Genome assembly, transcriptome and SNP database for chum salmon (*Oncorhynchus keta*)

Eric B. Rondeau[1,2,3*], Kris A. Christensen[1,2*], Dionne Sakhrani[1], Carlo A. Biagi[1], Mike Wetklo[3], Hollie A. Johnson[2], Cody A. Despins[2], Rosalind A. Leggatt[1], David R. Minkley[2], Ruth E. Withler[3], Terry D. Beacham[3], Ben F. Koop[2], Robert H. Devlin[1§]

[1]Fisheries and Oceans Canada, 4160 Marine Dr., West Vancouver, British Columbia, V7V 1N6, Canada

[2]Department of Biology, University of Victoria, Victoria, British Columbia, V8W 3N5, Canada

[3]Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, British Columbia, V9T 6N7, Canada

[§]Corresponding author

*Authors contributed equally to results of manuscript

Email addresses:

EBR: eric.rondeau@dfo-mpo.gc.ca

KAC: kris.christensen@wsu.edu

DS: Dionne.Sakhrani@dfo-mpo.gc.ca

CAB: Carlo.Biagi@dfo-mpo.gc.ca

MW: mike.wetklo@dfo-mpo.gc.ca

HAJ: holliej@uvic.ca

CAD: cdespins@uvic.ca

RAL: Rosalind.Leggatt@dfo-mpo.gc.ca

DRM: dminkley@uvic.ca

REW: RWithler@shaw.ca

TDB: Terry.Beacham@dfo-mpo.gc.ca

BFK: bkoop@uvic.ca

RHD: Robert.Devlin@dfo-mpo.gc.ca

Abstract

Chum salmon (*Oncorhynchus keta*) is the species with the widest geographic range of the anadromous Pacific salmonids,. Chum salmon is the second largest of the Pacific salmon, behind Chinook salmon, and considered the most plentiful Pacific salmon by overall biomass. This species is of significant commercial and economic importance: on average the commercial chum salmon fishery has the second highest processed value of the Pacific salmon within British Columbia. The aim of this work was to establish genomic baseline resources for this species. Our first step to accomplish this goal was to generate a chum salmon reference genome assembly from a doubled-haploid chum salmon. Gene annotation of this genome was facilitated by an extensive RNA-seq database we were able to create from multiple tissues. Range-wide resequencing of chum salmon genomes allowed us to categorize genome-wide geographic variation, which in turn reinforced the idea that genetic differentiation was best described on a regional, rather than at a stock-specific, level. Within British Columbia, chum salmon regional groupings were described at the conservation unit (CU) level, and there may be substructure within particular CUs. Genome wide associations of phenotypic sex to SNP genetic markers identified two clear peaks, a very strong peak on Linkage Group 15, and another on Linkage Group 3. With these new resources, we were better able to characterize the sex-determining region and gain further insights into sex determination in chum salmon and the general biology of this species.

Keywords

Chum salmon, Genome assembly, *Oncorhynchus keta*, RNA-seq, Resequencing, SNP database.

Background

Pacific salmon of the genus *Oncorhynchus* are iconic, culturally important keystone species spawning across freshwater watersheds that feed the Northern Pacific Ocean. Predominately anadromous, members of most species spend years at sea, consuming marine nutrients that are eventually deposited into coastal ecosystems where they provide a valuable source of food to numerous marine and terrestrial species as the salmon spawn and then die [1].

Chum salmon (*Oncorhynchus keta*) are the second largest of the Pacific salmonids and may have historically represented up to 50% of the salmonid biomass in the Pacific Ocean [2]. It is the most widely distributed of the Pacific salmonid species [3, 4], with spawning grounds ranging from Japan and the eastern coast of the Korean Peninsula through to Northern Russia, and from the Mackenzie River south through Central California in North America [5]. Among the most significant species of Pacific salmon in commercial fisheries – in an analysis of British Columbian commercial fisheries 2012-2015, chum salmon was the most plentiful species by weight in 3 out of 4 years analyzed, and second most valuable by processed value when averaged across the four year period ($31 million per year) [6].

A key and fascinating biological feature in salmonids is homing, whereby adults demonstrate an ability to return to the same riverine sites where they were spawned, although not all species show the same degree of site fidelity (reviewed in [7]). Some species, such as Sockeye, have been observed to return to within metres of where they were hatched (e.g., [8]), but other species vary in their fidelity to site of return and stray rate. Reasons for straying are likely varied (reviewed in [7]), but significant factors are thought to be juvenile freshwater residence time and freshwater migration distance, both of which lead to reduced imprinting. With chum salmon having relatively short freshwater residence (they migrate to sea as fry) and short migration distances (on average), it is perhaps not surprising that chum tend to have higher than average stray rates among the Pacific salmonids [7]. The consequences of such straying are that while regional-level differentiation (e.g., [9, 10]) and run-timing differentiation between summer and fall runs (e.g., [11–13]) can be observed, population-level genetic differentiation is not often seen within chum salmon.

The genomes of salmonids, including chum salmon, possess a key feature shared by all salmonid genomes, a salmon-lineage specific whole-genome duplications (WGD). WGDs very likely play one of the more significant roles in evolutionary innovation [14–17] and are found in plants (reviewed in [18]), fungi [19, 20] , arthropods [21, 22], basal vertebrates ~500 million years ago (mya) [15, 23, 24], fishes ~300 mya [25–27], and more recently in ancestral salmonids ~90 mya [28, 29] . These major genome expansions have been proposed to allow for adaptations to new niches or conditions, particularly in times of major environmental change (reviewed in [30]). The occurrence of over 70 different salmonid species lineages stemming from the relatively recent ancestral WGD [29] offers a valuable system to i) observe evolutionary consequences of a relatively recent autopolyploid WGD, ii) identify ensuing mechanisms for regaining stable meiosis and cell division by regaining a functional diploid state through re-diploidization, and iii) draw associations between mechanisms of re-diploidization to potential genetic specialization that allow for species adaptation such as disease resistance. Additionally, each species has evolved unique morphology, life history strategies, and responses to common salmon pathogens (e.g., varied resistance to salmon aquaculture from pathogens such as the sea louse [31, 32]). This phenotypic variety provides future opportunities for exploring the biology and genetics behind the genomic architecture of whole-genome duplication have shaped these unique species.

The presence of these duplications, however, can present major technological challenges to genome assembly, due to limited differentiation between duplicated portions of the genome. Salmonids offer additional hurdles in that a significant portion of the genome still remains in a tetraploid-like state [33, 34], and may show lineage-specific re-diploidization patterns [35], or chromosome architecture through species-specific fusions [36]. While many challenges remain, the technological barriers to assembly of salmonid genomes are beginning to fall, as evidenced by the relatively rapid recent release of salmonid genomes [37–44]. A fully-annotated reference chum salmon genome will enhance development of genomics-based technologies to improve the effectiveness of fisheries management of the wild chum salmon fishery. This has already been performed for other Pacific salmon species in British Columbia (e,g., [45, 46]), and a genome assembly for chum would provide the ability to adopt similar management tools based on emerging high-throughput sequencing technologies.

Genetic resources in chum salmon have, as in many other species, been in a state of transition as genetic tools have advanced and become more widespread. Early work on population genetic structure in chum salmon utilized allozymes [47, 48] and microsatellite markers [9, 49] and provided the first range-wide studies on genetic diversity [10]. Recently, genetic stock identification tools have been shifting from microsatellites to single-nucleotide polymorphisms (SNPs), providing increased accuracy of genetic discrimination with increasing marker numbers [50]. Early identification of SNPs in chum salmon [51–55] led to the development of a SNP panel for assessing genetic diversity and population structures in chum salmon [13]; development of expanded SNP panels for fisheries management continues to occur with increased marker density and improving genetic baselines allow for increased power ([56]; Beacham T.D. and Sutherland B.J.G, Personal Communication). Restriction-site Associated DNA sequencing (RADseq) has recently enabled a much more rapid throughput for SNP discovery [57, 58], and studies in chum have utilized this technique to enable researchers to develop linkage maps to explore regions of residual inheritance associated with the aforementioned genome duplication event [34]. This advance in technology has further allowed for the identification of extended patterns of linkage disequilibrium, demonstrating the power of increased marker density on the identification of genomic features of large effect [59]. Despite this significant effort, unlike in other *Oncorhynchus* species (e.g., Rainbow trout [60]; Chinook salmon [41], sockeye salmon [61]), neither a whole-genome catalog of SNP markers nor whole-genome resequencing data has been available as a resource for chum salmon to date. The development of such a resource will further allow genetic resources, such as SNP panels, to be placed in context relative to genes or other annotated genomic features.

In this work, we have sequenced and assembled the genome of a mitotic gynogen doubled haploid chum salmon to eliminate allelic variation but retain paralog differences. Extensive multi-tissue RNA-seq was generated to provide the base for annotation of the genome as well as a tissue-specific expression atlas for future comparative studies. Finally, whole-genome resequencing was performed across 59 individual chum salmon from a select distribution of the species' range to catalogue genome-wide diversity in this species. The utility of the dataset is further demonstrated by the genetic association of the sex phenotype onto the expected chromosome in a narrow window of elevated linkage disequilibrium.

Methods

Data availability

All raw sequencing reads and the assembled genome described in this project have been submitted to NCBI under BioProject PRJNA556729. SNP variant sets described below are available through Dryad repository.

Comment [RE1]: TBD on acceptance

Animal care and sample collection

All animals were reared in compliance with Canadian Council on Animal Care Guidelines, under oversight from the Fisheries and Oceans Canada Pacific Region Animal Care Committee (PRACC). Chum salmon for genome sequencing and assembly and for transcriptome assembly were from Chehalis River Hatchery parents and reared at Fisheries and Oceans Canada in West Vancouver. Chum salmon mitotic gynogen doubled haploids were produced following procedures described by [62]. Briefly, eggs were fertilized with UV-irradiated sperm and pressure shocked (10,000 psi for 5 minutes) in batches at 30 min intervals between 4 and 7 hours post-fertilization. One individual from the 7h pressure shock group (Oke142-1, NCBI BioSample: SAMN12367893; Supplementary Table 1) was confirmed to be homozygous for maternal alleles using a panel of 14 microsatellites [49], and was used for genome sequencing and assembly (see below). The individual was euthanized in a bath of 200 mg/L tricaine methanesulfonate (TMS) buffered in 400 mg/L sodium bicarbonate prior to first feeding stage, and stored in ethanol before DNA extraction and whole genome sequencing.

For transcriptomic data, control Chehalis River Hatchery chum salmon produced from the same parents as Oke142-1 but without UV milt treatment or pressure shock were grown in aerated fresh well water in 200–3700 L tanks and fed hourly as fry and to satiation 3 times daily as parr with stage-appropriate manufactured salmon feed (Skretting Canada Ltd.). At approximately 7 months post-ponding, a single selected chum female (86.9g with a 19.3cm fork length) was euthanized with TMS as above, then rapidly (< three min, PRACC management procedure 3.7) team dissected to harvest 18 tissues (see Supplementary Table 2) for RNA extraction, with an additional tissue (testes) sampled from an juvenile male. All tissues were stored in RNAlater at -20°C until extraction. RNA extractions were performed using the Qiagen RNeasy Mini Kit following the manufacturer's protocol.

For individuals used in resequencing, samples were obtained primarily through non-lethal sampling of fin clips or operculum punches from Fisheries and Oceans Canada hatchery brood programs. Additional samples were obtained from archived tissue sets used for genetic stock ID baseline development to supplement the dataset. In total, 59 individuals were utilized in this assessment, with DNA obtained via Qiagen DNeasy Animal tissue kit's following manufacturer's protocol) or phenol/chloroform extractions (following Thermo Fisher Scientific's protocol for genomic DNA preparation [63]. Tissue types, sex, collection dates and locations are summarized in Supplementary Table 3.

Genome sequencing and Assembly

DNA was isolated from RNAlater or ethanol preserved tissues using a phenol/chloroform extraction as per Thermo Fisher Scientific's protocol for genomic DNA preparation [63]. Extracted DNA was submitted for genome sequencing across multiple library types, using both Illumina and PacBio sequencing instruments (summarized in Supplementary Table 1: SRA chum Gynogen). Extracted DNA was submitted to the McGill University and Génome Québec Innovation Centre (now the Centre d'expertise et de services Génome Québec) for construction of overlapping (library size estimate = 497 base pairs (bp) and non-overlapping (library size estimate = 620bp) IDT dual-indexed Illumina Shotgun libraries. Each library was sequenced twice on an Illumina HiSeq2500 on RAPID mode PE250. Extracted DNA was also submitted to the McGill University and Génome Québec Innovation Centre for construction of a single library of 10X Chromium linked-reads. Following library construction, the library was sequenced across three lanes of Illumina HiSeqX PE150. Extracted DNA was also submitted to the National Research Council Plant Biotechnology Institute Genome Core for Illumina mate-pair library construction and sequencing. Mate-pair libraries targeting 2-3kb, 4-6kb and 7-12kb were constructed, and sequenced on a lane each of Illumina HiSeq2500 PE125. Finally, extracted DNA was submitted to McGill University and Génome Québec Innovation Centre for construction of a Pacific Biosciences SMRT library using a sheared large insert library type, and the MagBead OneCellPerWell v1 collection protocol. The library was ultimately sequenced across 16 total SMRT cells.

Assembly protocols followed successful strategies utilized for Northern Pike e.g., [40, 61, 64, 65]. See Supplementary Table 4 for specific parameters to assembly and trimming that were tested. Reads were first trimmed for quality, adapters and minimum length using Trimmomatic [66], and BBmap's FilterByTile was utilized to remove poorly performing portions of the Illumina reads (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/; [67]). Allpaths-LG v52488 [68] was utilized with overlapping Illumina overlapping PE250 and Illumina mate-pair libraries using a 3.0 TB memory node on the Compute Canada cluster Cedar. Non-overlapping libraries were also included in two assembly attempts, but ultimately exceeded the memory availability on the node in the MergeNeighbourhoods2 module and were dropped in successful assemblies. Assembly parameters were primarily adjusted for coverage of each of the library types as had been performed in other species; additional modifications were made to read filtering to improve the assemblies.

Following Allpaths-LG assembly, scaffolds were passed into PB Jelly 2 v 15.8.24 [69] along with all subreads produced in PacBio sequencing. Nodes on Compute Canada's Cedar cluster were used for all stages, with on-node temp directory and 48 cores used in all steps where allowed. Blasr parameters were `-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -nproc 48 -maxScore -500 –noSplitSubreads`. Extraction.py was modified to `MAXGAPHOLD= 1000000` to take advantage of memory available. Collection.py was run with `–m 3`. All other parameters remained default. Finally, the assembly was polished with Pilon

[70] using the trimmed paired-end data, aligned to the genome utilizing `bwa mem –M` and default parameters.

Scaffolds were ordered and oriented into chromosome representations (i.e., Pseudomolecules) predominately following the methods described in Christensen et al. (2018) [40]. The sequences underlying the markers for the published chum linkage map from Waples et al. (2016) [34] were aligned to the scaffold assembly utilizing BLAST (-outfmt 6, -word_size 48, perc_identity 94, -max_hsps 100, -max_target_seqs 10 -evalue 1E-16). All scaffolds with a link to at least one marker on the map were retained for subsequent pseudomolecule inclusion. Scaffolds were ordered and oriented to the extent allowed by the linkage map, although regions of low recombination limited the effectiveness of the maps alone at this task. Therefore, the sequences underlying the markers for the linkage map were also aligned to a higher contiguity genome of a related species (coho; GCF_002021735.2), and ordering and orientation was further refined based on the conserved synteny between the two species via manual review. Where discrepancies were observed, the chum linkage map was taken as correct to ensure major species-specific rearrangements were captured. Finally, pseudomolecules were aligned to genomes of additional salmonids rainbow trout GCF_002163495.1 [39], Atlantic salmon (GCF_000233375.1) [38], Chinook salmon (GCF_002872995.1) [40] and the non-duplicated outgroup to the salmonids, northern pike (GCF_000721915.3) [71] using Symap v4.2 [72] to ensure linearity was generally conserved, and where it was not, was supported by rearrangements observed in the linkage map.

A BUSCO v4.0.2 [73] analysis utilizing the actinopterygii_odb10 dataset and `-m geno –c 10 –sp zebrafish` was used to analyze the gene representation within the assembly utilizing the RefSeq maintained assembly: GCF_012931545.1.

Gene Annotation

Raw reads for RNA-seq libraries were uploaded into NCBI under BioProject PRJNA556729 for inclusion in the Eukaryotic Genome Annotation pipeline. NEBNext dual-indexed mRNA stranded libraries were constructed from tissues described above by the McGill University and Génome Québec Innovation Centre, and sequenced on a half lane of NovaSeq

6000 S4 PE150 (additional libraries in the lane consisted primarily of RNA-seq of Pink and Chinook salmon from related projects). Sequences were uploaded under: SRP216443, with individual accessions: SRR9841162 (Adipose), SRR9841163 (Brain), SRR9841160 (Gill), SRR9841161 (Head Kidney), SRR9841166 (Heart), SRR9841167 (Hindgut), SRR9841164 (Left Eye), SRR9841165 (Liver), SRR9841168 (Lower Jaw), SRR9841169 (Midgut), SRR9841171 (Ovary), SRR9841172 (Pituitary), SRR9841170 (Pyloric Caeca), SRR9841174 (Red Muscle Skin), SRR9841176 (Spleen), SRR9841177 (Stomach), SRR9841173 (Testes), SRR9841175 (Upper Jaw Nares), and SRR9841178 (White Muscle).

Variant Calling

All individuals sequenced for variant calling (Supplementary Table 3) used Shotgun PCR Free IDT dual-indexed Illumina libraries, produced on a quarter lane of Illumina HiSeqX – library construction and sequencing were performed at the McGill University and Génome Québec Innovation Centre. Raw reads were uploaded to the NCBI BioProject PRJNA556729, with individual accessions listed in the supplementary table.

Variant calling followed the best practices pipeline of GATK 3.8 [74–76], and generally followed the methods previously outlined in Christensen et al. (2020) [61]. Raw paired-end reads were aligned to the scaffold-version of the genome (pre-pseudomolecule construction) using bwa (v0.7.17) mem [77] and the `-M` option. Samtools (v1.9) [78] was used to sort and index the alignment files, while Picard (v2.18.9) [79] was utilized with the MarkDuplicates option to identify likely PCR duplicates, and with ReplaceSamHeader to add read group information to the alignment files. GATK's HaplotypeCaller was then used `--genotyping_mode DISCOVERY –emitRefConfidence GVCF` to generate gvcf files, GenotypeGVCFs was used to generate vcf on intervals, and CatVariants was used to concatenate interval files into a single vcf. A training variant set was generated using a hard-filtered subset of the first round of genotyping, utilizing VariantFiltration and the parameters ` --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"` as well as the VCFtools (v0.1.14) [80] parameters `--maf0.1 –hwe0.01`. A truth set was generated by overlapping the linkage map SNPs with the hard-filtered training set to obtain SNPs found in both methods.

VariantRecalibrator was applied using these sets `-mode SNP -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an InbreedingCoeff`, and ApplyRecalibration run to generate a final SNP set ` --ts_filter_level 99.0`. Finally, as SNP calling began pre-pseudomolecule construction, vcfChromTransfer in the Genomics General repository (https://github.com/simonhmartin/genomics_general; commit: 9d12505) was used to lift over the VCF file based on the NCBI submission AGP file. This lifted-over VCF is included in the accompanying dataset as the "raw SNP" set (referred to as set 1 below).

VCFtools v1.14 [80] `--maf 0.05 --max-alleles 2 -- min-alleles 2 --max-missing 0.9 –remove-filtered-all --remove-indels` was used to retain only bi-allelic markers with little missing data and remove the rarest variants (referred to as set 2). The next filter utilized the VCF.Filter.v1.0.py script [61] to remove variants with allelic imbalance `-ab 0.2`, followed by VCFtools to select only the 37 pseudomolecules `--chr` (referred to as set 3). The final filter utilized BCFtools v.1.9 to filter variants for LD in a 20kb window ` +prune, -w 20kb, -l 0.4, -n 2` (referred to as set 4). Finally, VCFtools `–relatedness2` was run to detect closely related individuals. In light of the results, individuals ` Oke180104-Fert164` and ` Oke171107-D` are recommended to be used cautiously in further analysis as they were deemed most likely to be haploid progeny (expected) and sibling (unexpected) respectively of other individuals in the analysis (can be applied to all sets prior to further analysis using `vcftools –remove`; removed for Figure 3 below, not removed in Figure 2).

SNP dataset analyses

SNPhylo [81] was run on the "set 4" dataset, using additional options `-m 0.05 -P Oket_chroms_37_ld0.2 -b –B 1000 -a 37` in order to generate a bootstrapped phylogenetic tree of the chum salmon. Visualization was performed using the Figtree V1.4.4 package (http://tree.bio.ed.ac.uk/software/figtree/). PCA analyses were performed on the same dataset using the R package `SNPrelate`, with full and Canadian-only sample sets plotted– the set 3 is visualized in this work, with all visualization performed using the ggplot2 package [82].

The sex phenotypes associated with re-sequencing samples (Supplementary Table 3) were utilized as the basis for a genome-wide association analysis for sex. Utilizing the allele balanced SNP set (“set 2” above), VCFtools v1.14 was used to generate input for plink (chromosomes only). An association test was run in PLINK 1.9 [83] using the formatted output data, and resulting Manhattan plot visualized in R [84] using the qqman package [85] . Further visualization of identified SNPs were performed using the Adegenet package [86]. Counts of coverage utilized samtools v1.9 depth, using default parameters to calculate genome wide coverage over each individual *.bam alignment file, and using the `-b` option to restrict the calculation to only the region of the growth hormone 2 gene (*GH2*) demonstrating elevated coverage in the males following a manual review of the alignments using IGV viewer 2.9.4 [87].

Duplicated regions, presumably from the Salmon specific 4R duplication event, were identified by alignments using the default settings of SyMap v4.2 [72], using a repeat-masked version of the genome following prior methods [61], by masking WindowMasker-based repetitive regions using ` sed -e '/^>/! s/[[:lower:]]/N/g` from the RefSeq genome. Summary tracks were predominately generated using scripts from [40]: Orientation of the blocks were generated using Analyze_Symap_Block_Orientation.py; percent identity was determined using Analyze_Symap_Linear_Alignments.py; percent identify of repetitive regions identified using Percent_Repeat_Genome_Fasta.py. Linkage map markers from “Map 1” in [34] were aligned to the genome as previously described above using BLAST. Linkage disequilibrium (LD) was interpreted using the `--geno-r2` option in VCFtools [80], and outputting only for those comparisons exceeding `--min-r2 0.5` in order to identify the most highly linked SNPs – summaries were further limited to single chromosomes using the `--chr` option. LD calculations utilized the allele balanced set (set 3) described above. LD track utilized counts of markers in linkage disequilibrium across at least 100kb, and summarized as a log sum per 1 million base pairs. Circos v0.69.9 [88] was utilized to visualize the data tracks described.

Heterozygosity analyses followed the same parameters and method as in [61]. Runs of homozygosity were identified from the variants that had been filtered for allele balance using PLINK v1.9 (parameters:—homozyg) [83]. The number of heterozygous genotypes and alternative homozygous genotypes per individual were counted using the same custom script described in the supplementary data of the sockeye genome [61]. Heterozygotes per kbp was calculated as the number of heterozygous genotypes divided by the total nucleotides in the genome (1,853,104,330) multiplied by 1 kbp. The heterozygosity ratio was calculated as the number of heterozygous genotypes divided by the number of alternative homozygous genotypes.

Results and Discussion

Genome Assembly and Annotation

From a raw data set consisting of 59X coverage (110 billion bp) of overlapping 250bp Illumina reads and 60X coverage (114 billion bp) of total mate-pair Illumina reads of three insert sizes (2, 5 and 8kb mean), multiple assembly attempts were performed varying the parameters on read depth as well as read-trimming . Ultimately, three of the attempts resulted in a completed assembly (see Table 1), with the final attempt being the most successful, with a contig N50 of 13.1 kb and a scaffold N50 of 653 kbp. Following AllPaths-LG assembly, contig gaps were filled utilizing PB Suite and 53 billion bp of Pacific Biosciences Sequel long-reads. Following Pilon polishing, utilizing the short insert Illumina libraries, scaffolds were organized into pseudomolecules representing the 37 chromosomes in chum salmon, predominately guided by the publicly available linkage map [34]; ultimately, the linkage map allowed for 70% of the genome assembly to be assigned to a linkage group, slightly lower but approximately equivalent to prior attempts in salmonids using equivalent techniques (e.g., [40, 61]). The final assembly was uploaded to NCBI under BioProject PRJNA556729 and ultimately was included in the RefSeq database as GCF_012931545.1.

Busco scores indicate that most of the genome is represented within the family, with results similar to what has been seen in Sockeye, with 85.0% complete (25.1% duplicated), 3.2% fragmented and 11.8% missing. This likely reflects the slightly more fragmented nature of the genome as compared to prior attempts using the same technology in other species. We believe this is most likely due to some minor shearing observed in the DNA utilized for library preparation. We did attempt to use 10X chromium data as part of this assembly process, but, our scaffolding power was negligible – after review, it is likely that DNA shearing noted in the bioanalyzer trace prior to library construction limited the size of the fragments from which to generate the linked reads, thus limiting scaffolding power. The raw data from this attempt is included under the BioProject (see Supplementary Table 1), but further attempts would need to use a separate individual in order to increase length of the starting material. Given that sequencing and assembly technology has advanced rapidly since we began this project, it is likely further efforts to improve the genome may benefit from the use of long-read technologies, where incredible advances in contiguity have already been demonstrated in salmonids [44, 89]. Indeed, a long-read assembly for chum salmon is planned by the authors, and will eventually replace this reference, in due course.

Following inclusion in the RefSeq database, the genome was annotated utilizing the NCBI Eukaryotic Annotation pipeline, ultimately yielding Annotation Release 100 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oncorhynchus_keta/100/) – see Table 2 for a summary. Gene annotation, via chum-specific reads, primarily utilized the 19 tissue RNA-seq dataset sequenced as part of this work (see Supplementary Table 2), with additional contribution of sequences from two additional datasets with publicly accessible RNA-seq data [90, 91]. Overall, gene numbers are comparable to other salmonid genomes, and thus likely reflect a relatively complete representation of the coding sequence. It is likely that a future genome that utilized long-reads would result in a slightly increased number of genes (as observed between *Oncorhynchus kisutch* Annotation releases 100 and 101, for example).

The resulting assembly is summarized visually in Figure 1. Duplicated regions, as identified via self-alignment using Mummer [92] reflect re-diploidized segments of the genome from the salmon-specific 4R duplication event. There are observations of elevated percent identity on the ends of some chromosomes (Figure 1) that demonstrate partial re-diplodization as in Waples et al. (2016) [34] (e.g., LG05 and LG32), but the effect is not nearly as extensive as that observed in other species. The repetitive elements identified by Window Masker were elevated in regions likely overlapping with centromeres based on synteny with other species for which chromosome arms have been described (Figure 1). Figure 1 also shows Map 1 from Waples et al. (2016) [34] (which can be further visualized in more detail in Supplementary Figure 1 and map 2 in Supplementary Figure 2), and demonstrates the co-linearity of the map with the pseudomolecules. As the maps do contain regions of low-recombination, much of the ordering and orientation of the scaffolds into pseudomolecules (but crucially, not the assignment to the pseudomolecule itself) relies heavily in some positions on the long-read and Hi-C based assembly of coho salmon (GCF_002021735.2). Given the extensive conserved synteny and co-linearity between orthologous salmonid chromosome arms demonstrated elsewhere (e.g., [36, 40]), this would appear to be a reasonable approach, and has been part of the development of pseudomolecules for short-read assemblies in salmonids previously (eg. [40]). Regions of the genome with high LD generally overlap with regions of reduced recombination as observed in the linkage map (Figure 1). Further exploration of regions of high LD can be observed in Supplementary Figures 3 and 4.

As a final clarification on the assembly presented, we note that pseudomolecules have been named within the publicly available assembly based on the linkage group naming mechanism in Waples et al. (2016), [34] to allow for direct comparison between the two works. However, the authors also note, and are enthusiastic about, the naming convention suggested by Sutherland et. al., [36] to describe chromosomal arms, and indeed the adoption of the system into the grayling genome assembly [42]. We provide here in Table 3 the naming for the pseudomolecules that could be suggested by such a system. While the pattern of fusions do make this system less than ideal, and the resulting chromosome names are somewhat unwieldy, we provide them here as a quick reference and potential guide to re-naming of the linkage groups should such a system continue to prove popular as future assemblies are released. Presenting both names here will hopefully ease future reference, whichever naming scheme ends up being formally adopted in future works.

Population level variation

Given the extensive distribution of chum salmon, attempts were made to maximize geographic distribution of the samples selected within the study. We were able to take advantage of an extensive collection of samples [10] in the archive of the Molecular Genetics Lab (Pacific Biological Station, Fisheries and Oceans Canada), combined with more recent contributions from various Fisheries and Oceans Canada hatchery staff for recent brood. While the collection is focused on British Columbia, the addition of the Japanese samples originating from the Tokushibetsu River on the Island of Hokkaido give a glance at the degree of variation expected across the Pacific. Samples and available metadata are summarized in Supplementary Table 3. In total, 15,372,999 nucleotide variants have been described with this data in the raw dataset, with described filters leaving 8,868,081 in set 2, 2,135,295 in set 3, and 94,080 in set 4. A summary of statistics by individual is given in supplementary table 5 [61]. On average (and ignoring the haploid individual), total lengths of runs-of-homozygosity (ROH) averaged 12.4 Mbp [0 - 40.8 Mbp as determined using default parameters]. Heterozygous SNPs per 1kbp averaged 1.47 (Standard Deviation = 0.15), while the heterozygosity ratio averaged 2.23 (Standard Deviation = 0.45). Overall, results are relatively similar to what was observed utilizing a parallel analysis in Sockeye salmon, although the overall length of ROH is lower (12.4 Mbp in chum salmon vs. 35.5 Mbp in Sockeye salmon), whereas heterozygous SNPs per 1kbp are increased (1.47 in chum salmon vs. 0.67 in Sockeye salmon), and the heterozygous ratio was approximately equivalent (2.23 in chum salmon vs. 2.21 in sockeye salmon after removing outliers). Deviations below the mean for both heterozygosity calculations were predominately associated with average coverage, implying that depth of sequencing likely impacted to some extent these calculations. Regardless, we demonstrate in chum salmon that there is a general increase in heterozygosity as compared to sockeye salmon, and establishes a comparative metric to be carried through to future comparative analyses in other Pacific salmonids.

Analyses of the SNP set resulting from whole genome resequencing (targeted coverage of 15X) should be considered exploratory, as collections were focused on geographic coverage to maximize variants within the catalogues rather than addressing additional questions. Nevertheless, the geographic variation explored allowed us to better understand differentiation among British Columbia locations. To this end, a bootstrapped maximum likelihood tree was constructed using a linkage-disequilibrium thinned SNP-set using SNPhylo [81]. As can be seen in Figure 2, the dendrogram clusters samples by regions similar to past analyses with more comprehensive sampling but using older marker technologies (see above). Samples can be resolved into regions corresponding to descriptions from the comprehensive sampling of Beacham et al. (2009) [10], with individual samples resolvable into Japan – Hokkaido; BC Central Coast (Snootli and Kitimat); BC- Haida Gwaii (Deena Creek); BC – West Coast Vancouver

Island (Nitinat); BC – Strait of Georgia (Tenderfoot, Big Qualicum, Puntledge); and BC – Lower Fraser (Chilliwack, Inch, Chehalis). However, within clusters from multiple regions, we see a relative lack of resolution to the riverine level. Such observations are supported by Principal Component Analysis (PCA) as well (Figure 3); however, we do begin to see stronger delineation, possibly from the increased number of variant and dimensions in the PCA analysis. In Figure 3A, we observe differentiation across the Pacific Ocean (best described along PC1), and to a lesser degree geographically across the British Columbia coastline (along PC2). When described regionally, individuals from most populations can easily be resolved when focusing on the British Columbia coastline (Figure 3B), and we are able to see delineation among all collections, except those in the Fraser River Basin. Focusing on the Fraser River Basin sites alone, the pattern is less clustered (Figure 3C), although we do see some separation from salmon collected in different river systems of the Fraser River drainage.

Clustering techniques show that river-level resolution is not always observed. Such results have been noted in the past when considering fishery mixture resolution and describing assignments to region only (for example [9, 93]), but it is worth emphasizing that incomplete resolution among collected populations remains true when considering a relatively comprehensive genome-wide representation of variation. As part of the thinning procedure for SNPhylo, however, by default a significant number of SNPs are removed to increase the speed of the calculation. Alternatively, in the analysis of principal components, with just an LD threshold applied (0.2), a much greater number of SNPs were input into the resulting analysis, and it is likely that the number of SNPs in the end analysis played at least a partial role in the reduced delineation observed in the dendrogram relative to the PCA. While collection level differentiation does emerge in the PCA result, observations on reduced datasets (e.g., by chromosome) greatly inhibited the resolving power of the analysis (supplementary figure 5). Based on the results presented here, is is likely that collection level-specific SNPs could be identified in this dataset that maximize the population differentiation observed genome-wide, and that would further drive differentiation observed in the PCA. However, with such a small sampling size, it is likely that any such discovery would be more a representation of sampling depth, and the noise within a set would be high. However, this dataset is now available, should future researchers need to draw on a pool of potential SNPs from which to develop such assays.

Within BC, chum salmon regional groupings are described at the conservation unit (CU) level [94], and it is intriguing to note that there may be substructure to the results observed along those lines in the present analysis. For example, the Tenderfoot hatchery samples in the Howe Sound-Burrard Inlet CU do tend to cluster more strongly, relative to the other collection sites in the adjacent Georgia Strait CU suggesting that a greater sample size may allow recovery of further groupings. However, it is likely that straying, generally described as high in chum salmon, is playing a role in limiting genetic distinctiveness to the level of the CU (or higher) regional groupings. While sampling within the study focused primarily on large hatchery operations, it is also possible we are simply revealing a high degree of variation within each population due to a large effective population size, in which case sufficient additional sampling may coalesce around a mean per population. Still, even within the dataset here, the observation remains that individual population level resolution within a region may begin to be demonstrated with genome-wide representation.

Mapping the sex-determining region

Although limited metadata was collected for individuals sampled beyond geographic locations sampled, we were able to collect phenotypic sex information on hatchery brood samples. Thus, we were able to explore genome wide associations (GWAs) of phenotypic sex. As demonstrated in Figure 4A, two clear peaks were observed with the GWAS: a very strong peak on Linkage Group 15, and another, albeit somewhat weaker, association on Linkage Group 3. As shown in Figure 4B, the specific region overlaps with an area of increased linkage disequilibrium on the distal end of LG15 . In Figure 4C, the genotypes for each individual is displayed for the 20 SNPs seen as most associated with sex within the GWAS analysis. LG15 has been previously identified by McKinney et al. (2020), [59] as linked to sex during a RAD-seq based study of chum salmon populations within Alaska. In this prior work, linkage of sex to a particular region of the genome was complicated by two potential factors – a lack of a chromosome-level assembly for chum salmon, and the identification of a putative inversion along the chromosome that resulted in significant patterns of linkage. We utilized the sex-linked RAD loci to position the markers onto the new genome assembly and observed that while all were indeed placed along Oket_LG15, they appeared to be more dispersed along the chromosome, and were not strongly linked to sex within our geographically distinct sample set (Supplementary Table 7). Within the present study, we observed sex linked to a very narrow region along Oket_LG15; while some noise is observed, the peak is approximately in the 30.8 Mbp to 31 Mbp region and encompasses four annotated genes: potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2-like; E3 ubiquitin-protein ligase RNF126-like; SURP and G-patch domain-containing protein 1-like; and serine/threonine-protein kinase STK11-like. While we do not suggest any of these are the sex-determination gene – as with other Pacific salmonids it is presumed to be sdY [95] – given that the underlying genome assembly is female, this likely represents the approximate region where sdY is inserted on the Y-chromosome, and limited recombination surrounding the region has led to sex-specific markers extending to autosomal-like sequence flanking the insertion. This region (on chromosome 3.2 based on the naming scheme in Sutherland et al., 2016 [36] and Table 3) would appear to be a unique placement thus far in sdY mapping – however, the relatively common observation of sdY on chromosome arm 3.1 (sockeye salmon, coho salmon, lake whitefish; [96] and references therein) does suggest that inter-homeologue transfer between chromosome arms arising from the most recent salmon-specific duplication could be a mechanism for this transfer.

The strong secondary peak observed on Linkage group 3 is slightly more confounding and intriguing, as it does not appear to be linked to a known sex-determination orthologue in salmonids [96], and because potential sex-markers appear linked to those on LG-15. While it could be linked to a misplaced contig within the assembly, comparative mapping between additional species did not suggest anything was misplaced based on conserved synteny (data not shown; performed within Symap using default parameters) – if this is the case, it is likely that a future long-read based assembly will correct such a matter. It seems most likely in this case that it represents a repetitive element or otherwise duplicated sequence that is prominent in the Y-specific region but is not present in this female genome; thus, mis-mapping appears to occur elsewhere in the genome. A manual review of the region does imply a highly repetitive region, with great differentiation in depths indicative of collapsed repeats. Such mismapping based on collapsed repeats or a lack of sex-specific reference is not uncommon (e.g., as demonstrated in Chinook salmon by mapping of the Y-specific growth hormone pseudogene to the GH2 locus on a different chromosome [97]) and it may be that assembly of a male genome will reveal repeat patterns underlying this unexpected result observed here. There may be additional, more complex reasons based on the observance of multiple sdY regions seen in other species (e.g., Atlantic salmon [98]), although other explanations may be equally likely here. Observations have been made elsewhere that GH-Y, a commonly used proxy for genetic sex in salmonids [99], was found to be missing in males or present in females in some chum salmon populations [100]. While the presented genome is female-based (and thus not predicted to contain GH-Y), observation of relative coverage at the most closely related gene in the genome – GH2 – indeed implied that between 0-5 copies of GH-Y are observed in male individuals, with those males observed to be missing GH-Y being from Kitimat (2x), Snootli (1x) and Tenderfoot (1x): see supplementary table 6. These data do not suggest the phenotypes are mis-identified, however, as inclusion of a Rainbow trout sdY into the alignment phase demonstrated that the presence of sdY matches the phenotype, as would be predicted [95]. No copy-number differences could be interpreted from the sdY alignment unfortunately, as the underlying sequence from trout appeared too differentiated to obtain a reliable estimate of coverage; however, reads were observed aligned to the sequence in all male individuals and not in female individuals in a manual review utilizing IGV viewer. Still, the GH-Y results do indicate that there is variability in the genomic architecture surrounding sdY, and perhaps may indicate that alternate locations within the genome could be influential. Whatever the underlying genomic architecture of the sex-determination region may be in chum salmon, the result presented here underlines the usefulness and ease of use of the presented SNP dataset and reference genome in mapping a trait of interest to the appropriate chromosome and chromosomal region within the genome.

Conclusions

The genome assembly for chum salmon represents a relatively complete representation of the chum salmon genome: the first such resource for the species. Contiguity and completeness is likely most affected in regions with high residual tetraploidy or incomplete re-diploidization. While long-read based assemblies (and future sequencing technologies) are likely to generate a more complete picture, the current genome assembly represents a valuable resource for chum salmon on par with those available for Chinook, sockeye, and longstanding assemblies for Atlantic salmon and rainbow trout that allowed a transformation in genomic understanding of these commercially and culturally specific species (e.g., [101]). Complementing the presented genome is a pilot-scale catalogue of variation that provides a genome-wide resource for British Columbian chum salmon populations, and allows for contrasting variation in Western and Eastern Pacific lineages. Such a dataset will be explored further as a resource for SNP genotyping panel expansion, structural variation discovery, or as demonstrated here, in identification of the chromosome and position most likely to contain the sex-determination gene in chum salmon.

Acknowledgements

We would like to thank the staff at McGill University and Genome Quebec Innovation Centre (now the Centre d’expertise et de services Génome Québec; https://cesgq.com/) in Montreal, QC, Canada, and the NRC Plant Biotechnology Institute Sequencing Centre in Saskatoon, SK, Canada, for their work on library construction and sequencing on this project. Compute Canada (https://www.computecanada.ca/) provided much of the computing power for genome assembly and SNP discovery, primarily on the Cedar cluster. Support for this research from Fisheries and Oceans Canada, the Canadian Regulatory System for Biotechnology.

References

1. Helfield JM, Naiman RJ. Keystone Interactions: Salmon and Bear in Riparian Forests of Alaska. Ecosystems. 2006;9:167–80.

2. Salo EO. Life History of Chum Salmon (Oncorhynchus keta). In: Groot C, Margolis L, editors. Pacific salmon life histories. Vancouver: UBC Press; 1991.

3. Bakkala RG. Synopsis of Biological Data on the Chum Salmon, Oncorhynchus Keta (Walbaum) 1792. U.S. Fish and Wildlife Service; 1970.

4. Fredin RA, Major RL, Bakkala RG, Tanonaka GK. Pacific salmon and the high seas salmon fisheries of Japan. 1977.

5. Behnke R. Trout and salmon of north america. Free Press; 2010.

6. Gislason G, Lam E, Knapp G, Guettabi M. Economic Impacts of Pacific Salmon Fisheries. Pacific Salmon Commission. University of Alaska Anchorage Institute of Social & Economic Research.

7. Keefer ML, Caudill CC. Homing and straying by anadromous salmonids: a review of mechanisms and rates. Reviews in Fish Biology and Fisheries. 2014;24:333–68.

8. Quinn TP, Stewart IJ, Boatright CP. Experimental evidence of homing to site of incubation by mature sockeye salmon, Oncorhynchus nerka. Animal Behaviour. 2006;72:941–9.

9. Beacham T, Sato S, Urawa S, Le K, Wetklo M. Population structure and stock identification of chum salmon *Oncorhynchus keta* from Japan determined by microsatellite DNA variation. Fisheries Science. 2008;74:983–94.

10. Beacham TD, Candy JR, Le KD, Wetklo M. Population structure of chum salmon (Oncorhynchus keta) across the Pacific Rim, determined from microsatellite analysis. Fishery Bulletin. 2009;107:244–60.

11. Olsen JB, Flannery BG, Beacham TD, Bromaghin JF, Crane PA, Lean CF, et al. The influence of hydrographic structure and seasonal run timing on genetic diversity and isolation-by-distance in chum salmon (Oncorhynchus keta). Can J Fish Aquat Sci. 2008;65:2026–42.

12. Small MP, Frye AE, Von Bargen JF, Young SF. Genetic Structure of Chum Salmon (Oncorhynchus keta) Populations in the Lower Columbia River: Are Chum Salmon in Cascade Tributaries Remnant Populations? Conservation Genetics. 2006;7:65–78.

13. Small MP, Rogers Olive SD, Seeb LW, Seeb JE, Pascal CE, Warheit KI, et al. Chum Salmon Genetic Diversity in the Northeastern Pacific Ocean Assessed with Single Nucleotide Polymorphisms (SNPs): Applications to Fishery Management. North American Journal of Fisheries Management. 2015;35:974–87.

14. Crow KD. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? Molecular Biology and Evolution. 2006;23:887–92.

15. Ohno S. Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg; 1970.

16. Otto SP, Whitton J. Polyploid Incidence and Evolution. Annu Rev Genet. 2000;34:401–37.

17. Taylor JS, Raes J. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu Rev Genet. 2004;38:615–43.

18. Sankoff D, Zheng C. Whole Genome Duplication in Plants: Implications for Evolutionary Analysis. In: Setubal JC, Stoye J, Stadler PF, editors. Comparative Genomics. New York, NY: Springer New York; 2018. p. 291–315.

19. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997;387:708–13.

20. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 2004;428:617–24.

21. Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, et al. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity. 2016;116:190–9.

22. Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol. 2017;15:62.

23. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008;453:1064–71.

24. Dehal P, Boore JL. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol. 2005;3:e314.

25. Taylor JS, Van de Peer Y, Braasch I, Meyer A. Comparative genomics provides evidence for an ancient genome duplication event in fish. Phil Trans R Soc Lond B. 2001;356:1661–79.

26. Taylor JS. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. Genome Research. 2003;13:382–90.

27. Hoegg S, Brinkmann H, Taylor JS, Meyer A. Phylogenetic Timing of the Fish-Specific Genome Duplication Correlates with the Diversification of Teleost Fish. J Mol Evol. 2004;59:190–203.

28. Allendorf FW, Thorgaard GH. Tetraploidy and the Evolution of Salmonid Fishes. In: Turner BJ, editor. Evolutionary Genetics of Fishes. Boston, MA: Springer US; 1984. p. 1–53.

29. 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. 2014;281:20132881.

30. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nature Reviews Genetics. 2017;18:411–24.

31. Jones SR, Fast MD, Johnson SC, Groman DB. Differential rejection of salmon lice by pink and chum salmon: disease consequences and expression of proinflammatory genes. Dis Aquat Organ. 2007;75:229–38.

32. Sutherland BJ, Koczka KW, Yasuike M, Jantzen SG, Yazawa R, Koop BF, et al. Comparative transcriptomics of Atlantic Salmo salar, chum Oncorhynchus keta and pink salmon O. gorbuscha during infections with salmon lice Lepeophtheirus salmonis. BMC Genomics. 2014;15:200.

33. 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. Journal of Heredity. 2015;106:217–27.

34. Waples RK, Seeb LW, Seeb JE. Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). Mol Ecol Resour. 2016;16:17–28.

35. Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, et al. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biology. 2017;18:111.

36. Sutherland BJG, Gosselin T, Normandeau E, Lamothe M, Isabel N, Audet C, et al. Salmonid Chromosome Evolution as Revealed by a Novel Method for Comparing RADseq Linkage Maps. Genome Biology and Evolution. 2016;8:3600–17.

37. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.

38. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.

39. Pearse DE, Barson NJ, Nome T, Gao G, Campbell MA, Abadía-Cardoso A, et al. Sex-dependent dominance maintains migration supergene in rainbow trout. Nat Ecol Evol. 2019;3:1731–42.

40. Christensen KA, Leong JS, Sakhrani D, Biagi CA, Minkley DR, Withler RE, et al. Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. PLoS ONE. 2018;13:e0195461.

41. Narum SR, Di Genova A, Micheletti SJ, Maass A. Genomic variation underlying complex life-history traits revealed by genome sequencing in Chinook salmon. Proc R Soc B. 2018;285:20180935.

42. Sävilammi T, Primmer CR, Varadharajan S, Guyomard R, Guiguen Y, Sandve SR, et al. The Chromosome-Level Genome Assembly of European Grayling Reveals Aspects of a Unique Genome Evolution Process Within Salmonids. G3. 2019;9:1283–94.

43. Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, et al. The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. Genome Biology and Evolution. 2018;10:2785–800.

44. 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. preprint. Genomics; 2019.

45. Beacham TD, Wallace CG, Jonsen K, Sutherland BJG, Gummer C, Rondeau EB. Estimation of Conservation Unit and population contribution to Chinook salmon mixed-stock fisheries in British Columbia, Canada using direct DNA sequencing for single nucleotide polymorphisms. Can J Fish Aquat Sci. 2021. https://doi.org/10.1139/cjfas-2020-0462.

46. Beacham TD, Wallace C, Jonsen K, McIntosh B, Candy JR, Rondeau EB, et al. Accurate estimation of conservation unit contribution to coho salmon mixed-stock fisheries in British Columbia, Canada, using direct DNA sequencing for single nucleotide polymorphisms. Can J Fish Aquat Sci. 2020;77:1302–15.

47. Phelps SR, LeClair LL, Young S, Blankenship HL. Genetic Diversity Patterns of Chum Salmon in the Pacific Northwest. Can J Fish Aquat Sci. 1994;51:65–83.

48. Seeb LW, Crane PA. High Genetic Heterogeneity in Chum Salmon in Western Alaska, the Contact Zone between Northern and Southern Lineages. Transactions of the American Fisheries Society. 1999;128:58–87.

49. Beacham TD, Spilsted B, Le KD, Wetklo M. Population structure and stock identification of chum salmon (Oncorhynchus keta) from British Columbia determined with microsatellite DNA variation. Can J Zool. 2008;86:1002–14.

50. Smith CT, Seeb LW. Number of Alleles as a Predictor of the Relative Assignment Accuracy of Short Tandem Repeat (STR) and Single-Nucleotide-Polymorphism (SNP) Baselines for Chum Salmon. Transactions of the American Fisheries Society. 2008;137:751–62.

51. Smith CT, Baker J, Park L, Seeb LW, Elfstrom C, Abe S, et al. Characterization of 13 single nucleotide polymorphism markers for chum salmon: PRIMER NOTE. Molecular Ecology Notes. 2005;5:259–62.

52. Smith CT, Elfstrom CM, Seeb LW, Seeb JE. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon: SNPs IN PACIFIC SALMON. Molecular Ecology. 2005;14:4193–203.

53. Elfstrom CM, Smith CT, Seeb LW. Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. Mol Ecol Notes. 2007;7:1211–5.

54. Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, Harkins T, et al. Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids: PERMANENT GENETIC RESOURCES ARTICLE. Molecular Ecology Resources. 2011;11:335–48.

55. Petrou EL, Hauser L, Waples RS, Seeb JE, Templin WD, Gomez-Uchida D, et al. Secondary contact and changes in coastal habitat availability influence the nonequilibrium population structure of a salmonid (*Oncorhynchus keta*). Mol Ecol. 2013;22:5848–60.

56. Small M, Warheit K, Pascal C, Seeb L, Ruff C, Zischke J, et al. Chum Salmon Southern Area Genetic Baseline Enhancement Part 1 and Part 2: Amplicon Development, Expanded Baseline Collections, and Genotyping.

57. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE. 2008;3:e3376.

58. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research. 2007;17:240–8.

59. McKinney G, McPhee MV, Pascal C, Seeb JE, Seeb LW. Network Analysis of Linkage Disequilibrium Reveals Genome Architecture in Chum Salmon. G3: Genes|Genomes|Genetics. 2020;10:1553.

60. Gao G, Nome T, Pearse DE, Moen T, Naish KA, Thorgaard GH, et al. A New Single Nucleotide Polymorphism Database for Rainbow Trout Generated Through Whole Genome Resequencing. Front Genet. 2018;9:147.

61. Christensen KA, Rondeau EB, Minkley DR, Sakhrani D, Biagi CA, Flores A-M, et al. The sockeye salmon genome, transcriptome, and analyses identifying population defining regions of the genome. PLOS ONE. 2020;15:e0240935.

62. Quillet E, Garcia P, Guyomard R. Analysis of the production of all homozygous lines of rainbow trout by gynogenesis. J Exp Zool. 1991;257:367–74.

63. Genomic DNA Preparation from RNAlaterTM Preserved Tissues—CA [Internet]. https://www.thermofisher.com/ca/en/home/references/protocols/nucleic-acid-purification-and-analysis/rna-protocol/genomic-dna-preparation-from-rnalater-preserved-tissues.html. Accessed 18 Feb 2021.

64. Christensen KA, Rondeau EB, Minkley DR, Leong JS, Nugent CM, Danzmann RG, et al. The Arctic charr (Salvelinus alpinus) genome and transcriptome assembly. PLoS ONE. 2018;13:e0204076.

65. Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, et al. 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:e102089.

66. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

67. Marić J. Long Read RNA-seq Mapper. Master Thesis. University of Zagreb; 2015.

68. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences. 2011;108:1513–8.

69. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE. 2012;7:e47768.

70. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014;9:e112963.

71. Johnson HA, Rondeau EB, Minkley DR, Leong JS, Whitehead J, Despins CA, et al. Population genomics of North American northern pike: variation and sex-specific signals from a chromosome-level, long read genome assembly. bioRxiv. 2020;:2020.06.18.157701.

72. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Research. 2011;39:e68–e68.

73. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 227–45.

74. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018;:201178.

75. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

76. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;43:11.10.1-11.10.33.

77. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

78. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

79. Picard toolkit. Broad Institute; 2019.

80. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

81. Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15:162.

82. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.

83. 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.

84. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

85. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software. 2018;3:731.

86. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24:1403–5.

87. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;29:24–6.

88. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Research. 2009;19:1639–45.

89. Gao G, Magadan S, Waldbieser GC, Youngblood RC, Wheeler PA, Scheffler BE, et al. A long reads-based de-novo assembly of the genome of the Arlee homozygous line reveals chromosomal rearrangements in rainbow trout. G3 Genes|Genomes|Genetics. 2021. https://doi.org/10.1093/g3journal/jkab052.

90. Palstra AP, Fukaya K, Chiba H, Dirks RP, Planas JV, Ueda H. The Olfactory Transcriptome and Progression of Sexual Maturation in Homing Chum Salmon Oncorhynchus keta. PLOS ONE. 2015;10:e0137404.

91. Tatara Y, Kakizaki I, Kuroda Y, Suto S, Ishioka H, Endo M. Epiphycan from salmon nasal cartilage is a novel type of large leucine-rich proteoglycan. Glycobiology. 2013;23:993–1003.

92. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biology. 2004;5:R12.

93. SEEB LW, TEMPLIN WD, SATO S, ABE S, WARHEIT K, PARK JY, et al. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Molecular Ecology Resources. 2011;11:195–217.

94. Fishery & Assessment Data Section, Pacific Biological Station. Chum Salmon (Oncorhynchus keta) Conservation Units, Sites & Status. 2017.

95. Yano A, Nicol B, Jouanno E, Quillet E, Fostier A, Guyomard R, et al. The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. Evol Appl. 2013;6:486–96.

96. Sutherland BJG, Rico C, Audet C, Bernatchez L. Sex Chromosome Evolution, Heterochiasmy, and Physiological QTL in the Salmonid Brook Charr Salvelinus fontinalis. G3 (Bethesda). 2017;7:2749–62.

97. Micheletti SJ, Narum SR. Utility of pooled sequencing for association mapping in nonmodel organisms. Molecular Ecology Resources. 2018;18:825–37.

98. Eisbrenner WD, Botwright N, Cook M, Davidson EA, Dominik S, Elliott NG, et al. Evidence for multiple sex-determining loci in Tasmanian Atlantic salmon (Salmo salar). Heredity. 2014;113:86–92.

99. Devlin RH, Biagi CA, Smailus DE. Genetic mapping of Y-chromosomal DNA markers in Pacific salmon. Genetica. 2001;111:43–58.

100. Muttray AF, Sakhrani D, Smith JL, Nakayama I, Davidson WS, Park L, et al. Deletion and Copy Number Variation of Y-Chromosomal Regions in Coho Salmon, Chum Salmon, and Pink Salmon Populations. Transactions of the American Fisheries Society. 2017;146:240–51.

101. Bobe J, Marandel L, Panserat S, Boudinot P, Berthelot C, Quillet E, et al. 2 - The rainbow trout genome, an important landmark for aquaculture and genome evolution. In: MacKenzie S, Jentoft S, editors. Genomics in Aquaculture. San Diego: Academic Press; 2016. p. 21–43.

Tables

Table 1: Assembly results for Allpaths and PBSuite based assemblies performed.

Table 2: Summary of Annotation Release 100 from the NCBI Eukaryotic Annotation pipeline. See https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oncorhynchus_keta/100/ for more details.

Table 3: Pike-like chromosome naming for the chum salmon pseudomolecules described in this work, based on Sutherland et al. (2016) [36]

Figures

Figure 1: Circos plot of the chum salmon genome GCF_012931545.1. Inner ribbons demonstrate ohnologous regions (regions duplicated at the salmon-specific genome duplication event). Working in to out, Track A describes the average percent identity between the duplicated regions, in 1 Mbp bins. Track B describes the average percent identity in the chromosomes, in 1 Mbp bins. Track C describes the relationship to “Map 1” chum linkage map from Waples et al. (2016) [34]. Track D describes SNPs demonstrating elevated LD (R-squared >= 0.5) and >= 100 kb apart, demonstrated as a log10 based count, in 1Mbp bins.

Figure 2: Dendrogram produced by SNPhylo, utilizing set 4 SNP data described in the text. Values at nodes indicate bootstrapping. Samples are coloured by geographic region.

Figure 3: Principal component analyses performed on set 3 SNPs described in the text, using

SNPrelate and plotted in ggplot2. Samples are coloured by collection and displayed in the legend. A) the full dataset (all samples) are presented. B), Japanese samples are removed from the analysis. C), the collections are reduced solely to the collections within the Fraser River drainage.

Figure 4: Association of the phenotypic sex to the genome utilizing SNP variant set 1. A) the results of the GWAS are presented, with Bonferroni-adjusted p-values shown at the 5% level (blue line) and 1% (orange line) levels. B) The SNPs with R-squared greater than 0.5 are counted, and plotted to show relationship of distance between SNPs being measured, for the region flanking the signal on Oket_LG15. C) The genotypes for each individual is displayed for the 20 SNPs seen as most associated within the GWAS analysis, with homozygous reference in blue, heterozygous in purple, homozygous alternate in red, and missing genotypes in white. Samples are sorted to group males, females and unknowns (Japanese samples—most likely females).

Supplementary Data

Supplementary Table 1: Biosample and SRA data for individual chum used in generating the genome assembly.

Supplementary Table 2: Biosample and SRA data for individual chum used in generating the Illumina RNA-seq data.

Supplementary Table 3: Biosample and SRA data for individual chum used in generating the Re-sequencing data.

Supplementary Table 4: Allpaths-LG parameters explored in attempting to obtain the highest contiguity assemblies.

Supplementary Table 5: Heterozygosity metrics by individual are described. Includes counts of missing genotypes, Homozygous Reference and Alternate genotypes, Heterozygous genotypes, average depth at called sites, the mean count of heterozygous SNPs per kbp, the ratio of Heterozygous genotypes to Homozygous alternate, and the total length of runs-of homozygosity as determined from PLINK using default parameters.

Supplementary Table 6: Depth of coverage across the alignments, and at the GH2 locus to approximate the count of GH-Y copies in each individual. GH2 is used for this calculation due to the lack of GH-Y in the reference genome, and therefore the alignment of GH-Y to the closest homologue.

Supplementary Table 7: Placement of SNPs associated with phenotypic sex from McKinney et al. [59] in Alaskan chum populations onto the current reference genome.

Supplementary Figure 1: Plotting the association between Linkage groups in Waples et al. (2016), [34] map 1, and the reference genome assembly presented in this work.

Supplementary Figure 2: Plotting the association between Linkage groups in Waples et al. (2016), [34] map 2, and the reference genome assembly presented in this work.

Supplementary Figure 3: Plotting the linkage disequilibrium along each chromosome. SNPs are only displayed if R-squared is greater than 0.5, and is plotted as a count of SNPs.

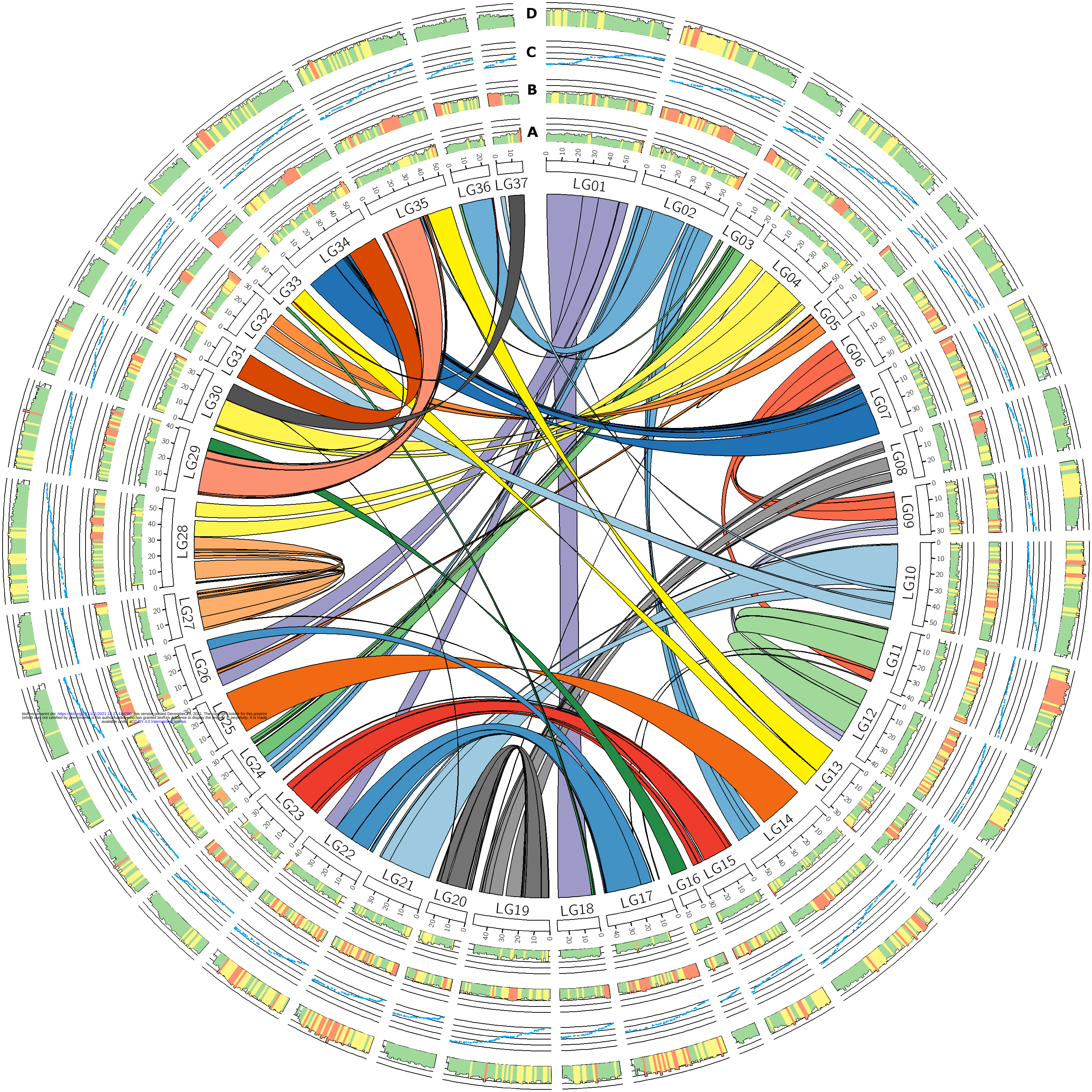
Supplementary Figure 4: Plotting the linkage disequilibrium along each chromosome. SNPs are only displayed if R-squared is greater than 0.5, and each SNP is plotted by R-squared value.

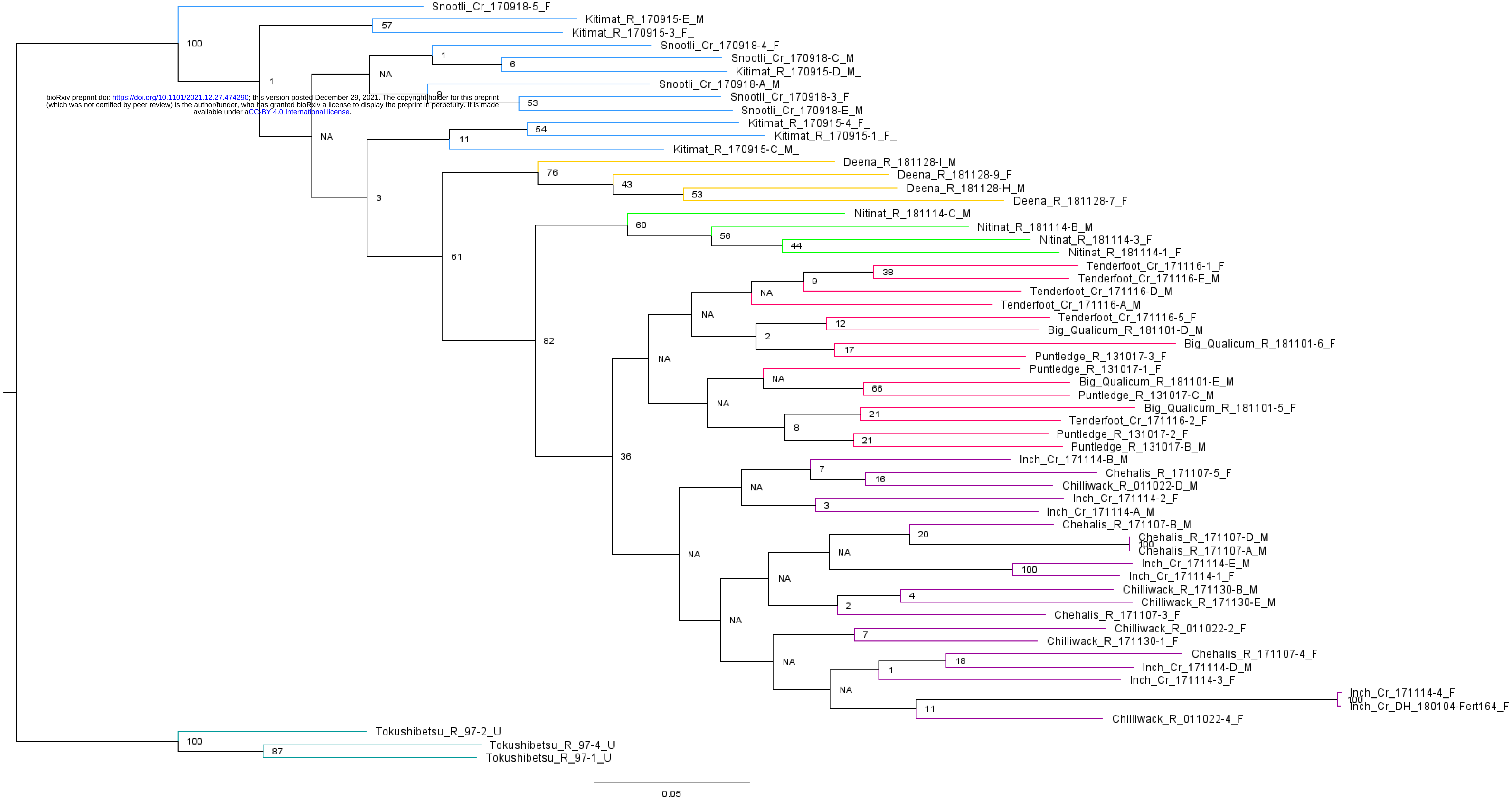
Supplementary Figure 5: Principal component analyses performed on set 3 SNPs described in the text, using SNPrelate and plotted in ggplot2 and reduced to only query LG15. Samples are coloured by collection, displayed in the legend. In panel A, the full dataset (all samples) are presented. In panel B, Japanese samples are removed from the analysis. In panel C, the samples are coloured by collection site rather than by region.

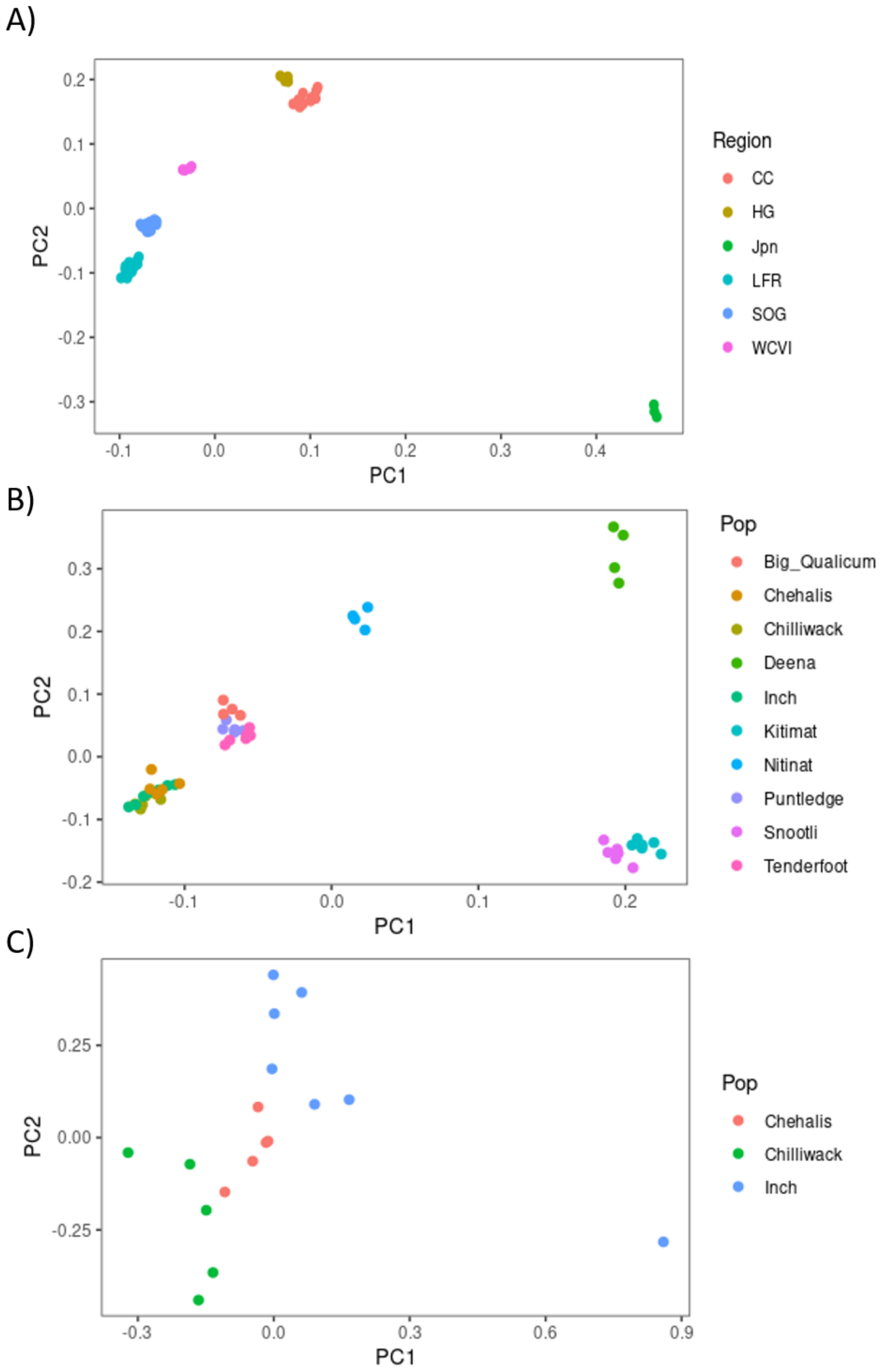
	Assembly size (Contigs)	Assembly size (Scaffolds)	Scaffold N50	Contig N50
Allpaths	1471097779	1813373414	653	13.1
PBSuite	1,766,907,823	1,852,809,593	665,581	52,191

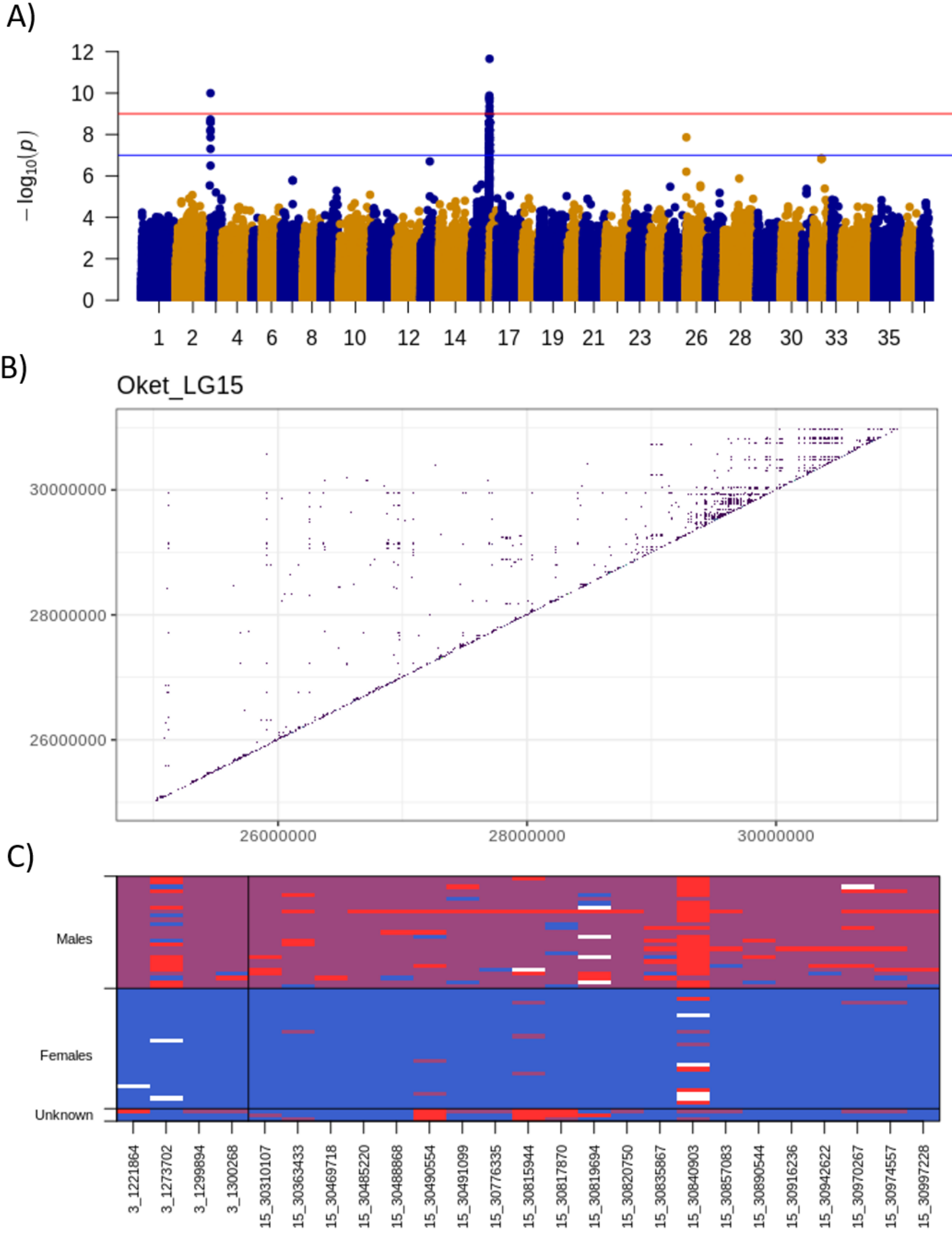
Feature	**Annotation Release 100**
Genes and pseudogenes	45643
protein-coding	36325
non-coding	6205
transcribed pseudogenes	222
non-transcribed pseudogenes	2821
genes with variants	14102
immunoglobulin/T-cell receptor gene segments	70

Linkage Group	Accession	Alternate Naming
Oket_LG01	NC_050106	1.2-8.2
Oket_LG02	NC_050107	6.1-2.2
Oket_LG03	NC_050108	5.1
Oket_LG04	NC_050109	21.1-4.2
Oket_LG05	NC_050110	11.2
Oket_LG06	NC_050111	14.2
Oket_LG07	NC_050112	16.1
Oket_LG08	NC_050113	18.2
Oket_LG09	NC_050114	14.1
Oket_LG10	NC_050115	17.1-9.2
Oket_LG11	NC_050116	7.2
Oket_LG12	NC_050117	7.1
Oket_LG13	NC_050118	24.1-23.1
Oket_LG14	NC_050119	13.1-2.1
Oket_LG15	NC_050120	3.2
Oket_LG16	NC_050121	25.1
Oket_LG17	NC_050122	19.1
Oket_LG18	NC_050123	1.1
Oket_LG19	NC_050124	10.2-18.1
Oket_LG20	NC_050125	10.1
Oket_LG21	NC_050126	17.2
Oket_LG22	NC_050127	19.2
Oket_LG23	NC_050128	3.1
Oket_LG24	NC_050129	5.2
Oket_LG25	NC_050130	13.2
Oket_LG26	NC_050131	8.1
Oket_LG27	NC_050132	12.1
Oket_LG28	NC_050133	12.2-21.2
Oket_LG29	NC_050134	15.2-25.2
Oket_LG30	NC_050135	4.1-22.2
Oket_LG31	NC_050136	20.1
Oket_LG32	NC_050137	9.1-11.1
Oket_LG33	NC_050138	23.2
Oket_LG34	NC_050139	16.2-20.2
Oket_LG35	NC_050140	15.1-24.2
Oket_LG36	NC_050141	6.2
Oket_LG37	NC_050142	22.1


D
C
B
A
LG01
LG02
LG03
LG04
LG05
LG06
LG07
LG08
LG09
LG10
LG11
LG12
LG13
LG14
LG15
LG16
LG17
LG18
LG19
LG20
LG21
LG22
LG23
LG24
LG25
LG26
LG27
LG28
LG29
LG30
LG31
LG32
LG33
LG34
LG35
LG36
LG37


Snootli_Cr_170918-5_F
Kitimat_R_170915-E_M
57
Kitimat_R_170915-3_F_
100
Snootli_Cr_170918-4_F
1
Snootli_Cr_170918-C_M
6
Kitimat_R_170915-D_M_
NA
1
Snootli_Cr_170918-A_M
9
Snootli_Cr_170918-3_F
53
Snootli_Cr_170918-E_M
Kitimat_R_170915-4_F_
54
Kitimat_R_170915-1_F_
NA
11
Kitimat_R_170915-C_M_
Deena_R_181128-I_M
76
Deena_R_181128-9_F
43
Deena_R_181128-H_M
53
3
Deena_R_181128-7_F
Nitinat_R_181114-C_M
60
Nitinat_R_181114-B_M
56
Nitinat_R_181114-3_F
44
Nitinat_R_181114-1_F
61
Tenderfoot_Cr_171116-1_F
38
Tenderfoot_Cr_171116-E_M
9
Tenderfoot_Cr_171116-D_M
NA
Tenderfoot_Cr_171116-A_M
NA
Tenderfoot_Cr_171116-5_F
12
Big_Qualicum_R_181101-D_M
2
82
Big_Qualicum_R_181101-6_F
17
Puntledge_R_131017-3_F
NA
Puntledge_R_131017-1_F
NA
Big_Qualicum_R_181101-E_M
66
Puntledge_R_131017-C_M
NA
Big_Qualicum_R_181101-5_F
21
Tenderfoot_Cr_171116-2_F
8
Puntledge_R_131017-2_F
21
Puntledge_R_131017-B_M
36
Inch_Cr_171114-B_M
7
Chehalis_R_171107-5_F
16
Chilliwack_R_011022-D_M
NA
Inch_Cr_171114-2_F
3
Inch_Cr_171114-A_M
Chehalis_R_171107-B_M
20
Chehalis_R_171107-D_M
100
Chehalis_R_171107-A_M
NA
NA
Inch_Cr_171114-E_M
100
Inch_Cr_171114-1_F
NA
Chilliwack_R_171130-B_M
4
Chilliwack_R_171130-E_M
2
Chehalis_R_171107-3_F
NA
Chilliwack_R_011022-2_F
7
Chilliwack_R_171130-1_F
Chehalis_R_171107-4_F
18
NA
Inch_Cr_171114-D_M
1
Inch_Cr_171114-3_F
NA
Inch_Cr_171114-4_F
100
Inch_Cr_DH_180104-Fert164_F
11
Chilliwack_R_011022-4_F
Tokushibetsu_R_97-2_U
100
Tokushibetsu_R_97-4_U
87
Tokushibetsu_R_97-1_U
0.05


A)
Region
CC
HG
Jpn
LFR
SOG
WCVI
PC2
0.2
0.1
0.0
-0.1
-0.2
-0.3
PC1
-0.1
0.0
0.1
0.2
0.3
0.4
B)
Pop
Big_Qualicum
Chehalis
Chilliwack
Deena
Inch
Kitimat
Nitinat
Puntledge
Snootli
Tenderfoot
PC2
0.3
0.2
0.1
0.0
-0.1
-0.2
PC1
-0.1
0.0
0.1
0.2
C)
Pop
Chehalis
Chilliwack
Inch
PC2
0.25
0.00
-0.25
PC1
-0.3
0.0
0.3
0.6
0.9


A)
$-\log_{10}(p)$
0
2
4
6
8
10
12
1
2
4
6
8
10
12
14
17
19
21
23
26
28
30
33
35
B)
Oket_LG15
26000000
28000000
30000000
C)
Males
Females
Unknown
3_1221864
3_1273702
3_1299894
3_1300268
15_30310107
15_30363433
15_30469718
15_30485220
15_30488868
15_30490554
15_30491099
15_30776335
15_30815944
15_30817870
15_30819694
15_30820750
15_30835867
15_30840903
15_30857083
15_30890544
15_30916236
15_30942622
15_30970267
15_30974557
15_30997228