

Population-size history inferences from the coho salmon (*Oncorhynchus kisutch*) genome

Eric B. Rondeau,^{1,2,3,†} Kris A. Christensen ,^{1,2,*†} David R. Minkley,^{1,2} Jong S. Leong,¹ Michelle T.T. Chan,^{2,4} Cody A. Despains,¹ Anita Mueller,¹ Dionne Sakhrani,² Carlo A. Biagi,² Quentin Rougemont,⁵ Eric Normandeau,⁵ Steven J.M. Jones,⁶ Robert H. Devlin,² Ruth E. Withler,³ Terry D. Beacham,³ Kerry A. Naish,⁷ José M. Yáñez,^{8,9} Roberto Neira,^{9,10} Louis Bernatchez,⁵ William S. Davidson,⁴ Ben F. Koop^{1,*}

¹Department of Biology, Centre for Biomedical Research, University of Victoria, Victoria, BC, V8W 2Y2, Canada

²Fisheries and Oceans Canada, 4160 Marine Drive, West Vancouver, BC, V7V 1N6, Canada

³Fisheries and Oceans Canada, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, BC, V9T 6N7, Canada

⁴Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

⁵Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC, G1V 0A6, Canada

⁶Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, V5Z 4S6, Canada

⁷School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, 98105, USA

⁸Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile, Santa Rosa 11735, La Pintana, Santiago, 8820808, Chile

⁹Millennium Nucleus of Austral Invasive Salmonids (INVASAL), Concepción, 4030000, Chile

¹⁰Facultad de Ciencias Agronómicas, Universidad de Chile, Santa Rosa 11315, La Pintana, Santiago, 8820808, Chile

[†]Present address: CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, 34293, France.

*Corresponding authors: Ben F. Koop, Department of Biology, Centre for Biomedical Research, University of Victoria, Victoria, BC, V8W 2Y2, Canada. Email: bkoop@uvic.ca; Kris A. Christensen, Fisheries and Oceans Canada, 4160 Marine Drive, West Vancouver, BC, V7V 1N6, Canada. Email: kris.christensen@wsu.edu

[†]Equal contributions.

Abstract

Coho salmon (*Oncorhynchus kisutch*) are a culturally and economically important species that return from multiyear ocean migrations to spawn in rivers that flow to the Northern Pacific Ocean. Southern stocks of coho salmon in Canada and the United States have significantly declined over the past quarter century, and unfortunately, conservation efforts have not reversed this trend. To assist in stock management and conservation efforts, we generated a chromosome-level genome assembly. We also resequenced the genomes of 83 coho salmon across the North American range to identify nucleotide variants and understand the demographic histories of these salmon by modeling effective population size from genome-wide data. From demographic history modeling, we observed reductions in effective population sizes between 3,750 and 8,000 years ago for several northern sampling sites, which may correspond to bottleneck events during recolonization after glacial retreat.

Keywords: coho, silver salmon, Pacific salmon, genome assembly, SNPs, demographic history, genomics

Introduction

Coho salmon have special cultural significance to the people of the First Nations in British Columbia and have traditionally been one of the highest-value Pacific salmon in the commercial and recreational fishery sectors. In 1977, a climatic regime shift in the North Pacific Ocean ushered in 3 decades of increasing Pacific salmon production that culminated in 2009, when over 600 million salmon (1.1 million metric tonnes) were harvested (Irvine and Fukuwaka 2011). However, this increased production of salmon masked substantial variability in regional abundances and species composition. Whereas the productivity and harvest of chum (*Oncorhynchus keta*), pink (*O. gorbuscha*), and sockeye salmon (*O. nerka*) increased throughout the North Pacific after 1977, the opposite was true for coho and Chinook salmon (*O. tshawytscha*). These declines became particularly acute after 1989 when marine survival for these species began a downward spiral that has yet to be reversed (Beamish et al. 2010; Irvine and

Fukuwaka 2011). A severe decline in the highly lucrative recreational coho salmon fishery in the Strait of Georgia saw the numbers of fish caught decrease from an average of over 500,000 to less than 100,000 throughout the 1990s (Beamish et al. 2011) (see Fig. 1 for a map of this region). In 2004, the recreational catch in the Strait of Georgia was just 9,500 coho salmon (Kristianson and Strongitharm 2006). Improved genetic resources and further studies may help us to better understand this decline and how to reverse it.

In a large-scale population structure analyses of coho salmon sampled from 318 localities, in 38 different regional groups in North America and Russia (representing most of the natural distribution of coho salmon), 17 microsatellite loci showed that salmon clustered geographically and regions could be delineated along a north–south gradient, with reduced variation to the north and isolated inland populations (Beacham et al. 2011). Follow-up studies, which used thousands of genetic markers in their analyses, observed that isolation-by-distance after the recolonization

Received: November 22, 2022. Accepted: January 26, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

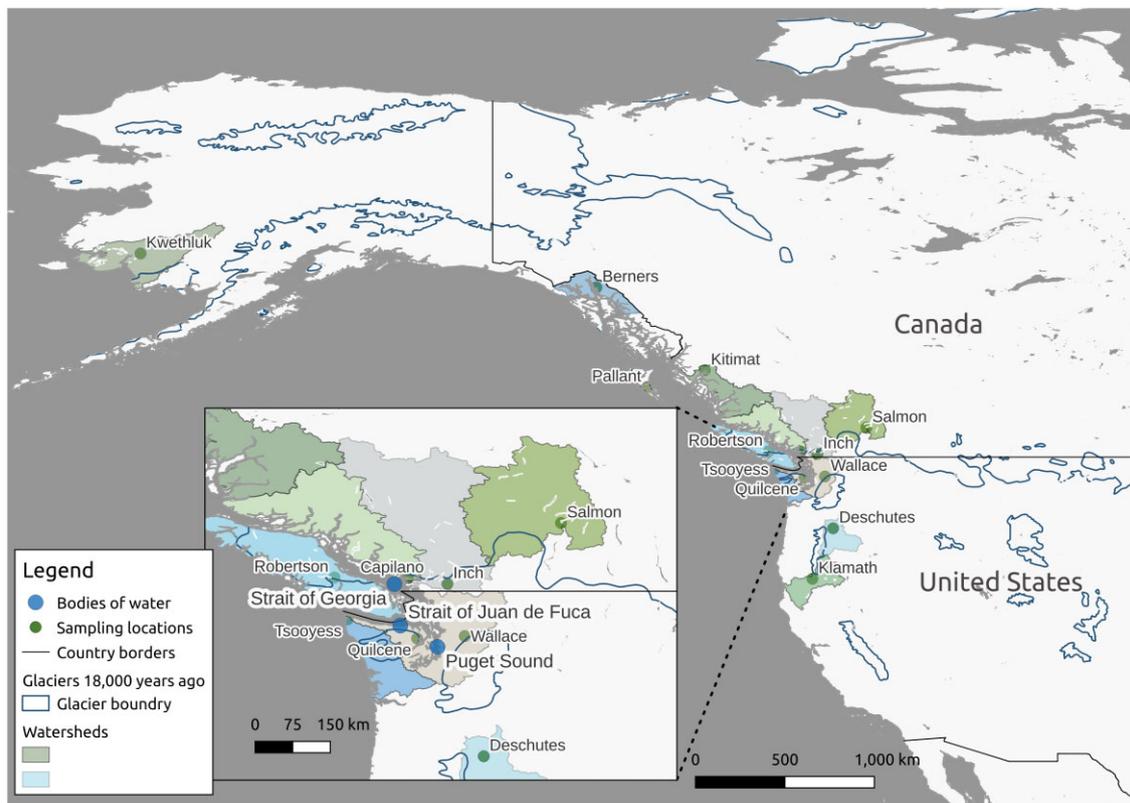


Fig. 1. North Pacific Ocean with sampling locations and the extent of glaciers 18,000 years ago. This map was drawn in qGIS (QGIS Development Team 2022) with data from the following resources: watersheds (Pasos), glaciers (Government of Canada 2015), and all other information (Natural Earth). Watersheds were shown only if they contained a sampling location. The last glacial maximum has been estimated as ~18,000 years ago [e.g. (Latch et al. 2009; Beatty and Provan 2010)]. The sampling locations for this study would be largely influenced by the Cordilleran Ice Sheet with an estimated maximum of ~19,000–20,000 years ago (Clark et al. 2009), and if so, the glacial borders could be slightly conservative in this figure but still represent the extent of the North American ice sheets.

of northern streams from a main southern glacial refugium since the last glacial maximum could explain most of the patterns of genetic diversity in modern coho salmon across their North American distribution (Rougemont et al. 2020, 2022) [see Fig. 1 for the extent of North American ice sheets 18,000 years ago (Clark et al. 2009)].

After glaciers receded, previously inaccessible areas were once again habitable. It has been suggested that recolonization of these areas has influenced the present-day distribution of many other species, including plants [e.g. (Beatty and Provan 2010)] and other animals [e.g. (Latch et al. 2009)]. As noted in Beatty and Provan (2010), multiple North American refugia were possibly used by these organisms. For example, Arctic char (*Salvelinus alpinus*) are thought to have persisted in refugia north of the North American ice sheets (Moore et al. 2015) rather than in a southern refugium (as with coho salmon). With nearly 45% of the current freshwater habitat of Pacific salmon species under ice sheets during the last glacial maximum (Pitman et al. 2020), recolonization also played a major role in the modern distribution of coho salmon.

With genetic resources for conservation efforts in mind, and with the objective of expanding our understanding of recolonization after the last glacial maximum, the goals of this study were to: (1) generate a high-quality chromosome-level reference genome assembly and (2) estimate the timing of coho salmon recolonization of northern streams. Our method to construct a high-quality genome assembly was to use multiple sequencing technologies and techniques to complement each other. Our method to estimate the timing of recolonization was to

resequence the genomes of 83 coho salmon and model effective population size from a comprehensive inventory of SNPs. This was done to identify genetic bottlenecks after the last glacial maximum corresponding to founding events during recolonization.

Materials and methods

Coho salmon samples for genome assembly

All animals were reared in compliance with the Canadian Council on Animal Care Guidelines, under permit from the Fisheries and Oceans Canada Pacific Region Animal Care Committee (under Ex.7.1). Using Inch Creek coho salmon (see Fig. 1 for location, NCBI BioSample accessions: SAMN05991263 and SAMN11962340), we generated fully homozygous diploid gynogenetic individuals [doubled haploids, see (Biagi et al. 2022) for full details] to help improve genome assembly quality. As noted in (Zhang et al. 2014), genome assemblies tend to benefit from the reduction of heterozygous genotypes, which might cause fragmentation during contig assembly.

Briefly, to produce doubled haploid individuals, a pressure shock (75.8 MPa for 5 minutes) was applied 6 hours post-fertilization (10°C) to coho salmon eggs that had been fertilized with UV-inactivated Atlantic salmon sperm (Chourout 1984). Surviving individuals from experiments with Inch Creek coho salmon were grown until they could be non-lethally biopsied and their genotypes determined using a panel of 9 highly polymorphic microsatellite markers (Beacham et al. 2001) to assess homozygosity, non-paternity, and sex. Females that were homozygous at all

loci and lacked any paternal alleles were retained and grown in freshwater for 8 months until they reached a size of 10–14 grams at which time selected fish were euthanized in 200 mg/L tricaine methanesulfonate (Syndel Laboratories, Canada) and sodium bicarbonate (400 mg/L). Tissue samples were stored at -80°C . Tissues from one coho salmon (NCBI BioSample SAMN05991263) were used for 10× Chromium, PacBio RS II and Sequel, and multiple Illumina sequencing libraries. Tissues from another individual (NCBI BioSample SAMN11962340) were used after the first coho salmon tissues were exhausted for Hi-C and further 10× Chromium sequencing.

DNA for Illumina sequencing was extracted from frozen muscle tissue using a Phenol:Chloroform extraction [adapted from (Sambrook et al. 2000)]; buffer ATL and Proteinase K (Qiagen) were used to digest tissue overnight. Phenol:chloroform:isoamyl alcohol (25:24:1) was added 50:50 to aqueous phase (3x) with a final extraction with chloroform:isoamyl alcohol (24:1). The aqueous phase was removed, and precipitated with 0.2 volumes of 10 M ammonium acetate and 2.5 volumes 95% EtOH. Precipitated DNA was extracted using a shepherd's crook and transferred to 70% EtOH to wash. DNA was resuspended overnight in buffer EB (Qiagen), and the DNA was quantified using a NanoDrop ND-1000 and a Qubit v2.0 Broad Range dsDNA kit (Life Technologies). The integrity of the DNA was confirmed by agarose gel electrophoresis.

High-molecular weight DNA was extracted from muscle tissue using a modified dialysis method. The tissue (550 mg) was ground into a powder with a mortar and pestle with liquid nitrogen. The powder was transferred to a 5 ml LoBind Eppendorf tube, along with 3600 μl buffer ATL (Qiagen), 400 μl proteinase K solution (Qiagen) and 40 μl RNase A solution (Qiagen). The sample was digested at 56°C for 3 hours, with rotation at ~ 4 rpm. A phenol-chloroform-isoamyl alcohol (25:24:1) purification was performed 3 times, followed by 1 round of chloroform-isoamyl alcohol (24:1). At each stage, 1 volume of organic was mixed with 1 volume of aqueous, inverted slowly for 3 minutes to mix thoroughly, spun for 15 minutes at 5000 g to separate the layers, and the aqueous top layer was transferred very slowly to a new tube using a 1000 μl pipette and wide bore tip. RNase A solution (2 μl , 20 mg/ml Qiagen) was added and incubated at room temperature for 1 hr, followed by 5 μl Proteinase K (20 mg/ml) overnight at 4°C . Approximately 750 μl was obtained from each tube after the final stage. This volume was transferred to a Spectra/Por Float-A-Lyzer G2 1000 kD (pink) dialysis device, and dialysis was performed in 1 L of 10 mM Tris-Cl, pH 8.5 at 4°C with gentle mixing for 1 week, changing buffer 5 times. DNA quantity was determined by Qubit and quality by 0.6% agarose gel at 60 Volts.

Library preparation and sequencing details for the various sequencing technologies and samples can be found under the following NCBI submissions: 10× Chromium (SRX5975809, SRX5975837–SRX5975838), Hi-C (SRX5975810), PacBio Sequel (SRX5975835–SRX5975836), PacBio RS II (SRX2338326–SRX2338334), Illumina mate-pair (SRX2336577–SRX2336579, SRX2333332), and Illumina overlapping paired-end (SRX2333080—SRX2333081).

Genome assembly

All PacBio data was assembled using Canu v1.8 (Koren et al. 2017). A genome size of 2.4 Gbp was estimated as input for Canu. Default settings were used with the exception of the following: “ovlMerThreshold = 2,000 corMhapSensitivity = normal corrected ErrorRate = 0.085 minReadLength = 2,500 corOutCoverage = 200 'batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50'.”

Arrow v 2.2.2 from SMRTlink 6.0.0.47841 (PacBio) was used to polish the assembly with PacBio data with the ArrowGrid wrapper program (Koren et al. 2017). The assembly was also polished with Pilon v1.22 (Walker et al. 2014) using all the Illumina data generated for this project.

We performed scaffolding in 3 stages. In the first stage, BESST (Sahlin et al. 2014, 2016) was used to scaffold the assembly with the Illumina mate-pair data [prepared using NXTrim (O'Connell et al. 2015)]. Unknown and mate-pair fractions from NXTrim output were aligned to the polished genome with BWA mem 0.7.17 (Li 2014) and sorted with SAM tools 1.9 (Li et al. 2009). BESST was executed under default parameters, with data input in the order of (1) unknown and (2) mate-pair with inputs ordered smallest to largest target fragments size (all “-orientation rf”).

In the second scaffolding stage, the Tigmint—arcs—links pipeline was utilized. Tigmint v1.1.2 (default options) (Jackman et al. 2018) was run using the “arcs” pipeline to run all 3 stages. Within the Arcs v1.0.5 (Yeo et al. 2018) and LINKS v1.8.6 (Warren et al. 2015) portions of the pipeline, default parameters were used except $a=0.4$ and $l=5$ were used after parameter optimization.

A custom gap-closing script was run to attempt to fill newly scaffolded gaps with error-corrected PacBio data, for full details see (Christensen et al. 2020). Purge Haplotigs (Roach et al. 2018) was used to remove scaffolds with extremely low coverage. PacBio subreads were aligned to the genome assembly with Minimap2 (Li 2018), and a minimum coverage limit of $15\times$ was used to remove potential artifacts. A second round of Arrow polishing was performed as above. As gap sizes were no longer meaningful at this stage, and Tigmint introduced some very short fragments, sed (a Unix program) was used to remove contigs smaller than 200 bp nested within scaffolds and to resize all remaining gaps to 100 Ns. Bioawk fastx (<https://github.com/lh3/bioawk>) was used to remove any remaining scaffolds smaller than 1001 bp (<100 total removed).

After the Tigmint pipeline, Hi-C data was aligned to the genome with Juicer 1.5.6 (Durand et al. 2016b). The parameters used were “-s Sau3AI and -S early” while also including Sau3AI cutsite file with the “-y” parameter. The resulting “merged_nodups.txt” file was input into 3d-dna v. 180,922 (Dudchenko et al. 2017) with “-i 50,000 -r 0.” A modified “.assembly” file, with ordered contigs/scaffolds, was created with data from a new genetic map generated from a 200 K Affymetrix SNP array (Barria et al. 2019). Scaffolds without mapped markers from the genetic map were identified and the order and orientation for these scaffolds were initially input into the assembly file based on synteny to the rainbow trout genome (Pearse et al. 2019) taken from the output of RaGOO (Alonge et al. 2019). Juicebox v1.8.8 (Durand et al. 2016a) was used to visualize and improve the assembly post-scaffolding. Purge_Haplotigs was used with the following settings—“contigcov -115 -h 300 -m 155” and “purge -m 400” to allow for identification of “haplotigs” that represent high-percent identity, homeologous sequence.

Scaffolds that remained unlinked following manual review of the Hi-C data were repeat masked and aligned to the northern pike genome assembly (Rondeau et al. 2014) using LastZ (--nochain --gextend --nogapped --identity = 75.0..100.00 --matchcount = 100 --format = general -hspthresh = 5,000). The northern pike was used due to its close relationship to salmonids and since it does not have a recent ancestral genome duplication, which is common to all salmonids (Leong et al. 2010). Scaffolds were ordered and oriented in the residually tetraploid-like regions of the genome to reflect the order in northern pike. The final genome was re-visualized with Hi-C in juicebox to remove any scaffolds for

Table 1. Whole-genome resequencing sampling locations.

Source	Country	State/province	Count female, male
Klamath River (Hatchery)	USA	CA/OR	1F, 4M
Deschutes River (Hatchery)	USA	CA/OR	2F, 3M
Big Quilcene River (Hatchery)	USA	WA	2F, 3M
Wallace River (Hatchery)	USA	WA	6 M, 4?
Tsoo-Yess River (Hatchery)	USA	WA	1F, 4M
Inch Creek (Hatchery)	Canada	BC	3F, 5M
Capilano River (Hatchery)	Canada	BC	5F
Robertson Creek (Hatchery)	Canada	BC	5M
Salmon River (Hatchery)	Canada	BC	5F
Pallant Creek	Canada	BC	5M
Kitimat River (Hatchery)	Canada	BC	5F, 5M
Berners River	USA	AK	2F, 3M
Kwethluk River	USA	AK	1F, 4M
AquaChile (Strain)	Chile	NA	5F

Table 2. SNP filtering for different analyses.

Parameters	Bi-allelic, < 10% missing, coverage 5–200	^a MAF 0.05	^a Linkage disequilibrium
Analysis(es)	SMC++	Pi, Fis, observed heterozygosity, private allele count	PCA, Mantel test
Number of variants	14,397,038	5,631,459	152,128
Average SNP coverage	21.6x	21.9x	21.6x
Average missing genotypes	17,665	11,780	514

^a Represents additional filters to the previous category.

which Hi-C evidence did not support inclusion in at least one of the 2 tetraploid homeologues, then the Phase Genomics juicebox_assembly_converter.py script (https://github.com/phasegenomics/juicebox_scripts/; commit 7692ad5) was used to generate the NCBI AGP files. After manual editing to rename linkage groups based on prior assemblies and the northern pike linkage map (Rondeau et al. 2014) [identified through LastZ alignments in Geneious (Kearse et al. 2012)], the genome assembly was submitted to the NCBI. The assembly was uploaded to NCBI under the BioProject ID PRJNA352719. A Circos plot (Krzyszowski et al. 2009) of the genome assembly was generated. Homeologous regions identified in this plot were detected using the pipeline from <https://github.com/KrisChristensen/HomeologousRegionIdentification>, and repetitive elements were detected using <https://github.com/KrisChristensen/NCBIGenomeRepeats>.

Whole-genome resequencing and nucleotide variant calling

Whole-genome resequencing was used to identify SNPs across the coho salmon's North American range. Figure 1 and Table 1 identify sample locations (see Supplementary File S1 for more information). We included 1 commercial aquaculture strain from Chile as well for comparison (Table 1). Many of the samples are from hatchery sources. While this likely influences population level metrics, our main goal with these samples is to better understand the timing of recolonization of northern streams after the last glacial maximum. As most modern hatchery stock originates from local sources [e.g. (Heard 2012)], the influence of using hatchery samples is not expected to impact our interpretation of

regional information. A more likely impact of sampling from hatcheries is that modern effective population size may differ from wild populations and we consider this in our interpretations of the data.

DNA for resequencing was extracted from fin-clips using the DNeasy Blood and Tissue extraction kit (Qiagen) or a MagMAX DNA Multi-Sample Ultra Kit with a KingFisher (ThermoFisher Scientific). Following DNA extraction, samples were quantified by Qubit BR DNA assay (ThermoFisher) and integrity validated by agarose gel electrophoresis. At McGill University and Genome Québec Innovation Centre (Montreal, QC, Canada), individual Illumina libraries were constructed with Illumina TruSeq LT sample preparation kits, and each individual was sequenced either on an Illumina HiSeq2500 (PE125) or a HiSeqXTen (PE150), targeting approximately 15–30x coverage. Resequenced genomes were submitted to the NCBI under BioProjects: PRJNA401427 and PRJNA808051 (Supplementary File S1).

BWA-MEM v0.7.17 (Li 2013) was used to align Illumina data to the reference genome, with the -M option for Picard compatibility. The Picard v2.18.9 (github.com/broadinstitute/picard) AddOrReplaceReadGroups program was used to add read group IDs, and the MarkDuplicates program was used to mark duplicates (default settings). GATK v3.8 (McKenna et al. 2010; Van der Auwera et al. 2013) was then used to call genotypes. Base and variant recalibration were each performed once (for 2 rounds through genotype). The variants used for recalibration were from 1) a reduced set of very high-confidence calls following default "hard-filtering" guidelines from GATK documentation from the first round of genotyping with a particular focus on coding regions and 2) validated SNPs from a 200 K Affymetrix SNParray [validated in a previous study (Barria et al. 2019)].

Following genotyping, VCFtools v0.1.15 (Danecek et al. 2011) was used to additionally filter data to only include biallelic SNPs with a minor allele frequency of 0.05 or greater, variants with fewer than 10% missing genotypes, and variants with a mean coverage between 5 and 200. Finally, the SNPs were filtered for linkage disequilibrium to reduce the influence of large haploblocks [bcftools (Danecek et al. 2021) version 1.9–102-g958180e + prune -w 20 kb -l 0.4 -n 2]. See Table 2 for SNP filtering parameters used for each of the analyses in the following paragraphs.

Basic population metrics and demographic history modeling

A principal components analysis (PCA) was performed using PLINK (Chang et al. 2015) v1.90b6.15 with default parameters (<https://www.cog-genomics.org/plink/1.9/>). Private allele counts

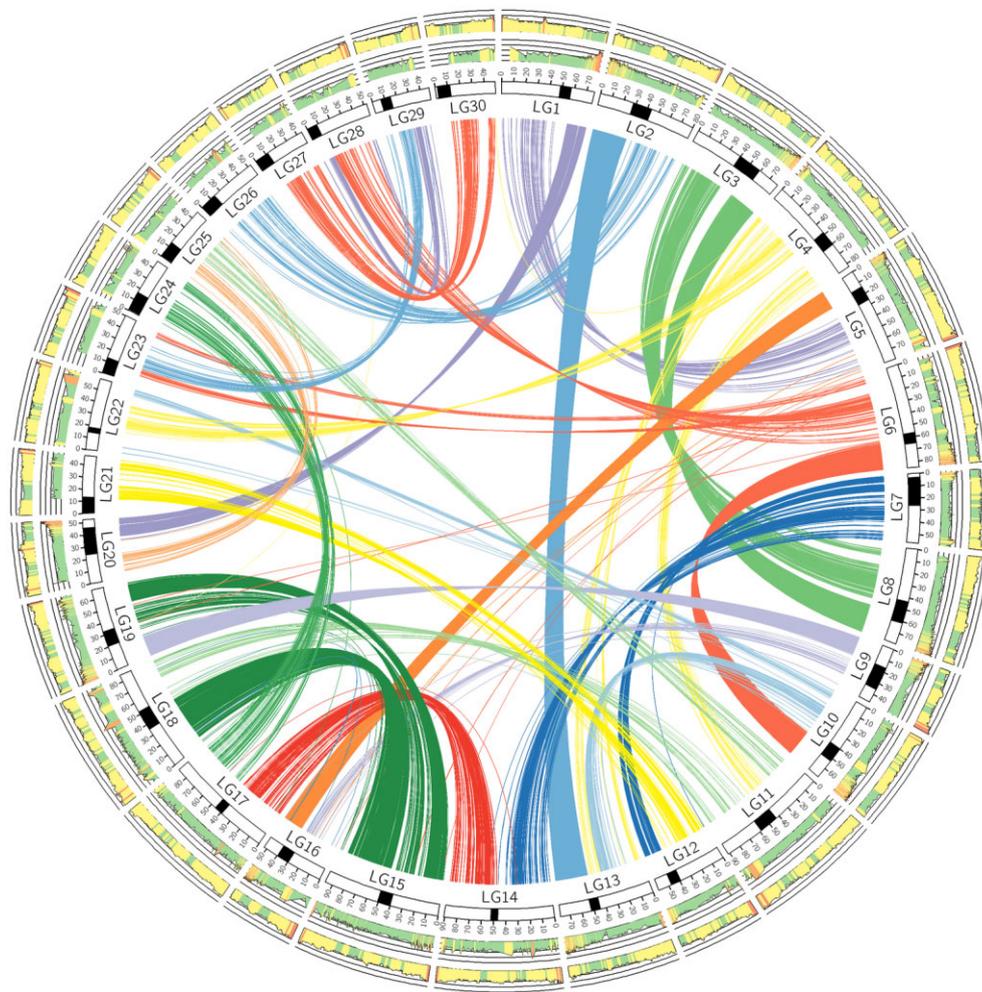


Fig. 2. Circos plot of the coho salmon genome assembly. In the interior of the Circos plot, the links between duplicated regions of the chromosomes/linkage groups (i.e. homeologous regions) are shown. The ring outside of these links contains representations of the chromosomes with the approximate position of the centromere marked by a filled region [centromere positions from (Kodama et al. 2014)]. The tick marks represent 10 Mbp intervals. The next ring contains the percent identity between duplicated regions of the chromosome. The red color represents very high similarity (> 98%), the orange color high similarity (94–98%), the yellow moderate (91–94%), and the green low (< 91%). On the outer ring, the fraction of repetitive elements is shown, with red representing high (> 70%), yellow as moderate (45–70%), and green as low (< 45%).

per river (river is used instead of sampling site throughout the rest of the text) were tallied using the populations module in Stacks (Catchen et al. 2011, 2013) version 2.54 with default parameters. Populations with more than 5 individuals were randomly subsampled to 5 to reduce the influence of uneven sampling on the number of private alleles identified. Stacks was also used to calculate other population level metrics such as observed heterozygosity, nucleotide diversity (π), and F_{is} with default settings. A Mantel test was performed in R (R Core Team 2020) with the following packages to test isolation-by-distance: adegenet (version 2.1.7) (Jombart 2008; Jombart and Ahmed 2011), ade4 (Dray and Dufour 2007; Bougeard and Dray 2018; Thioulouse et al. 2018), and vcfr (version 1.13.0) (Knaus and Grünwald 2017).

To infer demographic histories of the salmon from the various rivers, we used SMC++ (Terhorst et al. 2017) version 1.15.4.dev18 + gca077da. In this analysis, we set the mutation rate to $8e-9$ bp/generation and the generation time to 3 years. These parameters were previously used in another coho salmon study examining demographic histories (Rougemont et al. 2020). We used nucleotide variants that were not filtered for rare variants (e.g. $MAF < 0.05$). We also used the `-missing-cutoff` option (50 kbp) in SMC++ to reduce the influence of missing genotypes (e.g. in centromeres).

While each study site was sparsely sampled (Table 1), for this analysis we were mainly interested in understanding bottleneck events that occurred thousands of years ago. These events should be echoed in the genomes of a majority of individuals from a particular location as most presumably originated from the same founding event (geographically and chronological). Salmon tend to return to the same body of water or a suitable nearby location (Quinn 1993). For recolonization, we would expect major bottleneck events after the last glacial maximum in all of the streams covered by ice-sheets.

Results

Genome assembly

The size of the coho salmon genome assembly was 2.3 Gb, which is consistent with that for the closely related Chinook salmon genome assembly (NCBI accession: GCF_018296145.1). The contig N50 of the assembly was 1,159 kb and the percent of complete BUSCOs was 99%, which is comparable to the human genome assembly also at 99% (NCBI accession: GCF_000001405.39) (Table 2).

The coho salmon genome has extensive signatures of chromosomal duplication (Fig. 2), which have been retained from the

Table 3. Genome statistics.

Contig N50	Contig #	BUSCO	% Repeats	Genes
1,159,298	8,770	99.2%–57.1:42.2 ^{a,b}	53.12 ^b	60,330 ^b

^a Percent complete-single: duplicate.

^b Reported by NCBI (NCBI used actinopterygii_odb10 for BUSCO).
NCBI: GCF_002021735.2.

whole genome duplication common to all salmonids (Allendorf and Thorgaard 1984). The majority of homeologous regions from the salmonid-specific genome duplication have diverged by at least 9% (*i.e.* $\leq 91\%$ identity, Fig. 2), but certain sections of the genome have retained high sequence similarity (Fig. 2). Regions with very high sequence similarity remained as unplaced scaffolds (likely collapsed into 1 sequence) as it was not possible to resolve which sequence belonged to which duplicated region (see assembly methods available on the NCBI website, accession: GCF_002021735.2).

The coho salmon genome also has a high retention of repetitive elements (Fig. 2; Table 3), which is another commonality of studied salmonids [*e.g.* (Lien *et al.* 2016; Christensen *et al.* 2020)]. This is especially true in regions near the centromere where the fraction of repetitive elements is roughly 60–70% (Fig. 2). That value is high compared to the genome average of 53% (Table 3). For comparison, the most recent version of the Chinook salmon genome also has a repeat content of 53% (NCBI accession: GCF_018296145.1).

Basic population metrics and demographic history modeling

A PCA of 83 resequenced coho salmon genomes sampled from across North America (see Table 2 for the number of SNPs for analyses), revealed that coho salmon clustered by region with the exception of the Salmon River (Fig. 3). This river belongs to the Thompson River watershed, and coho salmon from this region have previously been observed to cluster in a similar manner (Rougemont *et al.* 2020). Clustering otherwise appeared to be dependent on latitude. However, isolation-by-distance was not detected (P -value = 0.068). In this study, we had fewer sampling sites and fewer samples in general than previous studies that did find a significant correlation of genetic distance and geographic distance (Rougemont *et al.* 2020, 2022).

The Salmon River group had the lowest private allele count (1,876 vs. a median of 4,188) and observed heterozygosity (0.22966 vs. a median of 0.285565, Fig. 3). The regions with the highest private allele count appears to be around the Puget Sound (*e.g.* Wallace River, private allele count = 5,546) and the Strait of Georgia (*e.g.* Capilano River, private allele count = 6,415). Most of the northern rivers have low private allele counts with the exception of the Kitimat River (private allele count = 6,341), which has the second highest count (Fig. 3).

To put the nucleotide variation generated by isolation-by-distance [identified previously (Rougemont *et al.* 2020, 2022), but not in the current study] into a broader context, we identified possible times when northern populations could have recolonized after the last glaciation period. By modeling demographic histories from resequenced genomes with the SMC++ program, we were able to identify major decreases in effective population size (N_e) that correspond with the Cordilleran Ice Sheet maximum and the presumed penultimate global glacial maximum (Fig. 4). We also observed that for some sampling sites, mostly northern,

there was an additional drop in effective population size between 3,750 and 8,000 years ago (Fig. 4).

Discussion

The BUSCO metric of this genome assembly suggests that it contains a near complete representation of the gene content of this species, and is comparable to model species. The other metrics demonstrate that this genome assembly is contiguous and suitable for downstream analyses. We performed several analyses on the genome assembly itself to identify homeologous regions and repetitive elements. We then measured several basic population metrics and modeled demographic histories from resequenced genomes.

As with previous analyses of salmonid genomes (Allendorf and Thorgaard 1984; Berthelot *et al.* 2014; Lien *et al.* 2016; Christensen *et al.* 2018, 2020; De-Kayne *et al.* 2020), the retention of duplicated chromosomes from the salmonid-specific whole genome duplication (Allendorf and Thorgaard 1984) and the retention of repetitive elements are defining features of the coho salmon genome. Some of the duplicated regions have likely retained very high sequence similarity for roughly 90 million years [time estimate from (Allendorf and Thorgaard 1984; Berthelot *et al.* 2014; Macqueen and Johnston 2014)]. A possible mechanism for high sequence similarity retention is through tetrasomic inheritance (Allendorf *et al.* 2015).

After completing the high-quality genome assembly, we measured population metrics from 83 resequenced genomes. One of the striking features of the PCA of coho salmon populations was how divergent Salmon River salmon were to all other populations. The Salmon River is part of the Thompson River watershed and coho salmon from this system were thought to be isolated from all other populations for potentially 150,000 years before secondary contact roughly 13,500 years ago (essentially during the previous glacial period) (Rougemont *et al.* 2020). This would be consistent with findings in kokanee (*O. nerka*, a landlocked sockeye salmon ecotype) in the upper Columbia River that similarly appear divergent from all other populations of sockeye salmon and kokanee (Christensen *et al.* 2020). Taken together, these pieces of evidence might provide support for a glacial refugium near the intersection of the Cordilleran Ice Sheet and the Laurentide Ice Sheet.

It is also possible that another unknown factor was influencing past analyses and the PCA from the current study. The Salmon River coho salmon have reduced heterozygosity and reduced private alleles, which are indicators of a recent and extensive bottleneck. We were also able to infer the demographic history from whole genome sequences of the Salmon River coho salmon and found evidence of a bottleneck (from $\sim N_e$ 16,227 to $\sim N_e$ 1,749) occurring in this population around 4,000 years ago. We only collected samples from one tributary of the Thompson River (a part of a much larger basin) and can only suggest that a plausible hypothesis from this data is that recolonization of the Salmon River from a small founding population took place after glaciers receded. We did not account for the influence of hatcheries, which could also influence some of the metrics discussed above. Also, we did not incorporate linked selection in demographic modeling as the type of analysis that we used was not amenable. Without linked selection accounted for, there could be biases in times and effective population sizes from our estimates (Pouyet *et al.* 2018).

Other streams clustered in the PCA based largely on latitude for both PC1 and PC2 of the PCA. With a much more extensive

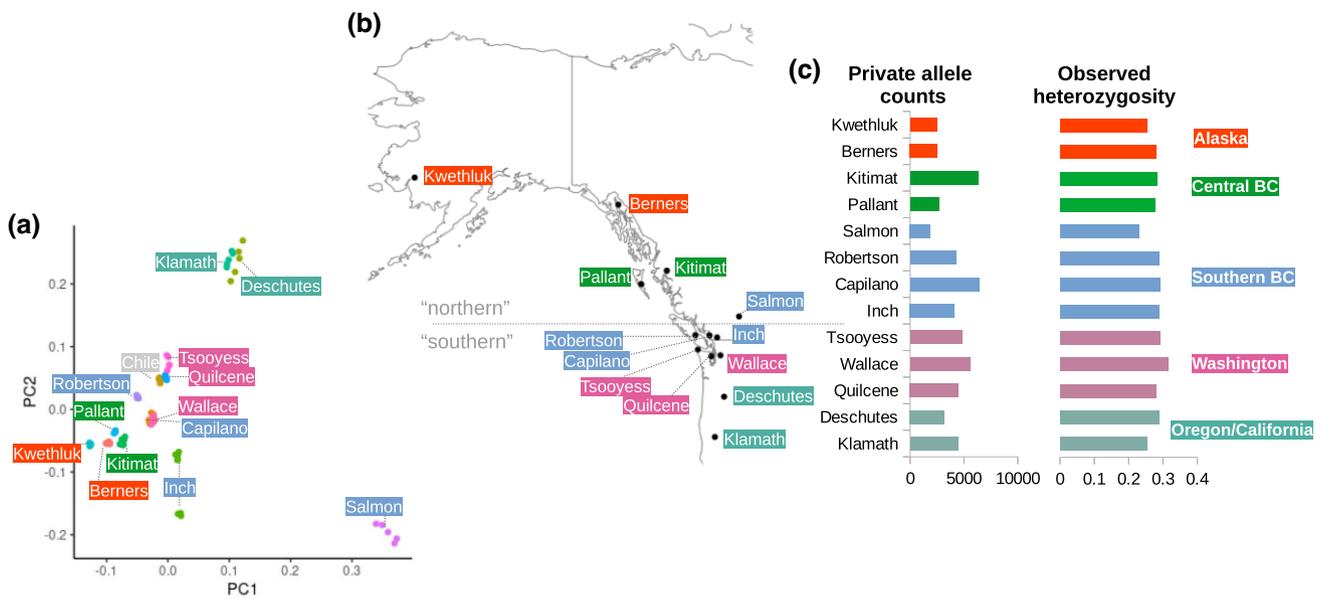


Fig. 3. Summary statistics of coho salmon sampling sites. a) A PCA of coho salmon individuals [plotted using ggplot (Wickham 2016)]. Individuals from the same sampling site have the same color. b) A map of the various river sampling sites of this study [plotted with the maps (Becker et al. 2018) package in R]. c) Private allele counts and observed heterozygosity values are also displayed to the side as bar graphs.

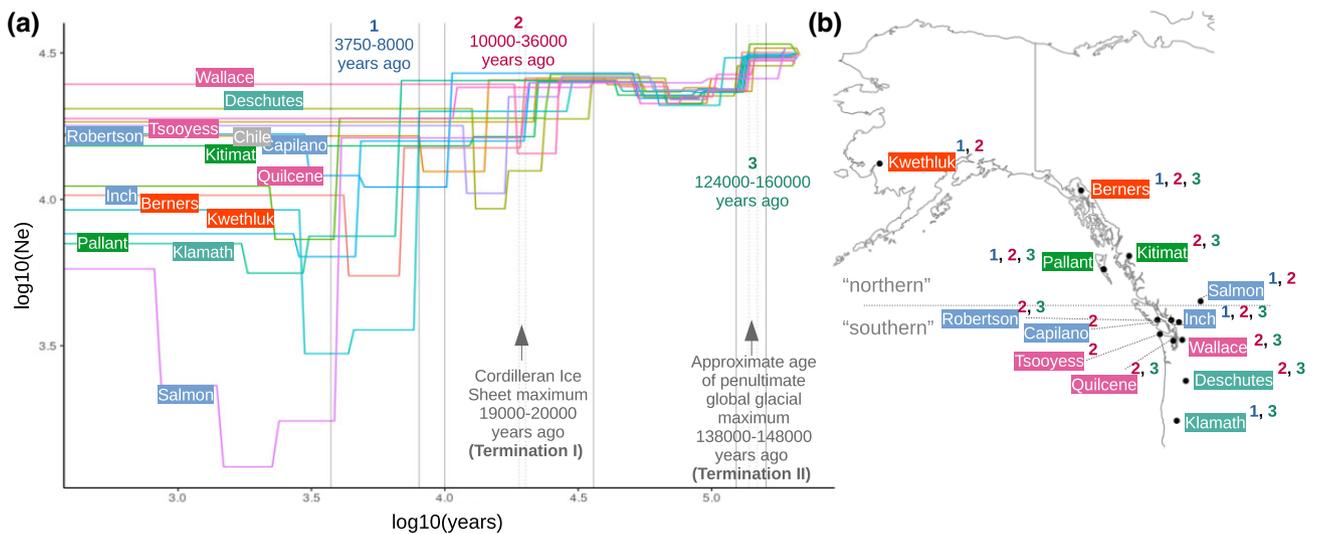


Fig. 4. Demographic histories of coho salmon based on genome resequencing. a) Each labeled line represents multiple individuals from the same river or strain. The X-axis represents calendar years based on a generation time of 3 years for coho salmon. The Y-axis is the effective population size (N_e) estimate. The estimated age of the Cordilleran ice sheet maximum was taken from (Clark et al. 2009). The approximate age of the last interglacial period was based on (Chappellaz et al. 1990). b) For each location, a number nearby indicates a drop of at least 5,000 in the N_e for one of the time intervals noted in A.

sampling strategy, Rougemont et al. (2020) found a similar trend and tested various demographic histories. The authors of that study found that the best supported model was a glacial refugium to the south with recolonization of the northern streams after glacial retreat—generating genomic signatures of isolation-by-distance. In our study we did not detect significant isolation-by-distance, but we had fewer sampling sites and samples in general. The private allele analysis identified that most of the northern streams did have low private allele counts compared to southern streams, which would be consistent with the hypothesis suggested by Rougemont et al. (2020).

From inferred demographic histories, we were able to estimate a recolonization date of some northern streams (based on the

founder effect that would be expected to accompany recolonization) to between 3,750–8,000 years ago. These values are based on assumptions of a mutation rate of $8e-9$ bp/generation and a generation time of 3 years. Linked selection may also bias our time and effective population estimates (Pouyet et al. 2018) as we did not account for them in modeling. However, these recolonization estimates are in-line with recolonization estimates of 4,000–12,000 based on continental glaciation and geographic data proposed previously (Beechie et al. 2001; Waples et al. 2008).

While it is important to remember that time and population estimates are influenced by many factors when inferring demographic histories from sequence data, multiple lines of evidence can be used to strengthen these inferences or put them in a

more realistic context. Radiometric evidence supports that the Cordilleran Ice Sheet maximum occurred between 19,000 and 20,000 years ago (Clark *et al.* 2009). Likewise, chemical properties of gases in Antarctic ice cores support the termination of the last glaciation period (Termination I) to roughly the same time period, as well as a previous termination of the penultimate glaciation period around 138,000 and 148,000 years ago (Termination II) (Kawamura *et al.* 2007). In the demographic histories of the coho salmon, we noted dramatic declines of nearly all salmon populations for both these time periods. This observation supports the parameters used for modeling the demographic histories as we expect that populations might decline in response to increased glaciation or rapid climate change.

The overall trend we observed from modeling demographic histories was major drops in effective population size at each transition from glaciation to inter-glaciation period with increases for nearly all populations after the penultimate glaciation period and uncommon increases for specific rivers after the most recent glaciation period. At a species level, these transitional drops in effective population size likely influence multiple aspects of coho salmon biology since genetic variability can contribute to many characteristics of a species.

Conclusions

In this study, we generated a high-quality reference genome assembly as a tool for conservation and management of coho salmon. Additionally, we resequenced the genomes of a wide distribution of coho salmon from rivers along North America to identify nucleotide variants that will have significant utility for other analyses of coho salmon genetics. We modeled demographic histories of the coho salmon with this data and estimated recolonization occurred between 3,750–8,000 years ago for some northern streams, and observed that major reductions in effective population size were related to changes between glacial and interglacial periods. This coho salmon genome assembly will facilitate research to better understand coho salmon biology in the future and will enhance management of this culturally and economically important species.

Data availability

Raw data for the genome assembly was submitted to the NCBI under the BioProject PRJNA352719. Whole genome resequencing data was submitted under PRJNA401427 and PRJNA808051 to the NCBI BioProjects (see [Supplementary File S1](#) for specific samples used in this study). The VCF file used for analyses in this study is available at figshare: <https://doi.org/10.25387/g3.20032082>.

[Supplemental material](#) available at G3 online.

Acknowledgements

We would like to acknowledge the extensive help from many individuals who contributed to work in sampling and sequencing in this study. Sample collection involved a number of people, many of which remain anonymous to the authors, but we thank them for all their efforts. Coordinating sample collections were Heather Hoyt of the Alaskan Department of Fish and Game for samples from Alaska, Christian Smith at the US Fish and Wildlife's Abernathy Fish technology Center for samples in Washington State and Oregon, John Carlos Garza at the NOAA Southwest Marine Fisheries Sciences Centre for samples collected in Klamath River, Justin Henry and Bruce Swift for samples from

Aquaculture broodstock from Northern Divine Aquafarms Ltd (formerly Target Marine), and from Riverence LLC for Aquaculture broodstock. For library preparation and sequencing, we thank the staff at the McGill University and Genome Québec innovation Center as well as the Michael Smith Genome Sciences Centre. We would also like to thank Compute Canada (Cedar) for computational resources. Finally, we would like to thank the editor and reviewers for their helpful comments.

Funding

This project was supported by a Genome Canada strategic grant entitle EPIC4-Enhanced Production in Coho: Culture, Community Catch (grant ID: 229COH, Genome Canada). In addition, funding and contributions were provided by Riverence LLC, Northern Divine Aquafarms. EBR was supported during this work by an NSERC PGSD3 grant, and both EBR and KAC were supported by a Natural Sciences and Engineering Research Council of Canada Visiting Fellow in a Government Laboratory (DFO) fellowship. MTC was supported by Simon Fraser University Graduate Dean Entrance Scholarship, SFU Provost Prize of Distinction and Garfield Weston Foundation. Funding from the Canadian Regulatory System for Biotechnology to RHD was provided to support production of doubled haploids.

Author contributions

E.B.R., K.A.C., J.S.L., and B.F.K. performed genome assembly, chromosome assembly, genome submission, and generated genome metrics. E.B.R., C.A.D., M.T.C., A.M., and D.S. performed wet-lab work including DNA extractions and mitochondrial sequencing. E.B.R., Q.R., E.N., D.R.M., and J.S.L. performed SNP calling and population genomic analyses. R.H.D., M.T.C., D.S., and C.B. generated, raised, and dissected doubled-haploid samples for the genome assembly and transcriptome. R.E.B., T.D.B., K.A.N., and J.M.Y. provided samples used in resequencing work. K.A.N. provided early access to linkage map and additional guidance on its use. R.H.D., R.E.W., T.D.B., K.A.N., J.M.Y., R.N., L.B., W.S.D., S.J.M.J., and B.F.K. initiated, planned and supervised the project. E.B.R., K.A.C., D.R.M., and B.F.K. wrote first draft of the manuscript.

Conflicts of interest statement

The authors declare no conflict of interest.

Literature cited

- Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J Hered.* 2015;106(3):217–227. doi:10.1093/jhered/esv015.
- Allendorf FW, Thorgaard GH. Tetraploidy and the evolution of salmonid fishes. In: Turner BJ, editors. *Evolutionary Genetics of Fishes*. Monographs in Evolutionary Biology. US: Springer; 1984. p. 1–53.
- Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 2019;20(1):224. doi:10.1186/s13059-019-1829-6.
- Barría A, Christensen KA, Yoshida G, Jedlicki A, Leong JS, Rondeau EB, Lhorente JP, Koop BF, Davidson WS, Yáñez JM. Whole genome

- linkage disequilibrium and effective population size in a coho salmon (*Oncorhynchus kisutch*) breeding population using a high-density SNP array. *Front Genet.* 2019;10:498. doi:10.3389/fgene.2019.00498.
- Beacham TD, Candy JR, Supernault KJ, Ming T, Deagle B, Schulze A, Tuck D, Kaukinen KH, Irvine JR, Miller KM. Evaluation and application of microsatellite and major histocompatibility complex variation for stock identification of coho salmon in British Columbia. *Trans Am Fish Soc.* 2001;130(6):1116–1149. <1116: EAAOMA>2.0.CO;2.
- Beacham TD, Wetklo M, Deng L, MacConnachie C. Coho salmon population structure in North America determined from microsatellites. *Trans Am Fish Soc.* 2011;140(2):253–270. doi:10.1080/00028487.2011.558782.
- Beamish RJ, McFarlane GA, Thomson RE. Recent declines in the recreational catch of coho salmon (*Oncorhynchus kisutch*) in the strait of Georgia are related to climate. *Can. J. Fish. Aquat. Sci.* 2011; 56(3):506–515. doi:10.1139/f98-195.
- Beamish RJ, Sweeting RM, Lange KL, Noakes DJ, Preikshot D, Neville CM. Early marine survival of coho salmon in the strait of Georgia declines to very low levels. *Mar Coast Fish.* 2010;2(1):424–439. doi: 10.1577/C09-040.1.
- Beatty GE, Provan J. Refugial persistence and postglacial recolonization of North America by the cold-tolerant herbaceous plant *Orthilia secunda*. *Mol Ecol.* 2010;19(22):5009–5021. doi:10.1111/j.1365-294X.2010.04859.x.
- Becker RA, Wilks AR, Brownrigg R, Minka TP, Deckmyn A. maps: Draw Geographical Maps 2018.
- Beechie TJ, Collins BD, Pess GR. Holocene and recent geomorphic processes, land use, and salmonid habitat in two north puget sound river basins. *Geomorphic Processes and Riverine Habitat.* Washington DC, US, : American Geophysical Union (AGU); 2001. p. 37–54.
- Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 2014;5(1): 3657. doi:10.1038/ncomms4657.
- Biagi CA, Leggatt RA, Sakhrani D, Wetklo M, Vandersteen WE, Christensen KA, Rondeau EB, Watson BM, Wellband KW, Koop BF, et al. Timing of postfertilization pressure shock treatment for the production of mitotic gynogens in six salmonid species. *North Am. J. Aquac.* 2022;84(4):505–515. doi:10.1002/naaq.10266.
- Bougéard S, Dray S. Supervised multiblock analysis in R with the ade4 package. *J Stat Softw.* 2018;86(1):1–17. doi:10.18637/jss.v086.i01.
- Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping loci De Novo from short-read sequences. *G3 (Bethesda) GenesGenomesGenetics.* 2011;1(3): 171–182. doi:10.1534/g3.111.000240.
- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. *Mol Ecol.* 2013;22(11):3124–3140. doi:10.1111/mec.12354.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015;4(1):7. doi:10.1186/s13742-015-0047-8
- Chappellaz J, Barnola JM, Raynaud D, Korotkevich YS, Lorius C. Ice-core record of atmospheric methane over the past 160,000 years. *Nature.* 1990;345(6271):127–131. doi:10.1038/345127a0.
- Chourrout D. Pressure-induced retention of second polar body and suppression of first cleavage in rainbow trout: production of all-triploids, all-tetraploids, and heterozygous and homozygous diploid gynogenetics. *Aquaculture.* 1984;36(1–2):111–126. doi:10.1016/0044-8486(84)90058-9.
- Christensen KA, Leong JS, Sakhrani D, Biagi CA, Minkley DR, Withler RE, Rondeau EB, Koop BF, Devlin RH. Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS One.* 2018;13(4): e0195461. doi:10.1371/journal.pone.0195461.
- Christensen KA, Rondeau EB, Minkley DR, Sakhrani D, Biagi CA, Flores A-M, Withler RE, Pavey SA, Beacham TD, Godin T, et al. The sockeye salmon genome, transcriptome, and analyses identifying population defining regions of the genome. *PLoS One.* 2020;15(10):e0240935. doi:10.1371/journal.pone.0240935.
- Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM. The last glacial maximum. *Science.* 2009;325(5941):710–714. doi:10.1126/science.1172873.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15): 2156–2158. doi:10.1093/bioinformatics/btr330.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021;10(2): giab008. doi:10.1093/gigascience/giab008.
- De-Kayne R, Zoller S, Feulner PGD. A de novo chromosome-level genome assembly of coregonus sp. “balchen”: one representative of the Swiss alpine whitefish radiation. *Mol Ecol Resour.* 2020; 20(4):1093–1109. doi:10.1111/1755-0998.13187.
- Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw.* 2007;22(4):1–20. doi:10.18637/jss.v022.i04.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. De novo assembly of the aedes aegypti genome using hi-C yields chromosome-length scaffolds. *Science.* 2017;356(6333):92–95. doi:10.1126/science.aal3327.
- Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for hi-C contact maps with unlimited zoom. *Cell Syst.* 2016a;3(1): 99–101. doi:10.1016/j.cels.2015.07.012.
- Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution hi-C experiments. *Cell Syst.* 2016b;3(1):95–98. doi:10.1016/j.cels.2016.07.002.
- Government of Canada N. R. C. GEOSCAN Search Results: Fastlink. geoscan.nrcan.gc.ca 2015.
- Heard WR. Overview of salmon stock enhancement in southeast Alaska and compatibility with maintenance of hatchery and wild stocks. *Environ Biol Fishes.* 2012;94(1):273–283. doi:10.1007/s10641-011-9855-6.
- Irvine JR, Fukuwaka M. Pacific salmon abundance trends and climate change. *ICES J. Mar. Sci.* 2011;68(6):1122–1130. doi:10.1093/icesjms/fsq199.
- Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM, et al. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics.* 2018;19(1):393. doi:10.1186/s12859-018-2425-6.
- Jombart T. ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinforma. Oxf. Engl.* 2008;24(11):1403–1405. doi:10.1093/bioinformatics/btn129.
- Jombart T, Ahmed I. ADEGENET 1.3–1: new tools for the analysis of genome-wide SNP data. *Bioinformatics.* 2011;27(21):3070–3071. doi:10.1093/bioinformatics/btr521.
- Kawamura K, Parrenin F, Lisiecki L, Uemura R, Vimeux F, Severinghaus JP, Hutterli MA, Nakazawa T, Aoki S, Jouzel J, et al.

- Northern hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature*. 2007;448(7156):912–916. doi:10.1038/nature06015.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinforma. Oxf. Engl.* 2012;28(12):1647–1649. doi:10.1093/bioinformatics/bts199.
- Knaus BJ, Grünwald NJ. Vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour.* 2017;17(1):44–53. doi:10.1111/1755-0998.12549.
- Kodama M, Briec MSO, Devlin RH, Hard JJ, Naish KA. Comparative mapping between coho salmon (*Oncorhynchus kisutch*) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. *G3 (Bethesda) Genes Genomes Genet.* 2014;4(9):1717–1730. doi:10.1534/g3.114.012294
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722–736. doi:10.1101/gr.215087.116.
- Kristianson G, Strongtharm D. The Evolution of Recreational Salmon Fisheries in British Columbia. Pacific Fisheries Resource Conservation Council; 2006. p. 32.
- Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009. doi:10.1101/gr.092759.109.
- Latch EK, Heffelfinger JR, Fike JA, Rhodes OE Jr. Species-wide phylogeography of North American mule deer (*Odocoileus hemionus*): cryptic glacial refugia and postglacial recolonization. *Mol Ecol.* 2009;18(8):1730–1745. doi:10.1111/j.1365-294X.2009.04153.x.
- Leong JS, Jantzen SG, von Schalburg KR, Cooper GA, Messmer AM, Liao NY, Munro S, Moore R, Holt RA, Jones SJM, et al. *Salmo salar* and *esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *BMC Genomics.* 2010;11(1):279. doi:10.1186/1471-2164-11-279.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* 2013.
- Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics.* 2014;30(20):2843–2851. doi:10.1093/bioinformatics/btu356.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* 2018;34(18):3094–3100. doi:10.1093/bioinformatics/bty191.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinforma. Oxf. Engl.* 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
- Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. The atlantic salmon genome provides insights into rediploidization. *Nature.* 2016;533(7602):200–205. doi:10.1038/nature17164.
- Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B Biol. Sci.* 2014;281(1778):20132881. doi:10.1098/rspb.2013.2881.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–1303. doi:10.1101/gr.107524.110.
- Moore J-S, Bajno R, Reist JD, Taylor EB. Post-glacial recolonization of the North American Arctic by Arctic char (*Salvelinus alpinus*): genetic evidence of multiple northern refugia and hybridization between glacial lineages. *J Biogeogr.* 2015;42(11):2089–2100. doi:10.1111/jbi.12600.
- Natural Earth www.naturalearthdata.com.
- O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. Nxtrim: optimized trimming of illumina mate pair reads. *Bioinformatics.* 2015;31(12):2035–2037. doi:10.1093/bioinformatics/btv057.
- Pasos, M. Commission for Environmental Cooperation—watersheds. www.cec.org.
- Pearse DE, Barson NJ, Nome T, Gao G, Campbell MA, et al. Sex-dependent dominance maintains migration supergene in rainbow trout. *Nat. Ecol. Evol.* 2019;3(12):1731–1742. doi:10.1038/s41559-019-1044-6
- Pitman KJ, Moore JW, Sloat MR, Beaudreau AH, Bidlack AL, Brenner RE, Hood EW, Pess GR, Mantua NJ, Milner AM, et al. Glacier retreat and pacific salmon. *BioScience.* 2020;70(3):220–236. doi:10.1093/biosci/biaa015.
- Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences (K. Veeramah, P. J. Wittkopp, & I. Gronau, eds.). *eLife.* 2018;7:e36317. doi:10.7554/eLife.36317
- QGIS Development Team. QGIS_software 2022.
- Quinn TP. A review of homing and straying of wild and hatchery-produced salmon. *Fish Res.* 1993;18(1–2):29–44. doi:10.1016/0165-7836(93)90038-9.
- R Core Team. R: A Language and Environment for Statistical Computing 2020.
- Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics.* 2018;19(1):460. doi:10.1186/s12859-018-2485-7.
- Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF. The genome and linkage map of the northern pike (*esox lucius*): conserved synteny revealed between the salmonid sister group and the neoteleostei. *PLOS ONE.* 2014;9(7):e102089. doi:10.1371/journal.pone.0102089.
- Rougemont Q, Moore J-S, Leroy T, Normandeau E, Rondeau EB, Withler RE, Van Doornik DM, Crane PA, Naish KA, Garza JC, et al. Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed pacific salmon. *PLOS Genet.* 2020;16(8):e1008348. doi:10.1371/journal.pgen.1008348.
- Rougemont Q, Xuereb A, Dallaire X, Moore J-S, Normandeau E, Perreault-Payette A, Bougas B, Rondeau EB, Withler RE, Van Doornik DM, et al. Long-distance migration is a major factor driving local adaptation at continental scale in coho salmon. *Mol Ecol.* 2022;32(3):542–559. doi:10.1111/mec.16339.
- Sahlin K, Chikhi R, Arvestad L. Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics.* 2016;32(13):1925–1932. doi:10.1093/bioinformatics/btw064.
- Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. BESST—efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics.* 2014;15(1):281. doi:10.1186/1471-2105-15-281.
- Sambrook J, Fritsch EF, Maniatis T. *Molecular Cloning: A Laboratory Manual*. Third Edition. New York: Cold Spring Harbor Laboratory Press; 2000.
- Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet.* 2017;49(2):303–309. doi:10.1038/ng.3748.
- Thioulouse J, Dray S, Dufour A-B, Siberchicot A, Jombart T, et al. *Multivariate Analysis of Ecological Data with ade4*. NY: Springer, New York; 2018.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J,

- et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 2013;43(1):11.10.1–11.10.33. doi:10.1002/0471250953.bi1110s43.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963. doi:10.1371/journal.pone.0112963.
- Waples RS, Pess GR, Beechie T. Evolutionary history of pacific salmon in dynamic environments. *Evol Appl.* 2008;1(2):189–206. doi:10.1111/j.1752-4571.2008.00023.x.
- Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, Birol I. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience.* 2015;4(1):35. doi:10.1186/s13742-015-0076-3.
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag; 2016.
- Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. *Bioinforma. Oxf. Engl.* 2018;34(5):725–731. doi:10.1093/bioinformatics/btx675.
- Zhang H, Tan E, Suzuki Y, Hirose Y, Kinoshita S, Okano H, Kudoh J, Shimizu A, Saito K, Watabe S, et al. Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Sci Rep.* 2014;4(1):6780. doi:10.1038/srep06780.

Editor: T. Pyhäjärvi