A chromosome-anchored genome assembly for Lake Trout (Salvelinus namaycush)

Seth Smith[1], Eric Normandeau[2], Haig Djambazian[3], Pubudu Nawarathna[3], Pierre Berube[3], Andrew Muir[4], Jiannis Ragoussis[3], Chantelle Penney[5], Kim Scribner[1], Gordon Luikart[6], Chris Wilson[7], and Louis Bernatchez[2]

[1]Michigan State University
[2]Université Laval
[3]McGill University
[4]Great Lakes Fishery Commission
[5]Trent University
[6]University of Montana
[7]Ontario Ministry of Natural Resources

April 8, 2021

Abstract

Here we present an annotated, chromosome-anchored, genome assembly for Lake Trout (Salvelinus namaycush) – a highly diverse salmonid species of notable conservation concern and an excellent model for research on adaptation and speciation. We leveraged Pacific Biosciences long-read sequencing, paired-end Illumina sequencing, proximity ligation (Hi-C), and a previously published linkage map to produce a highly contiguous assembly composed of 7,378 contigs (contig N50 = 1.8 mb) assigned to 4,120 scaffolds (scaffold N50 = 44.975 mb). 84.7% of the genome was assigned to 42 chromosome-sized scaffolds and 93.2% of Benchmarking Universal Single Copy Orthologs were recovered, putting this assembly on par with the best currently available salmonid genomes. Estimates of genome size based on k-mer frequency analysis were highly similar to the total size of the finished genome, suggesting that the entirety of the genome was recovered. A mitome assembly was also produced. Self-vs-self synteny analysis allowed us to identify homeologs resulting from the Salmonid specific autotetraploid event (Ss4R) and alignment with three other salmonid species allowed us to identify homologous chromosomes in other species. We also generated multiple resources useful for future genomic research on Lake Trout including a repeat library and a sex averaged recombination map. A novel RNA sequencing dataset was also used to produce a publicly available set of gene annotations using the National Center for Biotechnology Information Eukaryotic Genome Annotation Pipeline. Potential applications of these resources to population genetics and the conservation of native populations are discussed.

TITLE

A chromosome-anchored genome assembly for Lake Trout (*Salvelinus namaycush*)

RUNNING TITLE

The Lake Trout genome

TARGET JOURNAL

Molecular Ecology Resources

MANUSCRIPT TYPE

Permanent Genetic Resource

AUTHORS

Seth R. Smith[1,2], Eric Normandeau[3], Haig Djambazian[4], Pubudu M. Nawarathna[5], Pierre Berube[4], Andrew M. Muir[6], Jiannis Ragoussis[4], Chantelle M. Penney[7], Kim T. Scribner[1,2,8], Gordon Luikart[9,10], Chris C. Wilson[11], and Louis Bernatchez[3]

AFFILIATIONS

1 Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

2 Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA

3 Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Canada

4 McGill Genome Centre, Department of Human Genetics, Montreal, Quebec, Canada

5 Canadian Centre for Computational Genomics (C3G), Department of Human Genetics, McGill University, Montréal, QC, Canada

6 Great Lakes Fishery Commission, Ann Arbor, Michigan, USA

7 Environmental and Life Sciences Graduate Program, Trent University, Peterborough, Ontario, Canada

8 Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan, USA

9 Fish and Wildlife Genomics Group; University of Montana; Missoula, Montana, USA

10 Flathead Lake Biological Station; Division of Biological Sciences; University of Montana; Polson, Montana, USA

11 Aquatic Research and Monitoring Section, Ontario Ministry of Natural Resources and Forestry, Peterborough, Ontario, Canada

CORRESPONDING AUTHOR

Seth Smith (smit2959@msu.edu)

ABSTRACT

Here we present an annotated, chromosome-anchored, genome assembly for Lake Trout (*Salvelinus namaycush*) – a highly diverse salmonid species of notable conservation concern and an excellent model for research on adaptation and speciation. We leveraged Pacific Biosciences long-read sequencing, paired-end Illumina

sequencing, proximity ligation (Hi-C), and a previously published linkage map to produce a highly contiguous assembly composed of 7,378 contigs (contig N50 = 1.8 mb) assigned to 4,120 scaffolds (scaffold N50 = 44.975 mb). 84.7% of the genome was assigned to 42 chromosome-sized scaffolds and 93.2% of Benchmarking Universal Single Copy Orthologs were recovered, putting this assembly on par with the best currently available salmonid genomes. Estimates of genome size based on k-mer frequency analysis were highly similar to the total size of the finished genome, suggesting that the entirety of the genome was recovered. A mitome assembly was also produced. Self-vs-self synteny analysis allowed us to identify homeologs resulting from the Salmonid specific autotetraploid event (Ss4R) and alignment with three other salmonid species allowed us to identify homologous chromosomes in other species. We also generated multiple resources useful for future genomic research on Lake Trout including a repeat library and a sex averaged recombination map. A novel RNA sequencing dataset was also used to produce a publicly available set of gene annotations using the National Center for Biotechnology Information Eukaryotic Genome Annotation Pipeline. Potential applications of these resources to population genetics and the conservation of native populations are discussed.

KEYWORDS

Lake Trout, Salvelinus, Salmonid, Genome Assembly, Genomics

I. INTRODUCTION

Many key questions in evolutionary and conservation biology can only be addressed using genomic approaches and appropriate study species. Lake Trout (*Salvelinus namaycush*) are a top predator in many lentic ecosystems across northern North America and express exceptional levels of ecotypic variation (Muir et al. 2014; Muir et al. 2016), making them an ideal study species for exploring the processes of ecological speciation and adaptive diversification. The post-Pleistocene parallel evolution of diverse Lake Trout ecotypes has been likened to the adaptive radiation of cichlid species in the Great Lakes of east Africa (Muir et al. 2016); however, the radiation of Lake Trout ecotypes appears to have occurred over a relatively short evolutionary timescale (Harris et al. 2015, ˜8000 years). At least three distinct Lake Trout ecotypes (lean, siscowet, and humper) once existed throughout the Laurentian Great Lakes (Hansen 1999) and anecdotal evidence suggests that as many as 10 easily differentiable forms once existed in Lake Superior (Goodier 1981). High levels of ecotypic variation have also been documented in contemporary populations across the species range (Blackie et al. 2003; Zimmerman et al. 2006; Hansen et al. 2012; Chavarie et al. 2015), with as many as five trophic ecotypes being found in a single lake (Marin et al. 2016).

Lake Trout are also ancestrally autotetraploid, with the common ancestor of all salmonids having undergone a whole genome duplication event (WGD) roughly 60-100 million years ago (Crête-Lafrenière et al. 2012; Macqueen and Johnston 2014). For this reason, Salmonids have long been considered ideal study species for understanding the evolutionary consequences of WGD (Ohno 1970; Allendorf and Thorgaard 1984). Given the high levels of ecotypic diversity observed in Lake Trout, and the potential for WGD to facilitate the evolution of novel phenotypes (Ohno 1970; Macqueen and Johnston 2014; Van De Peer et al. 2017) and reproductive isolation (Lynch and Force 2000), research exploring the genetic basis for ecotypic differentiation and incipient speciation in Lake Trout could provide important insights about the role of relatively recent WGD events in adaptive radiations.

Furthermore, many Lake Trout populations, particularly those in the Laurentian Great Lakes, have been severely reduced in abundance or distribution, or extirpated, due to invasive species introductions and overfishing (Smith 1968). Following the basin-wide collapse of the lake whitefish (*Coregonus clupeaformis*) commercial fishery in the Great Lakes during the early 20th century, fishing pressure was transferred to Lake Trout populations, which partially contributed to population declines starting in the 1930s (Hansen 1999). A novel predator, the sea lamprey (*Petromyzon marinus*), also invaded the Great Lakes during this time,

leading to further increases in adult Lake Trout mortality and functional extirpation from all lakes except Lake Superior and a small, isolated, population in Lake Huron (Hansen 1999). The restoration program that commenced largely focused on reducing sea lamprey predation, reducing fishing pressure, creating aquatic refuges, and stocking juvenile Lake Trout from a diverse collection of domesticated strains originating from multiple source populations (Krueger et al. 1983; Hansen 1999). Lake Trout populations in Lake Superior rebounded relatively quickly; however, the re-emergence of natural reproduction in other lakes was hindered by high levels of lamprey predation on adult Lake Trout (Pycha et al. 1980), predation on juveniles by invasive alewife (Madenjian et al. 2008), reduced juvenile survival caused by thiamine deficiency (Fitzsimmons et al. 2010), and potentially reduced hatching success associated with PCB contamination (Mac and Edsall 1991). Today, Lake Superior populations remain relatively stable and recruitment has been observed in lakes Huron (Riley et al. 2007), Michigan (Hanson et al. 2013), and Ontario (Lantry 2015). Recent research suggests that domesticated strains used for reintroduction have variable fitness in contemporary Great Lakes environments (Scribner et al. 2018; Larson et al. 2021), and may be differentially contributing to recent recruitment, however, the biological mechanisms that underly these differences in fitness and recruitment remain unclear.

Genomic and transcriptomic approaches have been widely used to identify loci associated with adaptive diversity and ecotypic divergence in salmonids (Prince et al. 2017; Veale and Russelo 2017; Willoughby et al. 2018; Rougeux et al. 2019). This work has been partially driven by the publication of high-quality genome assemblies and linkage maps for numerous salmonid species (Gagnaire et al. 2013; Lien et al. 2016; Christensen et al. 2018a, Christensen et al. 2018b; Pearse et al. 2019; De-Kayne et al. 2020); however, genomic resources are notably lacking for Lake Trout. An annotated, chromosome-anchored, genome assembly is arguably the most valuable resource for advancing genomic research on any species. A publicly available reference genome for Lake Trout would eliminate many challenges associated with conducting conservation-oriented genetic research aimed at restoring ecotypic diversity and viable wild populations. Until recently, the assembly of non-model eukaryotic genomes was prohibitively expensive, computationally challenging, and required the collaborative efforts of large genome consortia; however, the development of long-read ('third generation') sequencing technologies has to some extent eliminated these hurdles (Hotaling and Kelley 2020; Whibley et al. 2021).

Long-read sequencing data can be useful for scaffolding and filling gaps in existing, fragmented, short-read assemblies (English et al. 2012). A number of assembly algorithms also seek to assemble contigs directly from long-read sequencing data (Falcon, Chin et al. 2016; Canu, Koren et al. 2017; wtdbg2, Ruan and Li 2020) and recent work suggests that this approach can be highly effective for assembling chromosome-anchored salmonid genomes when combined with additional scaffolding information (De Kayne et al 2020; also see RefSeq: GCF_002021735.2).

Salmonid genomes are highly complex and relatively difficult to assemble owing to ancestral autotetraploidy (Maqueen and Johnston 2014) and high repeat content (Lien et al 2016; De-Kayne et al. 2020; Kajitani et al 2014). Sequencing low-diversity individuals from inbred lines or homozygous individuals produced via chromosome set manipulations provides one route for simplifying assembly in such species. Previous salmonid genome assemblies have made use of doubled haploid individuals (Lien et al. 2016; Christensen et al. 2018b; Pearse et al. 2019) because these individuals are theoretically homozygous at all loci (but see Lien et al. 2016). However, it should be noted that the highly contiguous assembly produced by DeKayne et al. (2020) for European Whitefish (*Coregonus sp. balchen*) was produced using data from an outbred, wild-caught individual.

Here we present a chromosome-anchored reference genome for a double haploid Lake Trout that was assembled using Pacific Bioscience long-read sequencing data and scaffolded using a high-density linkage map (Smith et al. 2020) and genome-wide chromatin conformation capture followed by massively parallel sequencing (Hi-C). We also produced a number of complementary resources including a custom repeat library, an interpolated recombination map, and a set of publicly available gene annotations in order to facilitate additional research on this important species. Additionally, we identify Lake Trout homeologs resulting from the Salmonid specific autotetraploid event (Ss4R) and establish homologous relationships with chromosomes

from other salmonid species.

II. MATERIALS AND METHODS

IIA. CROSSING AND SAMPLE COLLECTION

Gynogenetic double haploids were produced by fertilizing eggs with UV irradiated sperm, then pressure shocking embryos immediately following the first mitotic division (as described in Thorgaard et al. 1983; Limborg et al. 2016). Double haploid (DH) offspring were created at Pendill's Creek National Fish Hatchery using eggs and sperm collected from captive adult Lake Trout from the Seneca Lake brood stock. Due to low survivorship of DH offspring (Komen & Thorgaard 2007), we tested multiple UV and pressure shock treatments on eggs from five different females. Batches of 900 eggs from each female were fertilized with sperm that was irradiated for 140, 280, or 1,260 seconds. Each batch was then split and sub-batches were pressure shocked at 11,000 PSI for five minutes at either 6.5, 7, 7.5, 8, 8.5, 9, 9.5, or 10 hours post-fertilization. One batch of 900 eggs from each female was also exposed to a control treatment which involved no sperm irradiation or pressure shock. Embryos were incubated in heath trays at ambient temperature until eye-up stage ($E^3 6$ per Balon 1980), with dead embryos being removed from trays on a daily basis. Individuals surviving past post-embryo stage (*sensu*Marsden et al. 2021) were euthanized using a lethal dose of tricaine methanesulfonate (MS-222) and flash frozen in liquid nitrogen. Prior to sequencing and assembly, we verified that the DH individual chosen for sequencing was completely homozygous at 15 microsatellite loci that are typically highly heterozygous in the Seneca Lake hatchery population (Valiquette et al. 2014). One of the two individuals was selected for high molecular weight DNA extraction and long-read sequencing.

IIB. LABORATORY METHODS

A long-read sequencing library was prepared for the selected individual using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, California), with the optional DNA Damage Repair step after size selection. Size selection was made for >10 kb using a Blue Pippin instrument (Sage Science, Beverly, Massachusetts) according to the manufacturer recommended protocol for 20kb template preparation. 5ug of concentrated DNA was used as input for the library preparation reaction. Library quality and quantity were assessed using a genomic DNA Tape Station assay (Agilent, Santa Clara, California), as well as Broad Range and High Sensitivity Qubit fluorometric assays (Thermo Fisher, Waltham, Massachusetts). Single-Molecule Real Time sequencing was performed on the Pacific Biosciences Sequel instrument at the McGill Genome Centre (McGill University, Montreal, Canada, https://www.mcgillgenomecentre.ca/) using an on-plate concentration ranging from 1.5-7.5pM and the Sequel Sequencing Kit 2.0 with diffusion loading. 38 SMRTCells were run with 600-minute movies and two SMRTCells were run with 1200-minute movies.

Hi-C proximity ligation libraries were generated to aid with assembly scaffolding. Two libraries were prepared from spleen and muscle tissue using library preparation kits manufactured by Kapa Biosystems (Wilmington, Massachusetts) and Lucigen (Middleton, Wisconsin), respectively. Each Hi-C library was spiked into a portion of an Illumina HiSeq lane in order to assess how effectively reads could be mapped against the draft contig assembly. Genpipes version 3.1.5 (Bourgey, Dali et al. 2019) and HiCUP version 0.7.2 (Wingett, Ewels et al. 2015) were used to map Hi-C sequencing reads. The Hi-C library prepared using muscle-derived DNA and prepared using the Arima-Hi-C Lucigen Kit (Arima Genomics, San Diego, CA), was selected for further sequencing. This kit employs a restriction enzyme cocktail that digests chromatin at N^GATC and G^ANTC sequence motifs.

DNA was also extracted from four Lake Trout from the Seneca Lake, Isle Royale, and Green Lake hatchery broodstocks using MagAttract HMW DNA extraction kits (Qiagen, Hilden, Germany) for the purpose of generating Illumina shotgun sequencing data. Sequencing reads from Seneca origin individuals were later used for contig polishing and correction (described below in *Assembly and Scaffolding*). Libraries were

prepared for these individuals using 100ng of input DNA and the NEBNext Ultra Library Preparation Kit for Illumina (New England Biolabs, Ipswich, Massachusetts). Libraries sheared to approximately 400 bp using a Covaris M220 Ultrasonicator, amplified for eight cycles, and quantified using Quant-It Picogreen dsDNA assays (Thermo Fisher, Waltham, Massachusetts) run in triplicate. Fragment size was assessed using a genomic DNA Tape Station assay (Agilent, Santa Clara, California). Libraries were multiplexed in equal concentrations and sequenced in two HiSeqX lanes using paired end 2x150 format by the Novogene Corporation (Beijing, China).

IIC. ASSEMBLY AND SCAFFOLDING

Assembly was carried out using the polished_falcon_fat assembly workflow run using the SMRT Analysis v3.0 pbsmrtpipe workflow engine provided with an installation of SMRT Link v5.0 (smrtlink-release_6.0.0.47841; https://github.com/PacificBiosciences/pbsmrtpipe). Read metadata were extracted using the SMRT Analysis v3.0 dataset tool with the merge option. Sequencing read metadata, pipeline settings, and an output directory were specified for the polished_falcon_fat pipeline option. Default assembly settings were used except genome size (HGAP_GenomeLength_str) was set to 3 gigabases (GB), seed coverage (HGAP_SeedCoverage_str) was set to 40X, and the minimum read length to use a read as a seed (HGAP_SeedLengthCutoff_str) was set to 1000. Multiple compute settings were also changed. The resulting assembly settings file, read metadata file, and commands used to run the pipeline are available in *Supplemental Material 1 -Assembly Parameters).*

The polished_falcon_fat workflow uses FALCON assembly algorithm (Chin et al. 2013) and the Quiver/Arrow consensus tool (https://github.com/PacificBiosciences/GenomicConsensus) to generate a polished contig assembly. The Falcon method operates in two phases: First, overlapping sequence reads were compared to generate accurate consensus sequences with read N50 greater than 10.9Kb. Next, overlaps between the corrected longer reads were used to generate a string graph. The graph was reduced so that multiple edges formed by heterozygous structural variation were replaced to represent a single haplotype. Contigs were formed by using the sequences of nonbranching paths. Two supplemental graph cleanup operations were applied to improve assembly quality by removing spurious edges from the string graph: tip removal and chimeric duplication edge removal. Tip removal discards sequences with errors that prevent 5' or 3' overlaps. Chimeric duplication edges may result from the raw sequence information or during the first sequence cleanup step and artificially increase the copy number of a duplication. In a second and final workflow stage, the polished_falcon_fat workflow used the Arrow consensus tool to perform error correction on the assembly and generate an initial polished assembly. The resulting contigs were passed through a second round of error correction using Pilon in order to resolve SNP, indel, and local assembly errors before proceeding with scaffolding (https://github.com/broadinstitute/pilon). The Illumina paired-end sequencing dataset described above was used as input for Pilon after removing adapters and trimming reads using the sliding window approach implemented in Trimmomatic v0.32 (Bolger et al. 2014).

We adopted a multifaceted scaffolding approach leveraging information from Hi-C sequencing and a high-density linkage map for Lake Trout (Smith et al. 2020). Hi-C reads were mapped to Pilon corrected contigs with default setting using the Arima Genomics Mapping pipeline (Arima Genomics, https://github.com/ArimaGenomics/mapping_pipeline), which included four primary steps. First, forward and reverse reads were mapped to the reference genome using bwa version 0.7.17 (Li 2013) separately. Next, the 5' end of the mapped reads were trimmed. Samtools version 1.9 (Li, Handsaker et al. 2009) was then used to filter reads with mapping quality (MAPQ) less than 10. Finally, Picard version 2.17.3 (https://broadinstitute.github.io/picard/) was used to add read group information and mark duplicate reads. The resulting BAM file was used as input for SALSA v2.2 (Ghurye et al. 2017) run with default settings (three iterations). We also tested Salsa2 using five iterations and compared results with those produced using default settings by calculating Spearman's rank order correlation coefficients between the order of loci on the Lake Trout linkage map (Smith et al. 2020) and the order of loci on the 50 largest scaffolds. Linkage mapped RAD contigs were aligned to the reference assembly using minimap2 using the -asm5 option. RAD

contigs with mapping qualities less than 60 were removed before calculating correlation coefficients using the *cor* function and the *method* argument set to "spearman."

Additional scaffolding was carried out using Chromonomer v1.13 (Catchen et al. 2020). The assembly was initially scaffolded using default settings, which yielded chromosome length scaffolds with a high degree of concordance with the linkage map; however, structural differences between the linkage map and scaffolds were apparent on six chromosomes. In order to resolve these inconsistences, we aligned the full set of PacBio subreads to the assembly using Minimap2 (Li 2018) using the preset option for PacBio data. The resulting bam file was sorted, indexed, and per-base coverage was calculated for all positions using samtools depth with the –a option. We then ran a second round of Chromonomer using the –rescaffold, –depth, and depth_-stdevs = 2 options, which allowed for gaps to be opened in contigs if the site-specific depth within a sliding window of 1000 base pairs was greater than 2 standard deviations from the mean, suggesting an assembly error. This resulted in an assembly with improved concordance with the linkage map; however, linkage group 41 still exhibited a large inversion relative to the scaffolds. We determined the approximate location of this assembly error by identifying the pair of linkage mapped loci for which the level of discordance between the linkage map and assembly was maximized. The scaffold was manually broken and reoriented using an existing gap that existed between these two loci.

Gaps were filled using PBJelly from PBSuite v15.8.24 (English et al. 2012). All PacBio reads were aligned to the draft assembly using Minimap2 using the -pb preset option and reads mapping within 5000 base pairs of a gap were retained for gap filling using bedtools intersect (Quinlan and Hall 2010). Retained reads were re-mapped with Blasr v5.3.2 (Chaisson et al. 2012) using the options –minMatch 11, –minPctIdentity 75, –bestn 1, –nCandidates 10, –maxScore -500, and –fastSDP. The "maxWiggle" argument was set to 100 kilobases (KB) for the PBJelly assembly stage in order to account for gaps of unknown length. After filling gaps, we corrected single nucleotide and short indel errors by running 3 iterations of Polca (distributed with MaSuRCA v. 3.4.2; Zimin and Salzberg 2020) using Illumina data from a Seneca strain female as input. Default settings were used except alignments overlapping gaps were removed from bam files using bedtools intersect (Quinlan and Hall 2010) prior to running the Polca variant calling step.

Illumina paired end data from the same individual used for genome polishing and PacBio data from one SMRTcell were aligned to the Arctic Char (*Salvelinus alpinus*) mitome (RefSeq: NC_000861.1) in order to obtain reads useful for assembling the Lake Trout mitome. Reads were aligned using Minimap2 using the sr and map-pb present options for short-reads and long-reads, respectively. Reads aligning to the Arctic Char mitome were extracted from original fastq files using seqtk subseq (https://github.com/lh3/seqtk) and hybrid assembly was conducted using Unicycler v0.4.8 (Wick et al. 2017) using the settings –min_fasta_-length 15000 and –keep 0. Unicycler implements a hybrid-assembly approach using Spades (Bankevich et al. 2012), SeqAn (Doring et al. 2008), and Pilon. First, Spades (v3.13.1 here) was used to assemble Illumina short-reads and contigs with graph coverage less than half the median coverage were removed due to potential contamination from the nuclear genome. Contigs were then scaffolded using long-reads and SeqAn (Doring et al. 2008) was used to generate gap consensus sequences. Finally, Pilon was used to resolve assembly errors using short-read alignments as input.

IIG. ASSEMBLY QUALITY CONTROL

We used multiple approaches to assess the accuracy, contiguity, and completeness of the genome assembly. First, we determined the proportion of the genome that was recovered in our assembly by comparing total assembly size with an estimate of genome size based on the distribution of k-mer frequencies from Illumina paired-end 2x150 data generated using DNA from a Seneca strain female. The frequency of all 19mers in the read data was calculated using the count function in Jellyfish v2.2.6 (Marcais and Kingsford 2012) with the options -m 19 and -C. K-mer counts were then exported to the histogram format using the *histo* function. This file was used as input for GenomeScope v1.0 (http://qb.cshl.edu/genomescope/; Vurture et al. 2017) with read length set to 150 bp and k-mer length set to 19.

Basic assembly statistics were calculated using the program summarizeAssembly.py from PBSuite v15.8.24 (English et al. 2012). Statistics included total assembly size, contig and scaffold N50s, and minimum and maximum contig and scaffold lengths. Assembly statistics were calculated with and without gaps. Contig and scaffold N50s and counts were obtained for 14 additional salmonid assemblies from NCBI for comparison. Single base consensus accuracy was estimated after each iteration of polishing with Polca.

Next, we calculated percentages of complete singleton, complete duplicated, fragmented, and missing Benchmarking Single Copy Orthologs (BUSCOs) for seven chromosome-level salmonid assemblies and compared these with scores for the Lake Trout assembly discussed here. These included genomes for Brown Trout (*Salmo trutta* ; GCA_901001165.1), European Whitefish (*Coregonus sp. balchen* ; GCA_902810595.1; DeKayne et al. 2020), Atlantic Salmon (Salmo salar; GCA_000233375.4; Lien et al. 2016), Coho Salmon (*Oncorhynchus kisutch* ; GCA_002021735.1), Rainbow Trout (*Oncorhynchus mykiss* ; GCA_002163505.1; Pearse et al. 2019), Chinook Salmon (*Oncorhynchus tshawytscha* ; GCA_002872995.1; Christensen et al. 2018b), and Dolly Varden (*Salvelinus malma* ; GCA_002910315.1; Christensen et al. 2018a). It should be noted that the assembly originally produced for Arctic Char (GCA_002910315.1; Christensen et al. 2018a, referred to as the Dolly Varden assembly here) was later found to be from a Dolly Varden or potentially a Dolly Varden – Arctic Char hybrid (see Shedko et al. 2019 and Christensen et al. 2021). BUSCO scores were also calculated for the Northern Pike genome (*Esox lucius* ; GCA_000721915.3), a member of the order *Salmoniformes* that is commonly used as a pre-Ss4R outgroup species. BUSCO scores were calculated using BUSCO v4.0.6, the actinopterygii_odb10 database (created November 20th, 2019), and the -genome option.

Finally, we aligned the linkage mapped contigs from Smith et al. (2020) to the final assembly and calculated Spearman's rank order correlation coefficients between physical mapping locations and the order of loci along linkage groups. Linkage mapped contigs were aligned to the reference assembly using minimap2 using the -asm5 preset parameters and the resulting sam file was filtered to exclude contigs with mapping qualities less than 60. Correlation coefficients were calculated using the *cor* function in R (R Core Team 2017) with the *method* argument set to "spearman." Correlation coefficients were then converted to absolute values using the *abs* function in order to compare chromosomes and linkage groups with reversed orientations.

III. REPETITIVE DNA

A custom repeat library was created using RepeatModeler v2.0.1 (Flynn et al. 2020) and repeats were subsequently classified using RepeatClassifier (Smit et al. 2015). Repeats were then masked using RepeatMasker (Smit et al. 2015) and the output of RepeatMasker was used to determine the genome-wide abundance of different repeat families and the relative density of repeat types across chromosomes. The density of the most abundant repeat type (Tcl-mariner) was visualized across chromosomes using the R-package circlize (Gu et al. 2014; Figure 2).

IIK. HOMEOLOG IDENTIFICATION AND SYNTENY

We performed a self-vs-self synteny analysis using SyMap v5 (Soderlund et al. 2006; Soderlund et al. 2011) to identify Lake Trout homeologs resulting from the Salmonid specific autotetraploid event (Macqueen and Johnston 2014; Lien et al. 2016). Prior to running SyMap, we hard-masked the genome using RepeatMasker v4.1.0 (Smit et al. 2015) using our custom repeat library as input and RMblast as the search engine (-e ncbi). Nucmer was used for SyMap alignments and options were set to –min-dots = 30, top_n = 2, and merge_blocks = 1. We then used Symap to identify blocks of synteny between Lake Trout and Dolly Varden, Rainbow Trout, and Atlantic Salmon. Alignments were conducted using Promer, and we used the options min_dots = 30, top_n = 1, merge_blocks = 1, and no_overlapping_blocks = 1. Results from self-vs-self synteny analysis were visualized using the R-package circlize (Gu et al. 2014). Results from the species-vs-species synteny analysis were visualized using the Chromosome Explorer option in Symap v5 (*Supplemental Material 4 – Syntenic Blocks and Between Species Circos Plots*).

IIF. RNA SEQUENCING AND GENE ANNOTATION

RNA samples were obtained from the offspring of Seneca Lake hatchery strain fish held within the Ontario Ministry of Natural Resources and Forestry (OMNRF) hatchery system. Offspring were produced using four males and four females in a full factorial mating cross, by dry-spawning anesthetized fish (anesthetic: 0.1 g L-1 MS-222; Aqua Life, Syndel Laboratories Ltd., B.C., Canada). Eggs (140 mL) were stripped from each female, divided evenly among four jars, and fertilized by pipetting milt directly onto them. After fertilization, embryos were transported to the Codrington Fish Research Facility (Codrington, Ontario, Canada) where they were transferred from the jars into perforated steel boxes with one family per box. These boxes were contained in flow-through tanks receiving freshwater at ambient temperature (5-6) and natural photoperiod under dim light. When the embryos fully absorbed their yolk sacs and were ready to feed exogenously (i.e. free embryos; approximately March 2016), 14 individuals from each family were randomly selected and split into two groups of seven, then transferred into one of four larger (200 L) tanks.

Tissue sample collection occurred between June 28 to August 9, 2016. Each fish was euthanized in a bath of 0.3 g L-1 of MS-222 and dissected to remove the whole liver. The liver was gently blotted on a lab wipe and stored in RNAlater (Invitrogen, Thermo Fisher Scientific) for 24-48 hours at room temperature. RNALater was pipetted from the liver tissue and the samples were stored at -80 until RNA isolation. Liver tissues were homogenized individually in 2 mL Lysing Matrix D tubes (MP Biomedicals) with 1 mL of Trizol reagent (Invitrogen, Thermo Fisher Scientific). RNA was extracted from the homogenate using phenol-chloroform extraction (Chomczynski & Sacchi, 2006). RNA was precipitated with RNA precipitation solution (Sambrook & Russel, 2001) and isopropanol, and washed with 75% ethanol. RNA samples were resuspended in nuclease-free water (Thermo Fisher Scientific). The purity and concentration of the RNA were initially determined using a NanoDrop-8000 spectrophotometer. RNA quality was also assessed using a Bioanalyzer (Agilent) and resulting RNA integrity numbers (RIN). All RNA samples met our minimum RIN threshold of 7.5.

RNA sequencing was performed over two years. Twenty-four samples were sent to The Centre for Applied Genomics (Sick Kids Hospital, Toronto, Ontario, Canada) in 2018, and another 30 samples were sent to the Centre d'expertise et de services Genome Quebec (Montreal, Quebec, Canada; https://cesgq.com/) in 2020. cDNA libraries were produced by enriching the poly(A) tails of mRNA with oligo dT-beads using the NEBNext Ultra II Directional polyA mRNA Library Prep kit (New England Biolabs; Ipswich, Massachusetts). The group of 24 individuals was sequenced in 2.5 Illumina HiSeq 2500 lanes using 2X126 bp paired end reads. The additional thirty individuals were sequenced in three Illumina HiSeq 4000 lanes using 2X126 bp paired end reads. Data were deposited in sequence read archives associated with BioProject PRJNA682236. These sequencing reads, along with those from two previous RNAseq experiments (Goetz et al. 2010; Goetz et al. 2016), were used as input for NCBI's Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2016).

IIE. RECOMBINATION RATES AND CENTROMERES

Sex averaged recombination rates were estimated across chromosomes using the sliding window interpolation approach implemented in MareyMap (Rezvoy et al. 2007). Restriction site associated DNA (RAD) contigs from the Lake Trout linkage map (Smith et al. 2020) were mapped to chromosomes using minimap2 using the -asm5 preset option and reads with mapping qualities less than 60 were removed. At this point, RAD loci overlapping centromere mapping intervals for each linkage group were extracted and the centromere center was considered to be the mean mapping position for centromere associated RAD tags. Centromere positions were visualized using the R-package circlize (Gu et al. 2014).

In order to remove contigs with anomalous mapping positions that could bias recombination rate estimates, we fit a loess model describing linkage map position as a function of physical position for each chromosome, extracted model residuals, and removed markers with residuals that were greater than one standard deviation from the mean. Loess models were fit using the *loess* function in R with the span argument set to 0.2 and

the degree argument set to 2. The remaining markers were output to MareyMap format and were manually curated using MareyMap Online (Siberchicot et al. 2017). A sex averaged recombination map was calculated using sliding window interpolation and exported from the program (*Supplemental Material 3 – Recombination Map*).

III. RESULTS

IIIA. SEQUENCING, ASSEMBLY, AND SCAFFOLDING

Of the 13,500 embryos exposed to UV irradiation and pressure shock treatments, two individuals survived beyond the post-embryo stage. The individual selected for assembly was found to be homozygous at all 15 genotyped microsatellite loci, suggesting that chromosome set manipulations were successful at inducing doubled haploidy. We proceeded with PacBio sequencing, and produced a dataset with an estimated genome coverage of 89X, with 53X coverage provided by reads longer than 12 KB in length.

The Falcon-based assembly pipeline and polishing with Arrow and Pilon yielded an initial assembly with 8,321 contigs, a total length of 2.3 GB, and a contig N50 of 1.3 megabases (MB) with a maximum contig length of 19.6 MB. Our analysis comparing the correlation between the Lake Trout linkage map and Hi-C scaffolds indicated that three iterations of Salsa (the default setting) produced moderately large scaffolds, while yielding a mean map versus scaffold correlation of 0.89. Thirty-three of the 50 largest scaffolds had correlations greater than 0.95 and 42 had correlations greater than 0.8. We opted to use these settings for scaffolding. Salsa v2.2 split multiple contigs, resulting in 8,367 contigs with an N50 of 1.25 MB and 5,171 scaffolds with an N50 of 5.15 MB. Additional scaffolding with Chromonomer v1.13 increased scaffold N50 to 44 MB and reduced the total number of scaffolds to 4,122. Chromonomer v1.13 also reduced contig N50 to a small degree due to the insertion of additional gaps at likely misassembles. Scaffolding with Hi-C and the Lake Trout linkage map ultimately allowed us to assign 84.7% of the genome to chromosomes. Gap filling with PBJelly increased scaffold N50 to 44.97 MB, increased the total assembly size to 2.345 GB, and increased contig N50 to 1.8 MB. Gap filling increased the maximum contig length to 34.78 MB and the maximum scaffold length to 98.19 MB. The estimated consensus accuracy after three rounds of error correction with Polca was 99.9959 %. The polished assembly was submitted to GenBank for public use (accession GCA_016432855.1).

IIIB. ASSEMBLY QUALITY CONTROL

We estimated the total haploid genome size for Lake Trout to be between 2.119 and 2.122 GB using k-mer analysis and GenomeScope v1.0, with 38% of the genome composed of unique sequence and 62% composed of repetitive sequence. Heterozygosity for the sample used for polishing was estimated to be between 2.78 and 2.9 heterozygous sites per 1000 base pairs. It should be noted that the individual used for polishing was a diploid and not a gynogenetic double haploid. The estimated coverage for the sample used for genome-size estimation was 16X, which should be sufficient for k-mer based methods (Williams et al. 2013).

We recovered 93.2% of BUSCO genes with 60.3% and 32.9% being present as singletons and duplicates, respectively (Figure 3). The salmonid genomes evaluated recovered between 88.1% and 95.3% complete BUSCOs with between 25.3% and 34.9% being duplicated and between 58.3% and 65% being singletons. The proportion of duplicated BUSCOs in the Lake Trout genome was the second highest among salmonid genomes (32.9%) and appears to be comparable to the Brown Trout genome (GCA_901001165.1; River Trout), which was also assembled using Falcon (Falcon-unzip) and polished using a method based on the Freebayes variant caller (Garrison and Marth 2012).

Spearman's rank order correlations between the genome assembly and the Lake Trout linkage map ranged from 0.89 to 1.0 for the 42 Lake Trout chromosomes. The mean correlation was 0.98 and 39 of 42 chromosomes

had correlations greater than or equal to 0.96, suggesting that the final genome assembly provides an accurate representation of the order of loci along Lake Trout chromosomes.

IIIC. REPETITIVE DNA

RepeatModeler 2 identified 2,810 interspersed repeats and 462 of these were classified by RepeatClassifier. RepeatMasker reported that 53.8% of the Lake Trout genome is composed of sequences from this repeat library. A total of 13.04% of the genome was composed of retroelements, with 10.47% being LINEs and 2.57% being LTR elements, and 9.97% of the genome was composed of DNA transposons. As has been observed in other salmonids, TcMar-Tc1 was the most abundant superfamily and these repeats were most abundant near centromeres (Figure 2; Lien et al. 2016; Pearse et al. 2019). A total of 30.79% of the genome was composed of interspersed repeats that were not classified by RepeatClassifier.

IIID. HOMEOLOG IDENTIFICATION AND SYNTENY

Self-vs-self synteny analysis conducted using Symap v5 identified 126 syntenic blocks shared between putative Lake Trout homeologs (Figure 2). Blocks ranged in size from 477,153 bp to 57,126,662 bp. Fifty-two blocks were longer than 10 MB and 70 were longer than 5 MB (Figure 2, inner links). We identified 50 syntenic blocks shared between Rainbow Trout and Lake Trout and identified homologous rainbow trout chromosomes for all Lake Trout chromosomes. Syntenic blocks shared between these two species ranged in size from 1.9 MB to 97.2 MB. Symap identified homologous chromosomes in Atlantic Salmon for all chromosomes except 32 and 39. However, we expect that Lake Trout chromosome 39 is homologous to a region of Atlantic Salmon chromosome 2 and chromosome 32 is homologous with a region of chromosome 14 based on the size of missing synteny blocks. Fifty-four syntenic blocks were detected between the two species that ranged in size from 208,516 bp to 88 MB. We identified 42 syntenic blocks shared between Dolly Varden and Lake Trout and identified homologs for all chromosomes except chromosome 41. Syntenic blocks ranged in size from 6.8 MB to 79.9 MB (*Supplemental Material 4 – Syntenic Blocks and Between Species Circos Plots).*

IIID. GENOME ANNOTATION

We generated a total of 3.45 billion RNA-seq reads that were subsequently used as input for the NCBI Eukaryotic Genome Annotation Pipeline v8.5 (July 9, 2020 release date). An additional 528,760 reads were used from previous Lake Trout gene expression studies. A total of 86% of reads were aligned to the genome assembly, and 12 Lake Trout transcripts from GenBank and 3,547 known Atlantic Salmon transcripts from RefSeq were also used as input for the pipeline.

The pipeline produced annotations for 49,668 genes and pseudogenes. A total of 3,307 non-transcribed pseudogenes and two transcribed pseudogenes were identified. Gene length ranged from 53 to 1,198,409 bp, with a median length of 8,676 bp. Gene densities for chromosomes ranged from 15.45 to 31.39 genes/mb with an average genome-wide density of 21.07 genes/mb (Figure 2, C). A total of 422,014 exons were identified, with between 1 and 224 exons per transcript (mean=10.31, median=8).

IIIE. RECOMBINATION RATES AND CENTROMERES

We were able to map between 1 and 238 centromere-associated RAD contigs to their respective chromosomes and determine approximate centromere locations for all chromosomes except chromosome 42. Smith et al. (2020) did not determine the location of the centromere for chromosome 42, which prohibited us from identifying its location. Across all chromosomes, we mapped 35 centromere-associated RAD loci to each chromosome on average. Between 39 and 238 centromeric loci were mapped to metacentric chromosomes (mean = 93), while between 1 and 59 loci were mapped for acrocentric or telocentric chromosomes (mean = 21).

In all, 14,438 linkage mapped contigs were mapped to the genome with mapping qualities greater than 60. A total of11,232 loci were retained for recombination rate estimation after manual curation and filtering using loess model residuals. We determined the mean sex averaged recombination rate to be 1.09 centimorgans/mb, with recombination rates varying between 0 and 6.58 centimorgans/mb across the genome.

IV. DISCUSSION

The adoption of multiple complementary scaffolding approaches resulted in an assembly of similar quality to the best available salmonid genomes. Multiple lines of evidence suggest that the genome presented here represents a nearly complete and accurate model of the Lake Trout genome. First, the total size of the finished genome was slightly greater than the genome size estimate obtained from GenomeScope. Pflug et al. (2020) found that k-mer based methods for genome size estimation tend to underestimate genome size by 4.5% on average, so this result is not entirely unexpected. Additionally, BUSCO scores were similar to those obtained for the highest quality salmonid genomes available at the time of analysis (e.g. Coho Salmon, Brown Trout, Rainbow Trout). Scores were highly similar between Brown Trout and Lake Trout genomes; however, the proportion of missing BUSCOs was 1.9% higher for Lake Trout and the proportion of complete duplicated BUSCOs was 2% lower suggesting that some duplicated regions might be missing from the Lake Trout genome. Nonetheless, these two assemblies had the highest percentage of complete BUSCOs and the highest percentage of complete duplicated BUSCOs out of the genome assemblies examined. Furthermore, the order of loci on the Lake Trout linkage map and the order of loci on Lake Trout chromosomes was shown to be highly concordant, suggesting that contigs are accurately ordered and properly oriented. The genome presented here is also highly contiguous, with a contig N50 higher than any published salmonid genome (but see the recently released assembly for Arlee Strain Rainbow Trout; GCF_013265735.2). Interestingly, the PacBio data used for assembly were of similar coverage to the data used for assembling the European Whitefish genome (De-Kayne et al. 2020); however, the Lake Trout genome contig N50 is >3X higher (although scaffold N50 is lower). There are two reasonable explanations for this. First, the European Whitefish genome was assembled using DNA from a wild-caught, outbred individual rather than a double haploid. Second, the European Whitefish genome was not gap filled after scaffolding. Gap filling the Lake Trout genome with PBJelly increased contig N50 by 561,496 bp.

The Lake Trout genome will likely be sufficient for the majority of downstream uses; however, improvements could likely be made using supplementary scaffolding resources such as a higher density linkage map or optical map (Pan et al. 2020). The annotation could also be improved by generating additional RNA-seq data. The number of annotated genes and pseudogenes (n=49,668) is similar to what has been obtained for other salmonids (eg Chum salmon *Oncorhynchus keta* , Sockeye salmon*Oncorhynchus nerka* , and Dolly Varden) using the same annotation pipeline. However, it is important to note that annotation completeness is markedly reduced relative to other assemblies with similar BUSCO scores such as Atlantic Salmon (57,783; GCF_000233375.1; Annotation Release 100), Coho Salmon (63,465; GCF_002021735.2; Annotation Release 101), Brown Trout (61,583; GCF_901001165.1; Annotation Release 100), Rainbow Trout (55,630, GCF_-002163495.1, Annotation Release 100), and Chinook Salmon (53,685, GCF_002872995.1, Annotation Release 100). These annotations were produced using RNA-seq evidence from a greater diversity of tissue types, which likely explains this discrepancy. The Lake Trout annotation, as well as annotations for other salmonids, could also be further improved by directly sequencing full length transcripts using long-read sequencing technologies (Workman et al. 2018). We predict that the completeness of the Lake Trout genome annotation will be improved as more gene expression data from a greater diversity of tissue types becomes available for the species (Salzberg 2019). Nonetheless, the current genome annotation will undoubtably aid in the interpretation of future findings by allowing researchers to link signals of selection and loci associated with phenotypes with putatively causal genes and biological processes. Publicly available gene expression and functional annotation resources, like those being developed by the Functional Annotation of All Salmonid Genomes (FAASG) initiative, will also aid in this effort (Macqueen et al. 2017).

The availability of a second high-quality assembly for a*Salvelinus* species will likely benefit comparative genomic research aimed at understanding the evolutionary consequences of genome duplication. Salmonids have long been appreciated as a model system for understanding evolution following whole genome duplication events (Ohno 1970) and the wealth of genomic resources for salmonids will hopefully continue to shed light on the evolutionary processes at play following autotetraploid genome duplication events. Additionally, multiple recent studies have highlighted the importance of structural genetic variation for promoting adaptive diversification within salmonid species (Pearse et al. 2019; Bertolotti et al. 2020), and chromosome-anchored genome assemblies are typically needed for detecting and genotyping structural variants (Merot et al. 2020).

Genomic methods have dramatically increased the precision of population genetic analyses and have enabled researchers to address qualitatively unique questions that require some knowledge of genome structure and function (Waples et al. 2020). Lake Trout have undergone repeated parallel adaptive radiations and ecotypic diversity appears to be heritable (Goetz et al. 2010); however, the genetic or epigenetic basis for ecotypic diversity is still unclear (Perreault-Payette et al. 2017). A genome assembly will greatly simplify the process of mapping loci associated with ecophenotypic differentiation and could enable identification of loci associated with reproductive isolation among ecotypes in populations where multiple ecotypes exist. Anecdotal evidence suggests that Lake Superior once harbored as many as ten ecotypes (Goodier 1981). Three ecotypes are contemporarily recognized (lean, siscowet and humper) and a fourth ecotype was recently identified (redfin; Muir et al. 2014). Interestingly, Muir et al. (2014) found that ecotypes collected near Isle Royale were moderately distinct, which is at odds with historical records suggesting that they were easy to identify visually (Rakestraw 1967). An improved understanding of the genetic basis for ecotypic differentiation could help determine if this is due to phenotypic plasticity, increased levels of hybridization between ecotypes, or other processes (Baillie et al. 2016). The ability to genotype historical collections and quantify levels of adaptive differentiation at different time points (Guinand et al. 2003) provides a particularly exciting avenue for future research on Lake Trout.

The Lake Trout genome assembly could also have important implications for ongoing Lake Trout restoration activities throughout the Great Lakes. The resources presented here will allow for the identification of loci associated with variation in fitness between Lake Trout hatchery strains in contemporary Great Lakes environments (Scribner et al. 2018) and the identification of loci that are adaptively diverged between hatchery strains. This information could help fisheries managers to maximize adaptive genetic diversity in re-emerging wild populations and prioritize hatchery populations for continued propagation.

V. ACKNOWLEDGEMENTS

We would like to thank hatchery personnel at Pendills Creek National Fish Hatchery (USFWS) and the OMNRF White Lake Fish Culture Station and Codrington Fish Research Facility for assistance with generating doubled haploid offspring and collecting tissue for RNA sequencing. We would also like the thank the Vertebrate Genomics and Evolution Group (VerGE, Spring 2021) at Michigan State University and the Population Genetics Seminar group (Spring 2021) at the University of Montana for helpful discussion and suggestions.

VI. FUNDING

SRS, KS, and GL were supported by award 2017_SCR_44067 from the Great Lakes Fisheries Commission. LB and EN were supported by award 2017_BER_44071 from the Great Lakes Fisheries Commission. LB was supported by the Canadian Research Chair in Genomics and Conservation of Aquatic Resources administered by the Canada Research Chair Program. JR was supported by CFI Grant 33408 from the Canada Foundation for Innovation, a Genome Technology Platform Grant from Genome Canada, and the CanSeq150 Sequencing Initiative. KTS was supported through the cooperative agreement Partnership for Ecosystem Research and

Management (PERM) between the Department of Fisheries and Wildlife at Michigan State University and the Michigan Department of Natural Resources.

VII. REFERENCES

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In Evolutionary genetics of fishes (pp. 1-53). Springer, Boston, MA.

Baillie, S. M., et al. (2016). "Loss of genetic diversity and reduction of genetic distance among lake trout Salvelinus namaycush ecomorphs, Lake Superior 1959 to 2013." Journal of Great Lakes Research 42(2): 204-216.

Balon, E. K. (1980). Charrs, salmonid fishes of the genus Salvelinus. Kluwer Boston.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., ... & Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. Nature communications, 11(1), 1-16.

Blackie, C.T., Weese, D.J., & Noakes, D.L.G. (2003). Evidence for resource polymorphism in the lake charr (Salvelinus namaycush) population of Great Bear Lake, Northwest Territories, Canada. Ecoscience 10(4), 509-514.

Bolger, A. M., Lohse, M., & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15), 2114-2120.

Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., ... & Bourque, G. (2019). GenPipes: an open-source framework for distributed and scalable genomic analyses. GigaScience, 8(6), giz037.

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. G3: Genes, Genomes, Genetics, 10(11), 4115-4128.

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics, 13(1), 1-18.

Chavarie, L., Howland, K., Harris, L., & Tonn, W. (2015). Polymorphism in lake trout in Great Bear Lake: intra-lake morphological diversification at two spatial scales. Biological Journal of the Linnean Society 114(1): 109-125.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., ... & Turner, S. W. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature methods, 13(12), 1050-1054.

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... & Koop, B. F. (2018). The Arctic Char (Salvelinus alpinus) genome and transcriptome assembly. PloS one, 13(9), e0204076.

Christensen, K.A., Leong, J.S., Sakhrani, D., Biagi, C.A., Minkley, D.R., Withler, R.E., Rondeau, E.B., Koop, B.F., & Devlin, R.H. (2018). Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. PloS One 13(4 (2018): e0195461.

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... & Koop, B. F. (2021). Retraction: The Arctic charr (Salvelinus alpinus) genome and transcriptome assembly.

Crete-Lafreniere, A., Weir, L. K., & Bernatchez, L. (2012). Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. PloS one, 7(10), e46662.

De-Kayne, R., Zoller, S., & Feulner, P. G. (2020). A de novo chromosome-level genome assembly of Coregonus sp."Balchen": one representative of the Swiss Alpine whitefish radiation. Molecular Ecology Resources.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., ... & Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS One, 7(11), e47768.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences, 117(17), 9451-9457.

Gagnaire P-A, Normandeau E, Pavey SA, Bernatchez L. 2013. Mapping phenotypic, expression and transmission ratiodistortion QTL using RAD marker in the Lake Whitefish (Coregonus clupeaformis). Molecular Ecology. 22: 3036-3048.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. BMC genomics, 18(1), 1-11.

Goetz, F., Rosauer, D., Sitar, S., Goetz, G., Simchick, C., Roberts, S., ... & Mackenzie, S. (2010). A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (Salvelinus namaycush). Molecular ecology, 19, 176-196.

Goetz, F., Smith, S. E., Goetz, G., & Murphy, C. A. (2016). Sea lampreys elicit strong transcriptomic responses in the lake trout liver during parasitism. BMC genomics, 17(1), 1-16.

Goodier, J. L. (1981). Native lake trout (Salvelinus namaycush) stocks in the Canadian waters of Lake Superior prior to 1955. Canadian Journal of Fisheries and Aquatic Sciences, 38(12), 1724-1737.

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics, 30(19), 2811-2812.

Guinand, B., K.T. Scribner, K.S. Page, and M.K. Burnham-Curtis. 2003. Genetic variation over space and time: analyses of extinct and remnant lake trout populations in the upper Great Lakes. Proc. Roy. Soc. Lond. 270: 425-434.

Hansen, Michael J. "Lake trout in the Great Lakes: basin-wide stock collapse and binational restoration." (1999): 417-453. Pages 417-453 in William W Taylor, C Paola Ferreri (eds). Great Lakes Fishery Policy and Management: A Binational Perspective. Michigan State University Press.

Hanson, S. D., Holey, M. E., Treska, T. J., Bronte, C. R., & Eggebraaten, T. H. (2013). Evidence of wild juvenile lake trout recruitment in western Lake Michigan. North American Journal of Fisheries Management, 33(1), 186-191.

Harris, L. N., Chavarie, L., Bajno, R., Howland, K. L., Wiley, S. H., Tonn, W. M., & Taylor, E. B. (2015). Evolution and origin of sympatric shallow-water morphotypes of Lake Trout, Salvelinus namaycush, in Canada's Great Bear Lake. Heredity, 114(1), 94-106.

Hotaling, S.; Kelley, J.L. The rising tide of high-quality genomic resources. Mol. Ecol. Resour. 2020, 19, 567–569.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome research, 24(8), 1384-1395.

Komen, H., & Thorgaard, G. H. (2007). Androgenesis, gynogenesis and the production of clones in fishes: a review. Aquaculture, 269(1-4), 150-173.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), 722-736.

Krueger, C. C., Horrall, R. M., & Gruenthal, H. (1983). Strategy for the use of lake trout strains in Lake Michigan. Wisconsin Department of Natural Resources, Administrative Report, 17.

Lantry JR (2015) Eastern basin of Lake Ontario warmwater fisheries assessment, 1976–2014. 2014 annual report, Bureau of Fisheries, Lake Ontario Unit and St Lawrence River Unit to the Great Lakes Fishery Commission's Lake Ontario Committee, pp. 1–35

Larson, W. A., Kornis, M. S., Turnquist, K. N., Bronte, C. R., Holey, M. E., Hanson, S. D., ... & Sloss, B. L. (2021). The genetic composition of wild recruits in a recovering lake trout population in Lake Michigan. Canadian Journal of Fisheries and Aquatic Sciences, 99(999), 1-15.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., ... & Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature, 533(7602), 200-205.

Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. Molecular Ecology. DOI: 10.1111/mec.13601

Lynch, M., & Force, A. G. (2000). The origin of interspecific genomic incompatibility via gene duplication. The American Naturalist, 156(6), 590-605.

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proceedings of the Royal Society B: Biological Sciences, 281(1778), 20132881.

Marcais, G., & Kingsford, C. (2012). Jellyfish: A fast k-mer counter. Tutorialis e Manuais, 1-8.

Marin, K., Coon, A., Carson, R., Debes, P. V., & Fraser, D. J. (2016). Striking phenotypic variation yet low genetic differentiation in sympatric lake trout (Salvelinus namaycush). PloS one, 11(9), e0162325.

Marsden, J. E., Noakes, D. L., & Krueger, C. C. (2021). Terminology Issues in Lake Charr Early Development. In A. M. Muir (Ed.), The Lake Charr Salvelinus namaycush: Biology, Ecology, Distribution, and Management (1st ed., Fish and Fisheries, Ser. 39, pp. 487-497). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-030-62259-6

Merot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. Trends in Ecology & Evolution, 35(7), 561-572.

Muir, A. M., Bronte, C. R., Zimmerman, M. S., Quinlan, H. R., Glase, J. D., & Krueger, C. C. (2014). Ecomorphological diversity of lake trout at Isle Royale, Lake Superior. Transactions of the American Fisheries Society, 143(4), 972-987.

Muir, A. M., Hansen, M. J., Bronte, C. R., & Krueger, C. C. (2016). If Arctic charr Salvelinus alpinus is 'the most diverse vertebrate', what is the lake charr Salvelinus namaycush?. Fish and Fisheries, 17(4), 1194-1207.

Ohno S. (1970). Evolution by gene duplication. New York: Springer-Verlag.

Pan, W., Jiang, T., & Lonardi, S. (2020). OMGS: optical map-based genome scaffolding. Journal of Computational Biology, 27(4), 519-533.

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadia-Cardoso, A., ... & Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. Nature Ecology & Evolution, 3(12), 1731-1742.

Perreault-Payette, A., Muir, A. M., Goetz, F., Perrier, C., Normandeau, E., Sirois, P., & Bernatchez, L. (2017). Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. Molecular Ecology, 26(6), 1477-1497.

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3: Genes, Genomes, Genetics, 10(9), 3047-3060.

Prince, D. J., O'Rourke, S. M., Thompson, T. Q., Ali, O. A., Lyman, H. S., Saglam, I. K., ... & Miller, M. R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. Science advances, 3(8), e1603198.

Pycha, R. L. (1980). Changes in mortality of lake trout (Salvelinus namaycush) in Michigan waters of Lake Superior in relation to sea lamprey (Petromyzon marinus) predation, 1968–78. Canadian Journal of Fisheries and Aquatic Sciences, 37(11), 2063-2073.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rakestraw, L. (1967). Post-columbian history of Isle Royale. Part II: fisheries. Master's thesis Houghton: Michigan Technological University

Doring, A., Weese, D., Rausch, T., & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. BMC bioinformatics, 9(1), 1-9.

Rezvoy, C., Charif, D., Gueguen, L., & Marais, G. A. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics, 23(16), 2188-2189.

Riley, S. C., et al. (2007). "Evidence of widespread natural reproduction by lake trout Salvelinus namaycush in the Michigan waters of Lake Huron." Journal of Great Lakes Research 33: 917-921.

Rougeux C, Gagnaire PA, Praebel K, Seehausen O, Bernatchez L. 2019. Polygenic selection drives the evolution of convergent transcriptomic landscapes across continents within a Nearctic sister-species complex. Molecular Ecology. 28:4388-4403.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature methods, 17(2), 155-158.

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. Genome Biology 20, 92. https://doi.org/10.1186/s13059-019-1715-2

Scribner, Kim, Iyob Tsehaye, Travis Brenden, Wendylee Stott, Jeannette Kanefsky, & James Bence. "Hatchery strain contributions to emerging wild lake trout populations in Lake Huron." Journal of Heredity 109, no. 6 (2018): 675-688.

Shedko, S. V. (2019). Assembly ASM291031v2 (Genbank: GCA_002910315. 2) identified as assembly of the Northern Dolly Varden (Salvelinus malma malma) genome, and not the Arctic char (S. alpinus) genome. arXiv preprint arXiv:1912.02474.

Siberchicot, A., Bessy, A., Gueguen, L., & Marais, G. A. (2017). MareyMap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome biology and evolution, 9(10), 2506-2509.

Smit, AFA, Hubley, R & Green, P. (2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Smith, S. H. (1968). "Species succession and fishery exploitation in the Great Lakes." Journal of the Fisheries Research Board of Canada 25: 667-693.

Smith, S. R., Amish, S. J., Bernatchez, L., Le Luyer, J., C. Wilson, C., Boeberitz, O., ... & Scribner, K. T. (2020). Mapping of Adaptive Traits Enabled by a High-Density Linkage Map for Lake Trout. G3: Genes, Genomes, Genetics, 10(6), 1929-1947.

Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome research, 16(9), 1159-1168.

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3. 4: a turnkey synteny system with application to plant genomes. Nucleic acids research, 39(10), e68-e68.

Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P. A., Murphy, T. D., ... & Souvorov, A. (2016). P8008 The NCBI Eukaryotic Genome Annotation Pipeline. Journal of Animal Science, 94(suppl_4), 184-184.

Thorgaard, Gary H., Fred W. Allendorf, & Kathy L. Knudsen. "Gene-centromere mapping in rainbow trout: high interference over long map distances." Genetics 103, no. 4 (1983): 771-783.

Valiquette E, Perrier C, Thibault I, Bernatchez L. 2014. Loss of genetic integrity in wild Lake Trout populations following stocking: Insights from an exhaustive study of 72 lakes from Quebec, Canada. Evolutionary Applications. 7: 625-644.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. Nature Reviews Genetics, 18(7), 411.

Veale, A. J., & Russello, M. A. (2017). An ancient selective sweep linked to reproductive life history evolution in sockeye salmon. Scientific Reports, 7(1), 1-10.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 33(14), 2202-2204.

Waples, R.S. K.A. Naish, and C.R. Primmer. 2020.Conservation and Management of salmon in the age of genomics. Ann. Rev. Anim. Biosci. 8: 117-143.

Whibley, A., Kelley, J., & Narum, S. (2020). The changing face of genome assemblies: guidance on achieving high-quality reference genomes. Molecular ecology resources.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS computational biology, 13(6), e1005595.

Williams, D., Trimble, W. L., Shilts, M., Meyer, F., & Ochman, H. (2013). Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genomics, 14(1), 1-11.

Willoughby, J. R., Harder, A. M., Tennessen, J. A., Scribner, K. T., & Christie, M. R. (2018). Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. Molecular Ecology, 27(20), 4041-4051.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. F1000Research, 4.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Zuzarte PC, et al. Nanopore native RNA sequencing of a human poly(a) transcriptome. bioRxiv; 2018. p. 459529. https://doi.org/10.1101/459529

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS computational biology, 16(6), e1007981.

Zimmerman, Mara S., Charles C. Krueger, & Randy L. Eshenroder. "Phenotypic diversity of lake trout in Great Slave Lake: differences in morphology, buoyancy, and habitat depth." Transactions of the American Fisheries Society 135, no. 4 (2006): 1056-1067.

VIII. DATA ACCESSIBILITY

The Lake Trout whole genome assembly project has been deposited at DDBJ/ENA/GenBank

under the accession JAEAGN000000000. The version described

in this paper is version JAEAGN010000000. The GenBank assembly accession is GCA_016432855.1. RNA sequencing data generated for annotation is available in Sequence Read Archives associated with BioProject PRJNA682236. Pacific Biosciences long-reads, Illumina paired-end reads, and Hi-C data will be made available in Sequence Read Archives associated with BioProject PRJNA682269 upon acceptance for publication. All references to the Lake Trout genome annotation are in reference to annotation release 100.

IX. AUTHOR CONTRIBUTIONS

SRS drafted the manuscript and assisted with genome assembly and scaffolding. GL and KS assisted with drafting the manuscript, provided funding support, and helped initiate the project. LB and AM oversaw all components of the project, provided funding support, and assisted with drafting the manuscript. CW coordinated creation of double haploid individuals, assisted with manuscript development, and oversaw the generation of RNA sequencing data. EN assisted with genome scaffolding, finishing, and assembly validation. JR coordinated and supervised long-read sequencing and Hi-C data generation and coordinated data analysis. PB carried out library preparation and PacBio sequencing. HD carried out long-read data quality control and long-read contig assembly. PNM carried out Hi-C scaffolding. CP generated RNA sequencing data and assisted with drafting the manuscript.

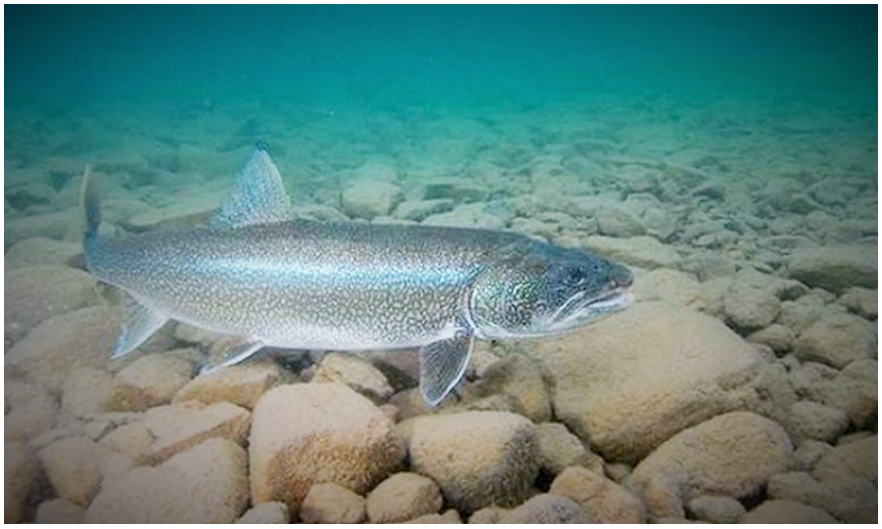
X. ANIMAL CARE AND USE

Experiments were conducted in accordance with the guidelines of the Canadian Council on Animal Care, and were approved by the Institutional Animal Care Committee of Trent University (Protocol # 24794) and the OMNRF Aquatic Animal Care Committee (Protocols #21 and #136).

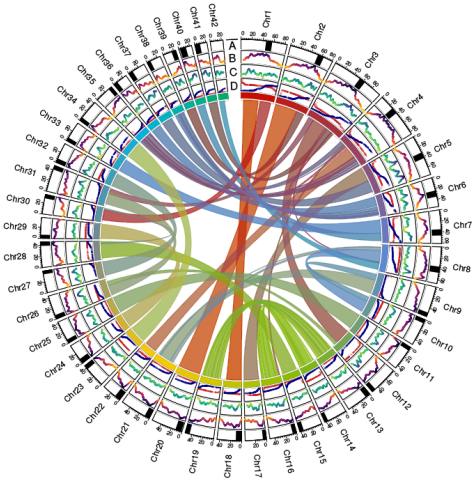
XI. FIGURES

XIA. FIGURE 1 – STUDY SPECIES

A photograph of an adult Lake Trout (*Salvelinus namaycush*) from Great Bear Lake, Northwest Territories, Canada. Photo credit: Andrew Muir.



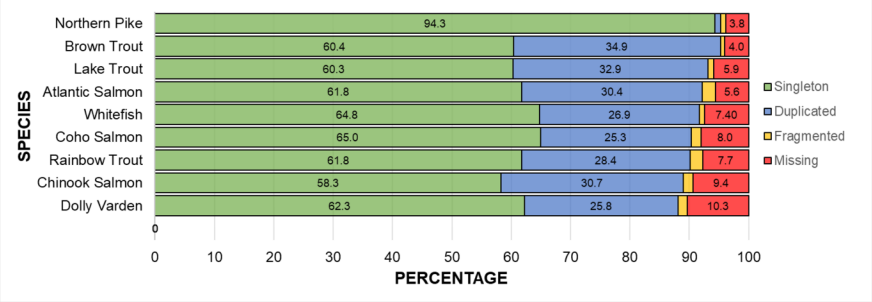
XIB. FIGURE 2 – THE LAKE TROUT GENOME



Circos plot displaying centromere positions, Tcl-Mariner abundance, density of annotated protein coding genes, male and female Lake Trout (*Salvelinus namaycush*) linkage maps and relationships between homeologs resulting from Ss4R. (A) Black boxes in the outside ring display the mean mapping positions (+/- 5 MB) for centromere associated RAD loci from Smith et al. (2020). (B) The second ring displays Z-transformed Tcl-Mariner repeat abundance in 5 megabase sliding windows with an offset of 100 kilobases. (C) The third ring displays the density of annotated genes in 5 megabase sliding windows with an offset of 100 kilobases. The fourth ring displays map distance (centimorgans) for male (red) and female (blue) linkage maps (y-axis) versus physical distance (x-axis) for each of the 42 chromosomes. Connections are drawn between syntenic blocks identified by SyMap v5 putatively resulting from the Salmonid specific autotetraploid event.

XIC. FIGURE 3 – BUSCO SCORES

Comparison of BUSCO scores across multiple chromosome-level salmonid assemblies. Scores for the pre-duplication outgroup species (Northern Pike; *Esox lucius*) are also included for comparison. Assemblies are listed top-to-bottom according to the total percentage of complete BUSCOs. Complete single-copy, complete duplicated, fragmented, and missing BUSCO percentages are delineated with green, blue, yellow, and red bars, respectively.



XII. TABLES

XIIA. TABLE 1 – ASSEMBLY STATISTICS

General summary statistics for the Lake Trout (*Salvelinus namaycush*) genome assembly. The total number of chromosomes, scaffolds (including chromosomes), and contigs are listed in the top row. Metrics reported for chromosomes and scaffolds include gaps of unknown length. Consensus accuracy was obtained from the output of POLCA after running three iterations of the program on contigs.

	Chromosomes	**Scaffolds**	**Contigs**	**Gaps**
Count	42	4,120	7,378	3,258
Minimum Length (bp)	22,041,605	9,606	84	100
Mean Length (bp)	47,175,710	569,295	317,859	100
Max Length (bp)	98,200,354	98,200,354	34,788,501	100
Total Length (bp)	1,981,379,816	2,345,496,355	2,345,170,555	325,800
N50 (bp)	48,336,861	44,976,251	1,804,090	100
N90 (bp)	34,530,387	249,999	114,532	100
N95 (bp)	26,015,404	84,453	61,568	100
Consensus Accuracy (%)	-	-	99.9959	-

XIIB. TABLE 2 – BUSCO SCORE COMPARISON

Total complete, complete single copy, complete duplicated, fragmented, and missing BUSCO percentages for 7 publicly available salmonid genomes, the Northern Pike (*Esox lucius*) genome, and the Lake Trout (*Salvelinus namaycush)* genome. Assemblies are ranked such that those with the highest percentage of complete BUSCOs are listed at the bottom. BUSCO scores were calculated using BUSCO v4.0.6 using the actinoptergii_odb10 database created on November 20th2019.

		Percent BUSCOs	**Percent BUSCOs**	**Percent BUSCOs**	**Percent BU**
Species	**Accession**	**Total Compete**	**Total Compete**	**Single Copy**	**Duplicated**
Dolly Varden	GCA_002910315.1	88.1	62.3	62.3	25.8

		Percent BUSCOs	Percent BUSCOs	Percent BUSCOs	Percent BU
Chinook Salmon	GCA_002872995.1	89.0	58.3	58.3	30.7
Rainbow Trout	GCA_002163505.1	90.2	61.8	61.8	28.4
Coho Salmon	GCA_002021735.1	90.3	65.0	65.0	25.3
European Whitefish	GCA_902810595.1	91.7	64.8	64.8	26.9
Atlantic Salmon	GCA_000233375.4	92.2	61.8	61.8	30.4
Lake Trout	GCA_016432855.1	93.2	60.3	60.3	32.9
Brown Trout	GCA_901001165.1	95.3	60.4	60.4	34.9
Northern Pike	GCA_000721915.3	95.3	94.3	94.3	1.0

XIIC. TABLE 3 - REPEATS

Number of elements, total sequence length, and percent of the Lake Trout (*Salvelinus namaycush*) genome occupied by retroelements, DNA transposons, and other repeat types.

Retroelements:		
	SINEs:	**SINEs:**
	Penelope:	**Penelope:**
	LINEs:	**LINEs:**
		CRE/SLACS
		L2/CR1/Rex
		R1/LOA/Jockey
		R2/R4/NeSL
		RTE/Bov-B
		L1/CIN4
	LTR Elements:	**LTR Elements:**
		BEL/Pao
		Ty1/Copia
		Gypsy/DIRS1
		Retroviral
DNA Transposons:		
		hobo-Activator
		Tc1-IS630-Pogo
		En-Spm
		MuDR-IS905
		PiggyBac
		Tourist/Harbinger
Other (Mirage, P-elements, Transib):	**Other (Mirage, P-elements, Transib):**	**Other (Mirage, P-elements,**
Rolling-Circles		
Unclassified:		
All Interspersed Repeats:	**All Interspersed Repeats:**	**All Interspersed Repeats:**

XIII. SUPPLEMENTAL MATERIALS

Supplemental Material 1 – GenomeScope Output and Plots

Supplemental Material 2 – Assembly Parameters

Supplemental Material 3 – Recombination Map

Supplemental Material 4 – Syntenic Blocks and Between Species Circos Plots

Supplemental Table 1 – BUSCO Comparison Between Species

Supplemental Table 2 – Contig and Scaffold N50 Comparison Between Salmonid Assemblies

Supplemental Table 3 – Centromere Locations