loci that were scored in 233 individuals were retained for linkage mapping in the five families. Sample size per family ranged from 31 to 62 (Supplementary File S1). The population and linkage mapping datasets require different filtration steps so there are loci that were unique to each dataset. A total of 12,140 loci were common to both datasets. The Alaska linkage map was 2,874.02 cM long and contained 15,798 loci; 13,084 were singleton (non-duplicated) and 2,714 were duplicate (Supplemental File S2). A total of 7,946 loci mapped in the Alaska families were present on the previous map of McKinney *et al.* (2016) that originated from Washington State; the low degree of shared markers is likely a due to a combination of both the different pools of standing genetic variation in the source populations (Templin et al. 2011) and the low number of families. The combined linkage map was 3,003.36 cM long and contained 23,715 loci; 19,762 loci were singleton and 3,953 loci were duplicate (Table 3; Supplemental File S2). A total of 630 of the newly ascertained RADseq loci that passed filters for inclusion in GT-seq panels (see below) were present on the linkage map. These were distributed approximately evenly across the linkage groups with an average of 18.5 GT-seq markers per linkage group and a range of 4-34 markers per linkage group (Supplemental File S3). Linkage groups differed in size; the proportion of markers on each linkage group that were included in GT-seq panels averaged 2.1% with a range of 1.1%-5.7%.

GT-seq Marker Selection and Optimization

We followed several steps of marker selection leading to GT-seq panel optimization. SNP location within the RAD tag was a limiting factor. Approximately 25% of the 1,500 RAD markers originally selected for panel design had to be discarded; the requirement of 16 bp on the 3' end of the SNP or 20 bases on the 5' end was a major limiting factor for primer design from the original sequence of 94bp in length. For SNPs that were within 20 bp of the 5' end of the sequence, paired-end contigs allowed primer design past the original 94 base limitation. Minor losses of SNPs to the panels were also due to transposable element annotation (4%) and identification as repetitive elements (3%).

A total of 1,343 loci passed initial filtering criteria and were developed into GT-seq loci for panels 2-4 (Figure 2) including 18 of 54 loci identified as outliers by *BayeScan*. Two rounds of test sequencing and optimization followed to remove loci due to over-amplification, excessive off-target sequence or cross-amplification between primers, or genotype discrepancies between RADseq and GT-seq genotyping. A total of 1,204 loci originating from the three new panels and that of Hess et al. (2016) were genotyped for all GT-seq samples (Figure 2F).

GT-seq Genotyping and Standardization of Datasets

Two additional filtering steps were conducted following genotyping. Individuals were removed from the analysis if they likely originated from a different population of origin (detailed in File S1). The genotype data also allowed us to examine the allele ratio plots for fit to expected genotype distributions; a total of 112 loci were removed if allele ratios did not fit the expected distributions (e.g. Figure S1D). The final set of GT-seq panels contained 1,092 loci (Figure 2G).

Because our study comparisons required standardization among three sources of genotypes (RADseq, GT-seq, TaqMan), a final quality control step to combine the datasets prior to MSA was necessary. Loci were excluded if they were absent in one of the datasets (84), were duplicate or diverged duplicate loci (70), exhibited greater than expected allele frequency differences between RADseq and GT-seq datasets (17), exhibited a low genotype rate (34) (Figure S2), or were monomorphic (40) (Table 4). Finally, samples were excluded if they exhibited a low genotype rate (Figure S3). A total of 847 loci and 1,545 individuals were available for evaluation with *GSIsim* following the standardization steps; 14% of these loci contained haplotype data.

Mixture Analysis

We first evaluated the power of the full set of more than 15,000 RAD loci. The full dataset increased resolution for Chinook salmon demonstrating > 95% accuracy to five reporting groups (Figure 3A). Individuals were proportionally assigned to populations and proportional assignments were summed for populations within a reporting group. The Kuskokwim and Nushagak drainages could be resolved with the full RADseq dataset. The Kuskokwim drainage could be split into three reporting groups: Upper Kuskokwim, Kuskokwim River, and Kuskokwim Bay (Figure 3A). Additional reporting groups included a combined Togiak/Goodnews reporting group and a Nushagak drainage reporting group.

We then evaluated the subset of 847 loci in the four GT-seq panels. We evaluated accuracy of four different reporting group scenarios of varying scales based on management objectives. Scenarios include: 1) all populations with reporting groups of Upper Kuskokwim, Kuskokwim River/Bay, Togiak/Goodnews, and Nushagak rivers (fine-scale bycatch scenario, Figure 3B), 2) all populations with reporting groups of Upper Kuskokwim River, combined Kuskokwim/Nushagak, and combined Togiak/Goodnews rivers (broad-scale bycatch scenario, Figure 3C). In-region scenarios include: 3) Kuskokwim drainage populations with reporting groups of Upper Kuskokwim River, Kuskokwim River, and Kuskokwim Bay (Kuskokwim

Bay/River scenario, Figure 3D), and 4) Togiak Bay/Nushagak drainage populations with reporting groups of Togiak, Iowithla/Stuyahok, and Koktuli rivers (Togiak Bay/Nushagak River scenario, Figure 3E). For some scenarios Togiak and Goodnews rivers were combined into a reporting group on the basis of genetic similarity even though they are not part of the same drainage. Reporting groups in each scenario had >90% accuracy with two exceptions, the Nushagak River reporting group in the fine-scale bycatch scenario (72%) and the Kuskokwim Bay reporting group in the Kuskokwim Bay/River scenario (87%) (Figure 3). Mean accuracy and 95% range for mixture estimates under each reporting group scenario are listed in Table S2 for reporting groups as a whole and for populations within reporting groups.

Individual Assignment

Individual assignment using the full RADseq dataset showed >95% self-assignment accuracy for all populations with the exception of Kwethluk and Togiak rivers (Table S3). Individual assignment using the GT-seq panels revealed that upriver populations for both the Kuskokwim and Nushagak drainages had >90% self-assignment accuracy with the exception of the Takotna River population (Table S4, S5, S6). Populations in the lower reaches of the Kuskokwim and Nushagak drainages misassigned to populations both within and between drainages while Togiak and Goodnews rivers tended to assign to themselves or to each other. Pitka Fork, Tatlawiksuk, Necons, Togiak, and Koktuli river populations all exhibited >90% accuracy for either at-sea or in-region sampling.

Discussion

Genotype data are now routinely a centerpiece in the mosaic of tools used by conservation practitioners for population assessment and sustainability planning. Until recently, the limited number of genetic markers available often limited the power of genotype data to resolve populations even though other biological information suggested that more resolution should be possible. The introduction of RADseq and other reduced-representation sequencing (RRS) protocols greatly increased the availability of genetic markers, followed in many situations by improved resolution of closely related populations after genotyping hundreds or thousands of loci (see RADseq data in Larson et al. 2014b, Candy et al. 2015). But the cost, limited analysis pipelines, and relatively slow throughput of RRS data often preclude their use in management situations where data from thousands of individuals (cf., Gilbey et al. 2017) or real-time data (cf., Dann et al. 2013) are needed. Amplicon sequencing approaches promise an intermediate solution where hundreds of loci may be rapidly and cost-effectively genotyped in thousands of individuals, offering conservation alternatives unavailable until now (Campbell et al. 2015, Beacham et al. 2018).

To evaluate resolution of populations we opted to build a baseline data set by adding GT-seq genotypic data to the RADseq data available from SNP discovery efforts from this study and from two previous studies (Larson et al. 2014b, McKinney et al. 2018). This approach appeared to be a practical use of a large amount of existing data when we started; however, two of the GT-seq panels incorporated data originating from TaqMan derived loci, requiring new TaqMan genotyping to backfill those same loci missing in the RADseq data sets. While robust data emerged after careful cross validation, a better and ultimately more cost-efficient choice would have been to regenotype all samples with identical GT-seq panels.

Genetic maps have become increasingly common and can be used as a foundation to integrate genomic resources for gene annotation and population genomic analyses (McKinney et al. 2016). However, genetic maps originating from one lineage or geographic location often don’t include a high proportion of polymorphic markers informative for a distant lineage or geographic region. We were able to leverage existing map resources for Chinook salmon from the Pacific Northwest and add variation from families that originated from Alaska to create a much denser map with much improved coverage for Alaska. As a result, 74% of the markers could be placed on the combined linkage map generated in this study while only 39% could be placed on the previous Washington-based linkage map.

The dense genetic map also allowed us to investigate the location of sex-associated loci. The sex-associated loci identified in this study were located in a 5 cM window of chromosome 17 which has been previously identified as the sex chromosome in Chinook salmon (Phillips et al. 2013). These loci flank a previously identified sex marker from Marblemount Hatchery Chinook salmon (University of Washington, unpublished data). The co-location of sex associated markers from multiple projects suggest that the sex determining gene is near this region of chromosome 17; however, the markers identified herein showed a maximum accuracy of 91% suggesting incomplete linkage between these SNPs and the true sex determining region. This is consistent with results from other molecular markers developed to assign sex to immature Chinook salmon which have displayed inconsistent accuracy when tested in populations throughout the species range (Nagler et al. 2001, Nagler et al. 2004, Chowen and Nagler 2005, Von Bargen et al. 2015).

GT-seq panel

The GT-seq panels increased MSA accuracy for Western Alaska Chinook salmon compared to earlier analyses. We were able to split the Kuskokwim/Bristol Bay reporting group identified by Larson et al. (2014a) into three reporting groups (Upper Kuskokwim, Kuskokwim/Nushagak, Togiak/Goodnews) with >95% accuracy. These panels also show utility for regional population discrimination with three reporting groups for the Kuskokwim Bay/River scenario and three reporting groups for the Togiak Bay/Nushagak River scenario. Individual assignment accuracy was generally >90% for upper river populations in both the Kuskokwim and Nushagak rivers.

A primary goal for fisheries management in the Kuskokwim River/Bristol Bay region is the discrimination of populations from the Kuskokwim and Nushagak drainages. Splitting the Kuskokwim/Nushagak reporting group by drainage yielded MSA accuracy of 93% for Kuskokwim drainage populations but only 72% for Nushagak drainage populations. A potential cause of this inaccuracy is the large difference in sample size between the two reporting groups. The Kuskokwim River reporting group contained eight populations and 728 samples, representing most of the production from the drainage, while the Nushagak River reporting group contained three populations and 252 samples. Differences in sample size can cause bias where some fish from low-sample reporting groups may assign to higher-sample reporting groups (Moran and Anderson 2018).

We were also unable to utilize the full panel in testing MSA accuracy. The panels included paralogs which may increase accuracy of genetic stock identification (Gilbey et al. 2016). We were able to successfully genotype paralogs in the GT-seq portion of the dataset, and the program we used for GSI (*GSIsim*) is capable of including paralogs in analysis. However, paralogs were excluded from MSA evaluation because read depth in the RADseq dataset was too low for reliable genotypes (McKinney et al. 2018). In addition, many loci from Panel 1 were excluded either due to monomorphic genotypes (34) or absence of baseline data (61). Genotyping additional Nushagak River populations would help determine if the low resolution we observed in that drainage is an artifact of unbalanced sampling or a true limitation of the panel. Sequencing of additional baseline samples using GT-seq will also obviate the need to include the RADseq samples in analysis, allowing paralogs to be incorporated as well as the Panel 1 loci missing in the baseline.

It may be possible to improve the efficiency observed in this study by working to design fewer panels with higher resolution loci. The primary source of high-F_{ST} marker loss was the use of SNP position thresholds at both ends of the marker sequence to allow space for primer design. Two Chinook salmon genome assemblies have now been deposited in NCBI (accessions: GCA_002872995.1, GCA_002831465.1); these can be used to design primers for these high-F_{ST} loci that were otherwise lost due to SNP position. Also, the additional power gained from haplotype loci, even when F_{ST} is reduced (McKinney et al. 2017, Baetscher et al. 2018), was recognized after our marker selection had concluded. The current panels include only 114 loci with haplotypes. Locus selection that enriches for haplotypes will allow smaller panels that achieve the same resolution as the current four panels and may yield further increases in resolution.

Recommendations

We leveraged RADseq, TaqMan, and GT-seq data originating from multiple sources, collected through time, to develop a comprehensive baseline of population data. While ultimately successful, this involved considerable effort to standardize datasets and ~20% of the loci were discarded from analysis due to incompatibilities among datasets. Regenotyping all populations using a consistent laboratory method would have been a better solution to develop a standardized dataset.

As genomic resources expand, conservation practitioners are faced with the problem of translating large amounts of available data into useful management applications. Challenges to enabling these applications include reducing the number of markers to a level easily assayed in a cost-effective manner, working with differing baseline datasets and missing loci, developing efficient analysis pipelines, and identifying appropriate accuracy and precision for particular applications. The method we employed in this study, RADseq SNP discovery followed by GT-seq panel development, is a straightforward pathway to develop genetic tools for fisheries management. The panels we developed increased resolution of reporting groups in Western Alaska Chinook salmon, a region with low population structure that has been historically difficult for stock discrimination. These panels offer a cost-effective method for improving genetic stock identification for fisheries management. During the study, we also realized the difficulty in not only handling large amounts of data with increasing numbers of SNPs but also leveraging data from varying laboratory techniques. We anticipate that computational pipelines such as *GTscore* will expedite processing of both RADseq, RAPTURE (Ali et al. 2016), and similar data sets and assist researchers in sharing and standardizing data not only across techniques, but among researchers and laboratories.

Acknowledgements

We thank the many biologists at the Alaska Department of Fish and Game (ADFG) who helped with the design and implementation of the project. They provided encouragement as well as many tissue samples, feedback on SNP selection, and input on analyses. In particular, we thank Chris Habicht, Judy Berger, Nick Decovich, Katie Howard, Chuck Brazil, and the members of the Gene Conservation Laboratory. From the University of Washington, we thank members of the Seeb Laboratory, particularly Wes Larson and Ryan Waples, and Daniel Schindler from the Alaska Salmon Program for insightful discussions and assistance with analyses. We also thank Eric Anderson at the NOAA Southwest Fisheries Science Center (SWFSC) for assistance with *GSIsim* and advice on MSA analyses. This research was partially funded by Alaska Sustainable Salmon Fund awards 44812, 44913, and 44914 from the National Oceanic and Atmospheric Administration, US 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 US Department of Commerce, National Oceanic and Atmospheric Administration, or the Alaska Department of Fish and Game.

References

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.

Allendorf, F.W., Hohenlohe, P.A., and Luikart, G. 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics **11**(10): 697-709.

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.

Baetscher, D.S., Clemento, A.J., Ng, T.C., Anderson, E.C., and Garza, J.C. 2018. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. Molecular ecology resources **18**(2): 296-305.

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.

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.

Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D.T., Barth, J.M.I., Beacham, T.D., Maes, G.E., Martinsohn, J.T., Miller, K.M., Naish, K.A., Ovenden, J.R., Primmer, C.R., Suk, H.Y., Therkildsen, N.O., and Withler, R.E. 2017. Harnessing the Power of Genomics to Secure the Future of Seafood. Trends in Ecology & Evolution **32**(9): 665-680.

Bolstad, G.H., Hindar, K., Robertsen, G., Jonsson, B., Saegrov, H., Diserud, O.H., Fiske, P., Jensen, A.J., Urdal, K., Naesje, T.F., Barlaup, B.T., Floro-Larsen, B., Lo, H., Niemela, E., and Karlsson, S. 2017. Gene flow from domesticated escapes alters the life history of wild Atlantic salmon. Nature Ecology & Evolution **1**(5).

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.

Candy, J.R., Campbell, N.R., Grinnell, M.H., Beacham, T.D., Larson, W.A., and Narum, S.R. 2015. Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, *Osmeridae*), an anadromous Pacific smelt. Molecular ecology resources **15**(6): 1421-1434.

Chowen, T.R., and Nagler, J.J. 2005. Lack of sex specificity for growth hormone pseudogene in fall-run Chinook salmon from the Columbia River. Transactions of the American Fisheries Society **134**(1): 279-282.

Clemento, A.J., Crandall, E.D., Garza, J.C., and Anderson, E.C. 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.

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.

Dann, T.H., Habicht, C., Templin, W.D., Seeb, L.W., McKinney, G.J., and Seeb, J.S. 2018. Identification of genetic markers useful for mixed stock analysis of Chinook salmon in Cook Inlet, Alaska. Alaska Department of Fish and Game, Division of Commercial Fisheries, Anchorage.

deReynier, Y.L. 1998. Evolving principles of international fisheries law and the North Pacific anadromous fish commission. Ocean Development and International Law **29**(2): 147-178.

Foll, M., and Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180**(2): 977-993.

Forseth, T., Barlaup, B.T., Finstad, B., Fiske, P., Gjoaester, H., Falkegard, M., Hindar, A., Mo, T.A., Rikardsen, A.H., Thorstad, E.B., Vollestad, L.A., and Wennevik, V. 2017. The major threats to Atlantic salmon in Norway. Ices Journal of Marine Science **74**(6): 1496-1513.

Gilbey, J., Cauwelier, E., Coulson, M.W., Stradmeyer, L., Sampayo, J.N., Armstrong, A., Verspoor, E., Corrigan, L., Shelley, J., and Middlemas, S. 2016. Accuracy of Assignment of Atlantic Salmon (Salmo salar L.) to Rivers and Regions in Scotland and Northeast England Based on Single Nucleotide Polymorphism (SNP) Markers. PloS one **11**(10): e0164327.

Gilbey, J., Wennevik, V., Bradbury, I.R., Fiske, P., Hansen, L.P., Jacobsen, J.A., and Potter, T. 2017. Genetic stock identification of Atlantic salmon caught in the Faroese fishery. Fisheries Research **187**: 110-119.

Gisclair, B.R. 2009. Salmon bycatch management in the Bering Sea walleye pollock fishery: threats and opportunityies 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.

Griffiths, J.R., Schindler, D.E., and Seeb, L.W. 2013. How Stock of Origin Affects Performance of Individuals across a Meta-Ecosystem: An Example from Sockeye Salmon. PloS one **8**(3): e58584.

Healey, M.C. 1991. Life history of Chinook salmon. *In* Pacific Salmon Life Histories. *Edited by* C. Groot and L. Margolis. UBC Press, Vancouver. pp. 311-394.

Hess, J.E., Campbell, N.R., Matala, A.P., Hasselman, D.J., and Narum, S.P. 2016. Genetic assessment of Columbia River stocks, 4/1/2014-3/31/2105 annual report, 2008-907-00. CRITFC. Available from http://www.critfc.org/wp-content/uploads/2016/04/16-03.pdf.

Holman, L.E., de la Serrana, D.G., Onoufriou, A., Hillestad, B., and Johnston, I.A. 2017. A workflow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon. Aquaculture **476**: 59-64.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9**(4): 357-359.

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*). Evol Appl **7**(3): 355-369.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16): 2078-2079.

Limborg, M.T., Larson, W.A., Seeb, L.W., and Seeb, J.E. 2017. Screening of duplicated loci reveals hidden divergence patterns in a complex salmonid genome. Molecular ecology.

Lin, P.C., Adams, R.M., and Berrens, R.P. 1996. Welfare effects of fishery policies: Native American treaty rights and recreational salmon fishing. Journal of Agricultural and Resource Economics **21**(2): 263-276.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20**(9): 1297-1303.

McKinney, G.J., Seeb, J.E., and Seeb, L.W. 2017. Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences **74**(4): 429-434.

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.

McKinney, G.J., Waples, R.K., Pascal, C.E., Seeb, L.W., and Seeb, J.E. 2018. Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. Molecular ecology resources **18**(3): 570-579.

Meek, M.H., Baerwald, M.R., Stephens, M.R., Goodbla, A., Miller, M.R., Tomalty, K.M.H., and May, B. 2016. Sequencing improves our ability to study threatened migratory species: Genetic population assignment in California's Central Valley Chinook salmon. Ecology and Evolution **6**(21): 7706-7716.

Miller, B.G. 1993. The press, the Boldt decision, and indian-white relations. American Indian Culture and Research Journal **17**(2): 75-97.

Moran, B.M., and Anderson, E.C. 2018. Bayesian inference from the conditional genetic stock identification model. Canadian Journal of Fisheries and Aquatic Sciences.

Myers, K.W., and Rogers, D.E. 1988. Stock Origins of Chinook Salmon in Incidental Catches by Groundfish Fisheries in the Eastern Bering Sea. North American Journal of Fisheries Management **8**(2): 162-171.

Myers, K.W., Walker, R.V., Davis, N.D., Armstrong, J.L., and Kaeriyama, M. 2009. High seas distribution, biology, and ecology of Arctic-Yukon-Kuskokwim salmon: direct information from high seas tagging experiments 1954-2006. *In* Pacific Salmon: Ecology and Management of Western Alaska's Populations. *Edited by* C.C. Krueger and C.E. Zimmerman. American Fisheries Society, Bethesda. pp. 201-240.

Nagler, J.J., Bouma, J., Thorgaard, G.H., and Dauble, D.D. 2001. High incidence of a male-specific genetic marker in phenotypic female chinook salmon from the Columbia River. Environmental Health Perspectives **109**(1): 67-69.

Nagler, J.J., Cavileer, T., Steinhorst, K., and Devlin, R.H. 2004. Determination of genetic sex in chinook salmon (Oncorhynchus tshawytscha) using the male-linked growth hormone pseudogene by real-time PCR. Marine biotechnology **6**(2): 186-191.

Ohlberger, J., Ward, E.J., Schindler, D.E., and Lewis, B. 2018. Demographic changes in Chinook salmon across the Northeast Pacific Ocean. Fish and Fisheries **19**(3): 533-546.

Phillips, R.B., Park, L.K., and Naish, K.A. 2013. Assignment of Chinook salmon (*Oncorhynchus tshawytscha*) linkage groups to specific chromosomes reveals a karyotype with multiple rearrangements of the chromosome arms of rainbow trout (*Oncorhynchus mykiss*). G3 **3**(12): 2289-2295.

Pritchard, V.L., Erkinaro, J., Kent, M.P., Niemelä, E., Orell, P., Lien, S., and Primmer, C.R. 2016. Single nucleotide polymorphisms to discriminate different classes of hybrid between wild Atlantic salmon and aquaculture escapees. Evolutionary Applications **9**(8): 1017-1031.

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

Schoen, E.R., Wipfli, M.S., Trammell, E.J., Rinella, D.J., Floyd, A.L., Grunblatt, J., McCarthy, M.D., Meyer, B.E., Morton, J.M., Powell, J.E., Prakash, A., Reimer, M.N., Stuefer, S.L., Toniolo, H., Wells, B.M., and Witmer, F.D.W. 2017. Future of Pacific Salmon in the Face of Environmental Change: Lessons from One of the World's Remaining Productive Salmon Regions. Fisheries **42**(10): 538-+.

Siegel, J.E., McPhee, M.V., and Adkison, M.D. 2017. Evidence that Marine Temperatures Influence Growth and Maturation of Western Alaskan Chinook Salmon. Marine and Coastal Fisheries **9**(1): 441-456.

Smit, A., Hubley, R., and Green, P. 2013. RepeatMasker Open-4.0. available at: http://repeatmasker.org.

Smith, C.T., Elfstrom, C.M., Seeb, L.W., and Seeb, J.E. 2005a. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Molecular Ecology **14**(13): 4193-4203.

Smith, C.T., Seeb, J.E., Schwenke, P., and Seeb, L.W. 2005b. Use of the 5 '-nuclease reaction for single nucleotide polymorphism genotyping in Chinook salmon. Transactions of the American Fisheries Society **134**(1): 207-217.

Smith, C.T., Templin, W.D., Seeb, J.E., and Seeb, L.W. 2005c. Single Nucleotide Polymorphisms (SNPs) provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. North American Journal of Fisheries Management **25**(3): 944-953.

Sylvester, E.V.A., Bentzen, P., Bradbury, I.R., Clément, M., Pearce, J., Horne, J., and Beiko, R.G. 2017. Applications of random forest feature selection for fine-scale genetic population assignment. Evolutionary Applications: n/a-n/a.

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.

Thorgaard, G.H. 1977. HETEROMORPHIC SEX-CHROMOSOMES IN MALE RAINBOW-TROUT. Science **196**(4292): 900-902.

Von Bargen, J., Smith, C.T., and Reuth, J. 2015. Development of a Chinook salmon sex identification SNP assay based on the growth hormone pseudogene. Journal of Fish and Wildlife Management **6**(1): 213-219.

Walsh, V.M. 1998. Eliminating driftnets from the North Pacific Ocean: US-Japanese cooperation in the International North Pacific Fisheries Commission, 1953-1993. Ocean Development and International Law **29**(4): 295-322.

Waples, R.K., Seeb, J.E., and Seeb, L.W. 2017. Congruent population structure across paralogous and nonparalogous loci in Salish Sea chum salmon (*Oncorhynchus keta*). Molecular ecology **26**(16): 4131-4144.

Warheit, K.I., Seeb, L., Templin, W.D., and Seeb, J. 2013. Moving GSI into the next decade: SNP coordination for Pacific Salmon Treaty fisheries. Washington Department of Fish and Wildlife, Report FPT 13-09. http://wdfw.wa.gov/publications/01629/wdfw01629.pdf.

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., and Anderson, O.D. 2008. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics **9**: 253.

Table 1. Number of samples sequenced and retained per population for RADseq and GT-seq analyses.

		RADseq		GT-seq			
Population	Region	Sequenced	Retained	Sequenced	Retained	Total	Source[1]
Pitka Fork	Upper Kuskokwim River	0	0	95	95	95	1
Takotna	Upper Kuskokwim River	0	0	95	94	94	1
Tatlawiksuk	Upper Kuskokwim River	0	0	95	56	56	1
Necons	Upper Kuskokwim River	48	47	48	48	95	2
George	Kuskokwim River	48	32	49	40	72	2
Kogrukluk	Kuskokwim River	64	48	47	47	95	3
Aniak	Kuskokwim River	48	47	48	48	95	1
Kisaralik	Kuskokwim River	48	48	47	47	95	1
Kwethluk	Kuskokwim River	48	35	52	52	90	1
Eek	Kuskokwim River	0	0	95	93	93	1
Kanektok	Kuskokwim Bay	48	44	49	49	95	1
Arolik	Kuskokwim Bay	48	46	47	47	93	1
Goodnews	Goodnews Bay	48	47	48	48	95	2
Togiak	Togiak Bay	48	46	47	45	91	1
Iowithla	Nushagak River	48	47	47	17	64	1
Stuyahok	Nushagak River	48	48	47	45	93	1
Koktuli	Nushagak River	56	56	39	39	95	3
Total		648	591	995	910	1506	

[1] Sources: 1) this study, 2) McKinney et al. (2018), 3) Larson et al. (2014b).

Table 2. Putative sex loci identified by the presence of a two genotype classes (heterozygote and one homozygote) at approximately even frequencies in the population. Location of each locus on the Chinook salmon linkage map is listed for mapped loci. Accuracy was calculated by comparing genotypes with observed sex of 23 individuals from the Togiak River population.

Locus	Chromosome	Position (cM)	Accuracy
RAD67724	*Ots17*	4.7	83%
RAD29719	*Ots17*	9.4	78%
RAD27492	NA	NA	91%

Table 3. Size of Chinook salmon linkage groups and number of loci for Washington linkage map, Alaska linkage map, and combined linkage map.

	Washington Map				Alaska Map				Combined Map			
Linkage		Singleton	Duplicated	Total		Singleton	Duplicated	Total		Singleton	Duplicated	Total
Group	Size (cM)	Loci	Loci	Loci	Size (cM)	Loci	Loci	Loci	Size (cM)	Loci	Loci	Loci
Ots01	125.85	669	103	772	107.72	674	107	781	117.24	1066	159	1225
Ots02	120.4	428	140	568	119.94	491	155	646	118.78	729	231	960
Ots03	129.64	507	140	647	109.09	517	161	678	125.78	788	236	1024
Ots04	126.21	417	168	585	104.63	434	196	630	112.29	681	275	956
Ots05	116.49	499	15	514	107.42	592	19	611	107.80	868	31	899
Ots06	119.09	604	116	720	98.14	650	118	768	108.89	958	175	1133
Ots07	134.24	538	113	651	110.29	577	157	734	120.00	850	216	1066
Ots08	106.38	584	11	595	103.54	614	18	632	109.02	921	30	951
Ots09	126.14	560	183	743	127.46	590	205	795	116.79	890	295	1185
Ots10	122.51	410	23	433	106.76	504	22	526	111.70	739	37	776
Ots11	103.84	378	53	431	108.11	432	137	569	104.64	627	168	795
Ots12	127.67	462	159	621	102.06	511	195	706	120.69	780	268	1048
Ots13	114.07	574	21	595	110.24	596	29	625	111.91	901	43	944
Ots14	123.57	374	123	497	105.84	396	154	550	118.43	610	208	818
Ots15	96.29	295	139	434	107.36	288	156	444	98.24	453	232	685
Ots16	114.35	414	13	427	116.55	469	15	484	115.76	702	23	725
Ots17	66.53	164	129	293	64.20	166	122	288	70.34	260	194	454
Ots18	66.47	277	15	292	53.51	305	13	318	60.63	464	20	484
Ots19	73.98	445	18	463	85.89	449	14	463	79.15	677	25	702
Ots20	74.7	354	10	364	77.28	374	19	393	79.07	556	23	579
Ots21	70.02	223	12	235	52.24	260	8	268	60.58	391	18	409
Ots22	66.77	306	11	317	72.21	292	14	306	66.90	444	25	469
Ots23	60.51	131	148	279	64.51	156	148	304	60.35	235	218	453

	Washington Map				Alaska Map				Combined Map			
Linkage		Singleton	Duplicated	Total		Singleton	Duplicated	Total		Singleton	Duplicated	Total
Group	Size (cM)	Loci	Loci	Loci	Size (cM)	Loci	Loci	Loci	Size (cM)	Loci	Loci	Loci
Ots24	56.61	217	2	219	55.31	245	9	254	56.30	333	11	344
Ots25	73.55	303	14	317	70.35	311	16	327	69.16	482	25	507
Ots26	66.66	342	13	355	56.62	332	15	347	66.16	522	24	546
Ots27	69.63	145	150	295	56.01	152	166	318	62.69	231	250	481
Ots28	64.39	280	10	290	55.12	323	8	331	65.15	467	14	481
Ots29	57.75	234	6	240	51.66	265	8	273	55.40	378	11	389
Ots30	70.21	343	12	355	58.71	342	7	349	63.50	536	18	554
Ots31	48.8	239	9	248	55.18	227	6	233	50.89	365	13	378
Ots32	68.42	122	139	261	52.10	110	159	269	61.98	182	237	419
Ots33	77.29	335	3	338	88.48	337	6	343	79.96	508	9	517
Ots34	80.74	111	115	226	59.49	103	132	235	77.19	168	191	359
Total	3119.77	12,284	2,336	14,620	2874.02	13,084	2,714	15,798	3003.36	19,762	3,953	23,715

Table 4. Results of locus filtering prior to mixed-stock analysis (MSA). The number of total loci for each panel are in the Panel Loci column; the number of loci retained after filtering are in the *GSIsim* Loci column. The number of loci filtered at each step are listed in the remaining columns. The retained 847 singleton loci were evaluated for MSA using *GSIsim*.

Panel	Panel Number	Panel Loci	No data for RADseq samples	Duplicate	Allele Frequency Discrepancies	Genotype Rate <70%	Mono-morphic	*GSIsim* Loci
Hess et al. (2016)	1	293	61	6	2	33	34	157
Cook Inlet + TaqMan	2	274	23	13	4	1	4	232
Kuskokwim	3	254	0	24	6	0	2	219
Nushagak	4	271	0	27	5	0	0	239
Total		1,092	84	70	17	34	40	847

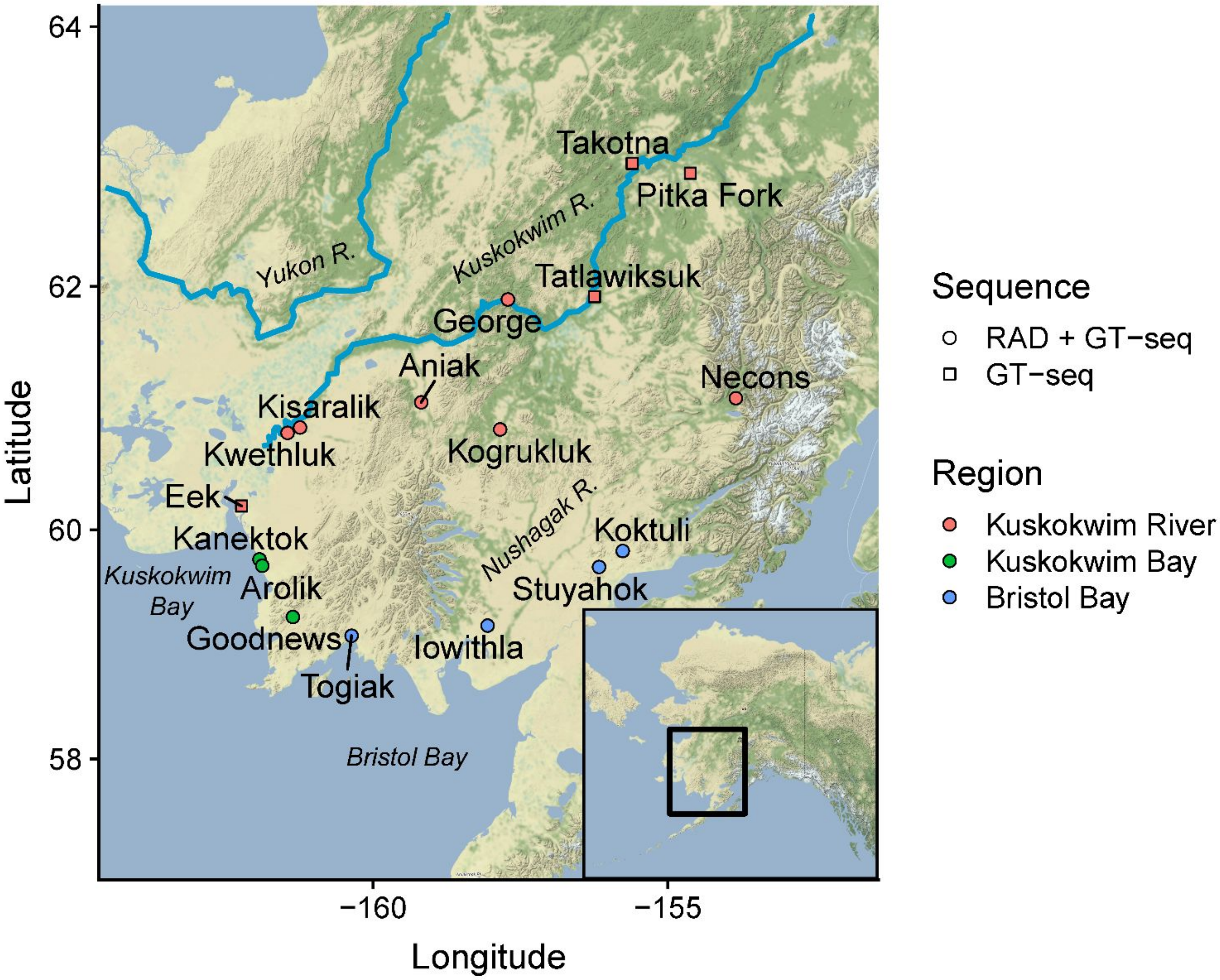


Figure 1. Sample sites for populations used in this study. Shapes denote type of sequencing done on the population; populations are color coded by region. Map tiles by Stamen Design, under CC BY 3.0.

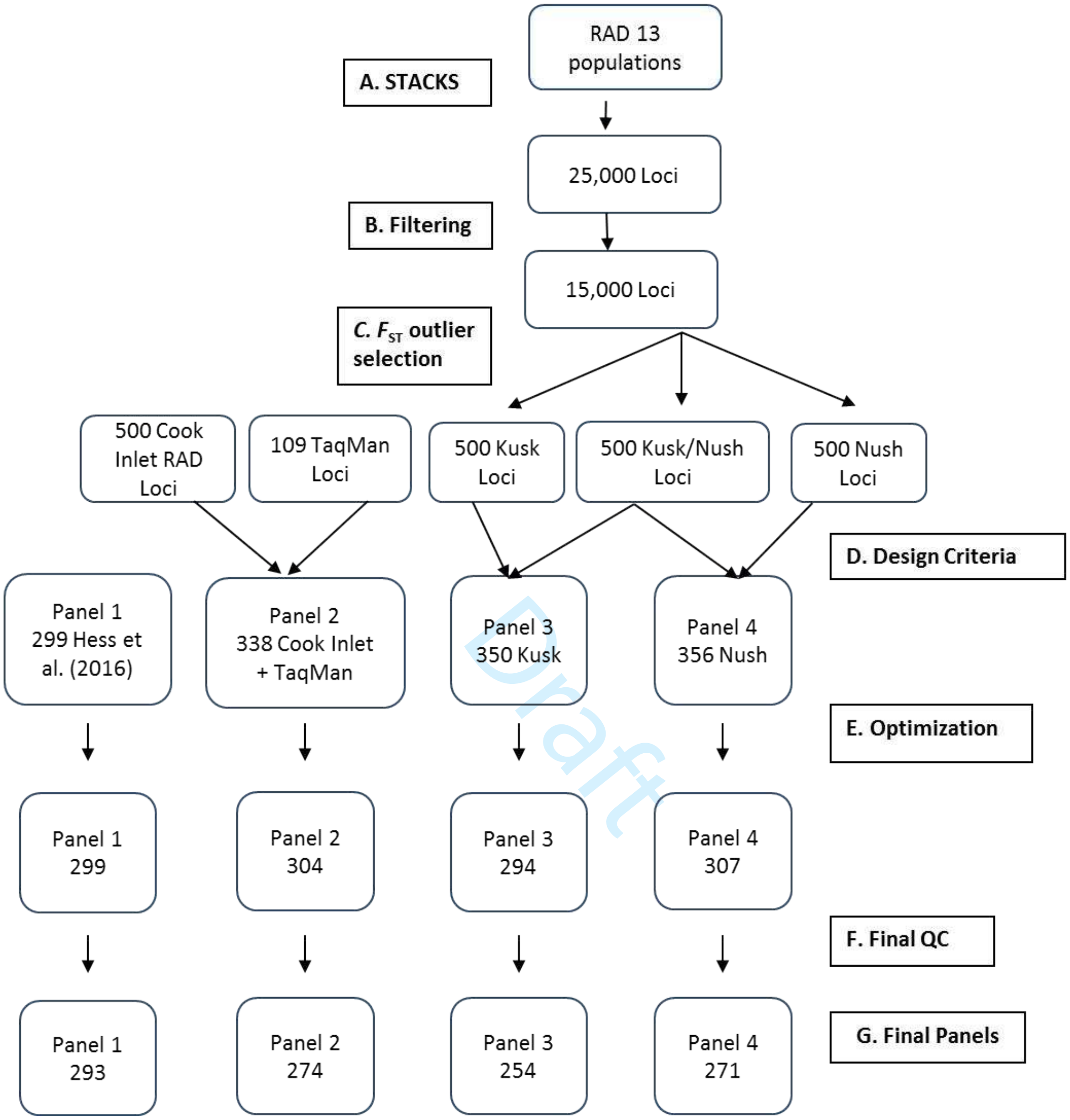


Figure 2. Flowchart of GT-seq panel development.

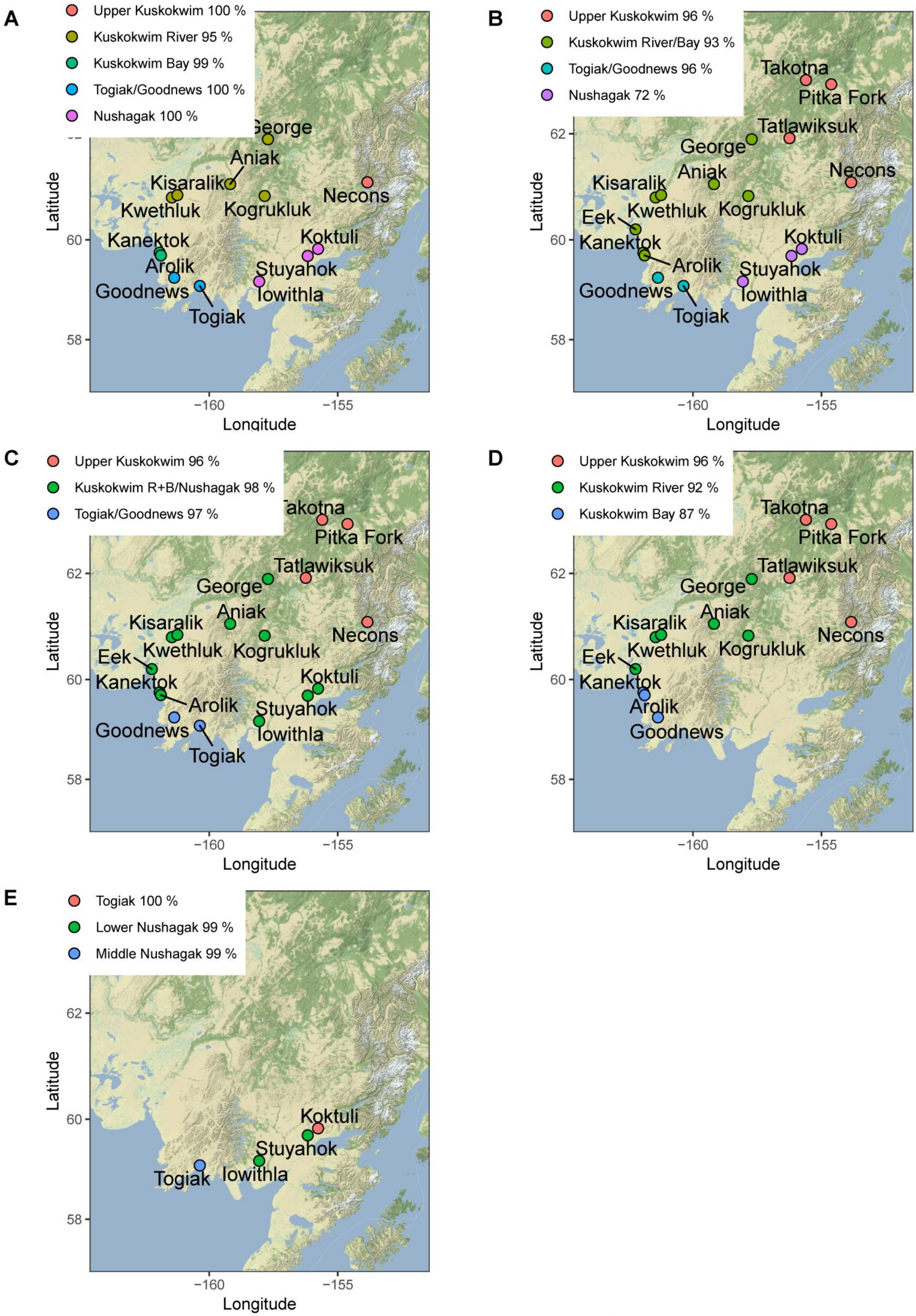
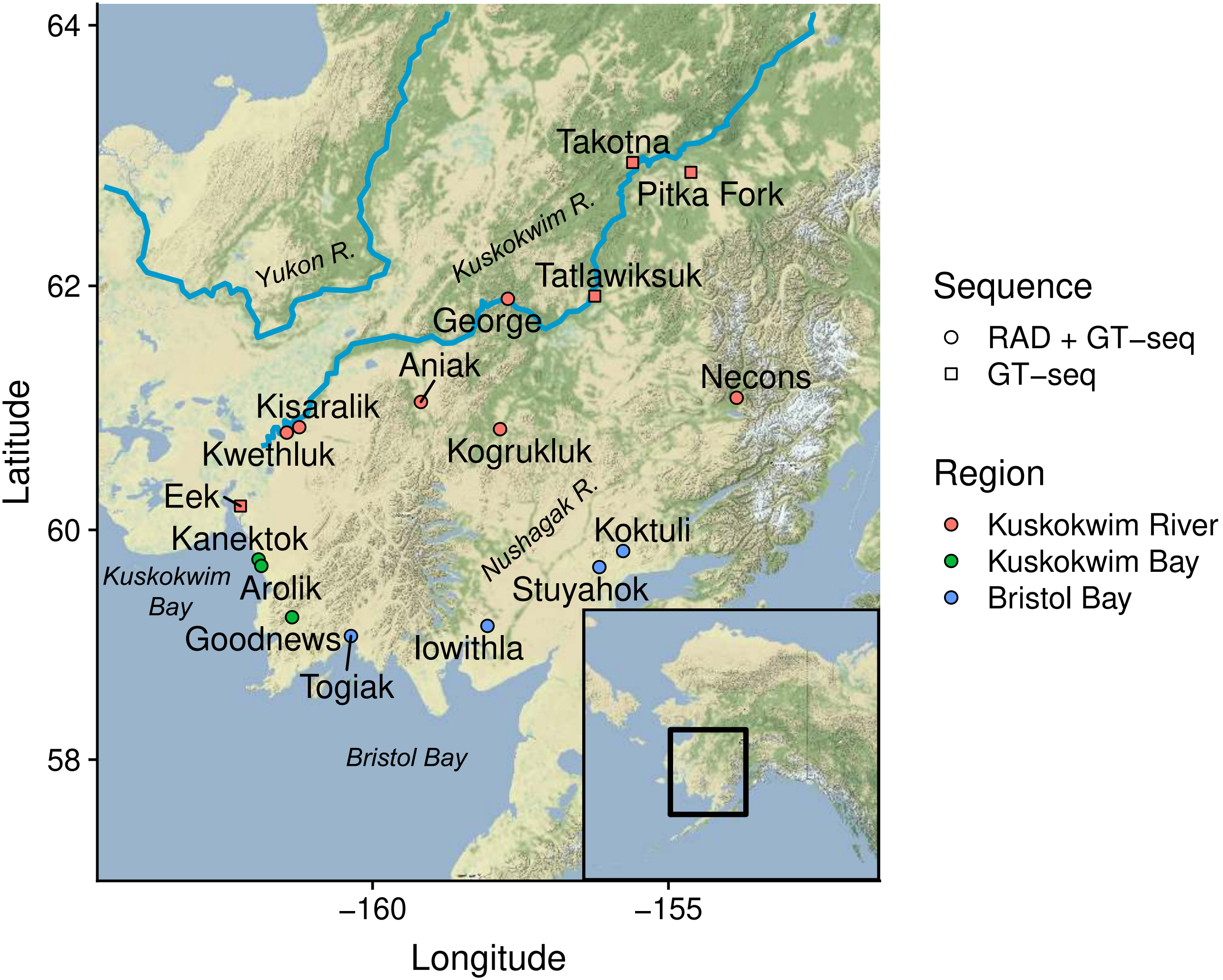

A
● Upper Kuskokwim 100 %
● Kuskokwim River 95 %
● Kuskokwim Bay 99 %
● Togiak/Goodnews 100 %
● Nushagak 100 %
George
Aniak
Kisaralik
Kwethluk
Kogrukluk
Necons
Kanektok
Arolik
Goodnews
Togiak
Koktuli
Stuyahok
Iowithla
Latitude
60
58
-160
-155
Longitude
B
● Upper Kuskokwim 96 %
● Kuskokwim River/Bay 93 %
● Togiak/Goodnews 96 %
● Nushagak 72 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Necons
Kisaralik
Kwethluk
Kogrukluk
Eek
Kanektok
Arolik
Koktuli
Stuyahok
Goodnews
Togiak
Iowithla
Latitude
62
60
58
-160
-155
Longitude
C
● Upper Kuskokwim 96 %
● Kuskokwim R+B/Nushagak 98 %
● Togiak/Goodnews 97 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Kisaralik
Kwethluk
Kogrukluk
Necons
Eek
Kanektok
Arolik
Koktuli
Stuyahok
Goodnews
Togiak
Iowithla
Latitude
62
60
58
-160
-155
Longitude
D
● Upper Kuskokwim 96 %
● Kuskokwim River 92 %
● Kuskokwim Bay 87 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Kisaralik
Kwethluk
Kogrukluk
Necons
Eek
Kanektok
Arolik
Goodnews
Latitude
62
60
58
-160
-155
Longitude
E
● Togiak 100 %
● Lower Nushagak 99 %
● Middle Nushagak 99 %
Koktuli
Stuyahok
Togiak
Iowithla
Latitude
62
60
58
-160
-155
Longitude

Figure 3. Map of populations used in this study and reporting groups tested. SNP discovery was conducted on ascertainment populations using RADseq. All RADseq populations were used for genetic stock assignment along with additional populations and individuals genotyped using GT-seq. For each reporting group scenario, populations on the map are color coded by reporting group, and mean accuracy of reporting groups are shown in the legend. The maximum possible resolution using the full RADseq dataset (>15,000 loci) is shown in A. Reporting group scenarios for GT-seq panel (847 loci) resolution are B) fine-scale resolution, C) broad-scale resolution, D) resolution within Kuskokwim Bay/River, and E) resolution within Togiak Bay/Nushagak River. Map tiles by Stamen Design, under CC BY 3.0.


Takotna
Pitka Fork
Kuskokwim R.
Yukon R.
Tatlawiksuk
George
Aniak
Necons
Kisaralik
Kwethluk
Kogrukluk
Eek
Nushagak R.
Koktuli
Kanektok
Kuskokwim Bay
Arolik
Stuyahok
Goodnews
Iowithla
Togiak
Bristol Bay
64
62
60
58
Latitude
−160
−155
Longitude
Sequence
○ RAD + GT−seq
□ GT−seq
Region
● Kuskokwim River
● Kuskokwim Bay
● Bristol Bay

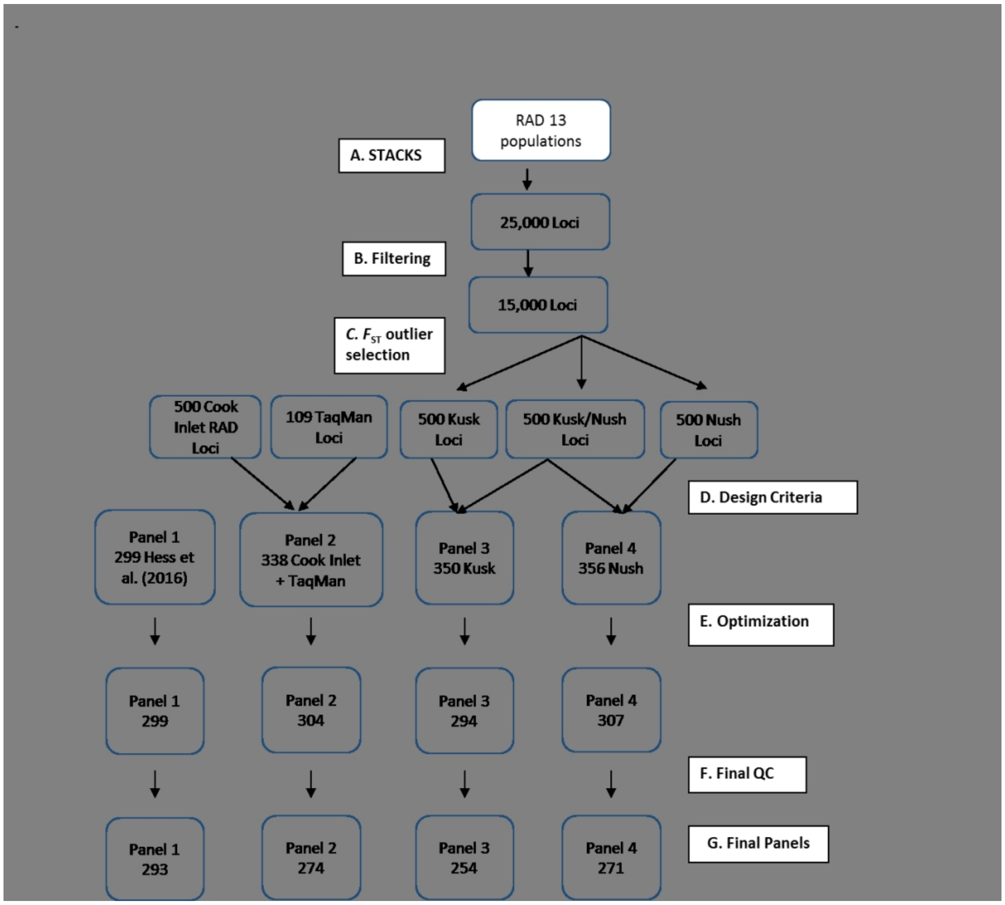
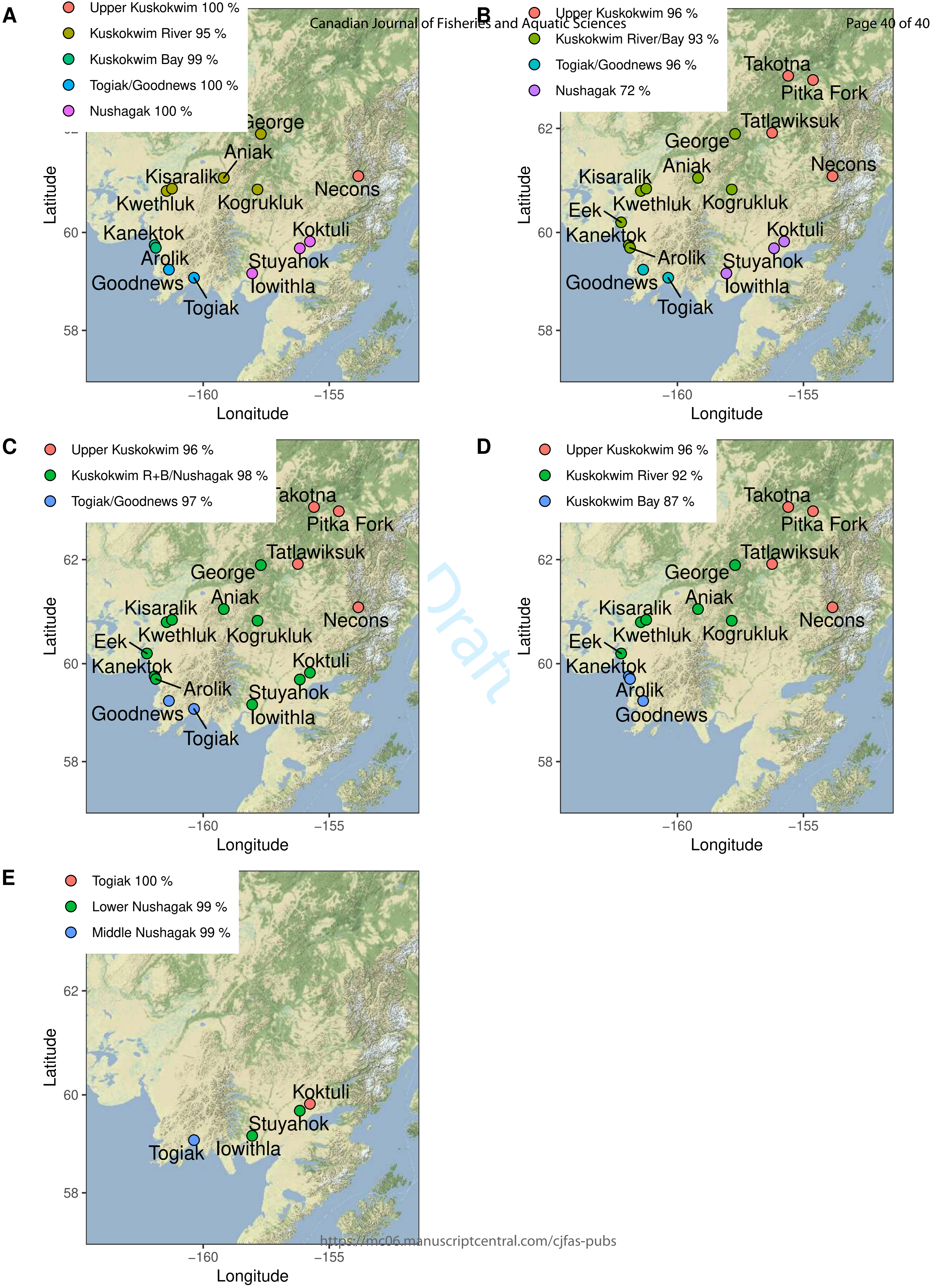


Figure 2. Flowchart of GT-seq panel development.


A
Upper Kuskokwim 100 %
Kuskokwim River 95 %
Kuskokwim Bay 99 %
Togiak/Goodnews 100 %
Nushagak 100 %
George
Aniak
Kisaralik
Kwethluk
Kogrukluk
Necons
Kanektok
Arolik
Goodnews
Togiak
Koktuli
Stuyahok
Iowithla
60
58
−160
−155
Latitude
Longitude
B
Upper Kuskokwim 96 %
Kuskokwim River/Bay 93 %
Togiak/Goodnews 96 %
Nushagak 72 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Necons
Kisaralik
Eek
Kwethluk
Kogrukluk
Kanektok
Arolik
Koktuli
Stuyahok
Goodnews
Iowithla
Togiak
62
60
58
−160
−155
Latitude
Longitude
C
Upper Kuskokwim 96 %
Kuskokwim R+B/Nushagak 98 %
Togiak/Goodnews 97 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Kisaralik
Necons
Eek
Kwethluk
Kogrukluk
Kanektok
Koktuli
Arolik
Stuyahok
Goodnews
Iowithla
Togiak
62
60
58
−160
−155
Latitude
Longitude
D
Upper Kuskokwim 96 %
Kuskokwim River 92 %
Kuskokwim Bay 87 %
Takotna
Pitka Fork
Tatlawiksuk
George
Aniak
Kisaralik
Necons
Eek
Kwethluk
Kogrukluk
Kanektok
Arolik
Goodnews
62
60
58
−160
−155
Latitude
Longitude
E
Togiak 100 %
Lower Nushagak 99 %
Middle Nushagak 99 %
Koktuli
Stuyahok
Togiak
Iowithla
62
60
58
−160
−155
Latitude
Longitude