Population genomics of North American northern pike: variation and sex-specific signals from a chromosome-level, long read genome assembly

Hollie A Johnson[1*], Eric B Rondeau[1,2*], David R Minkley[1], Jong S Leong[1], Joanne Whitehead[1], Cody A Despins[1], Brent E Gowen[1], Brian J Collyard[3], Christopher M Whipps[4], John M Farrell[5], Ben F Koop[1§]

[1]Department of Biology, Centre for Biomedical Research, University of Victoria, Victoria, British Columbia, V8W 3N5, Canada

[2]Centre for Aquaculture and Environmental Research, Fisheries and Oceans Canada, 4160 Marine Dr., West Vancouver, British Columbia, V7V 1N6, Canada

[3]Alaska Department of Fish and Game, Division of Sport Fish, 1300 College Rd, Fairbanks, Alaska, 99701-1599, USA

[4]Center for Applied Microbiology, Department of Environmental and Forest Biology, SUNY College of Environmental Science and Forestry, Syracuse, New York, 13210, USA

[5]Thousand Island Biological Station, Department of Environmental and Forest Biology, SUNY College of Environmental Science and Forestry, Syracuse, New York, 13210, USA

[§]Corresponding author

*Authors contributed equally to results of manuscript

Email addresses:

HAJ: holliej@uvic.ca

EBR: erondeau@uvic.ca

DRM: dminkley@uvic.ca

JSL: jong@uvic.ca

JW: joanne.whitehead@bcchr.ca

CAD: cdespins@uvic.ca

BEG: bgowen@uvic.ca

BJC: brian.collyard@alaska.gov

CMW: cwhipps@esf.edu

JMF: jmfarrell@esf.edu

BFK: bkoop@uvic.ca

Abstract

We present a chromosome-level, long-read genome assembly as a reference for northern pike (*Esox lucius*) where 97.5% of the genome is chromosome-anchored and N50 falls at 37.5 Mb. Whole-genome resequencing was genotyped using this assembly for 47 northern pike representing six North American populations from Alaska to New Jersey. We discovered that a disproportionate frequency of genetic polymorphism exists among populations east and west of the North American Continental Divide (NACD), indicating reproductive isolation across this barrier. Genome-wide analysis of heterozygous SNP density revealed a remarkable lack of genetic variation with 1 polymorphic site every 6.3kb in the Yukon River drainage and one every 16.5kb east of the NACD. Observed heterozygosity (H_o), nucleotide diversity (π), and Tajima's D are depressed in populations east of the NACD (east vs. west: H_o: 0.092 vs 0.31; π: 0.092 vs 0.28; Tajima's D: -1.61 vs -0.47). We confirm the presence of the master sex determining (MSD) gene, *amhby*, in the Yukon River drainage and in an invasive population in British Columbia and confirm its absence in populations east of the NACD. We also describe an Alaskan population where *amhby* is present but not associated with male gender determination. Our results support that northern pike originally colonized North America through Beringia, that Alaska provided an unglaciated refugium for northern pike during the last ice age, and southeast of the NACD was colonized by a small founding population(s) that lost *amhby*.

Keywords

Northern pike, *Esox lucius*, Resequencing, Population Genomics, Long-Read Assembly, Genetic Variation

Introduction

The northern pike (*Esox lucius*) belongs to the small order Esociformes (pikes and pickerels) and is the most studied species of the genus *Esox* (Forsman et al., 2015; Nelson, 2006; Skov & Nilsson, 2018). They inhabit fresh and brackish water and have a widespread

distribution across much of the northern hemisphere (Craig, 2008). As voracious apex predators, they influence species assemblage and are able to colonize new habitats with just a few individuals. Prized in sport fishing, they are highly valued in Canadian sport fisheries (Government of Canada, 2016). Used extensively in physiological, toxicological and ecological studies, it has been proposed that northern pike is approaching the status of a model organism in ecology and evolution (Forsman et al., 2015). Genetic resources available include microsatellite markers, full mitochondrial sequences, expressed sequence tags, a preliminary reference genome, and now, a chromosome-level genome.

Genetic variation is pivotal to the ability of a species to adapt to environmental change over space and time. Allelic diversity provides the needed genetic resources to increase potential for species adaptation when exposed to selective pressures (Barrett & Schluter, 2008; Höglund, 2009), thereby allowing populations to survive environmentally challenging conditions or to colonize new habitats. Northern pike is a three to five million year-old species (Grande, 1999) that often finds success in colonization (either natural or introduced). However, from the earliest studies of allozymes, microsatellites, and mitochondrial sequences, to the first version of the northern pike genome in 2014 , low levels of genetic variation have been encountered (Bosworth & Farrell, 2006; Miller & Kapuscinski, 1996, 1997; Rondeau et al., 2014; Senanan & Kapuscinski, 2000; Skov & Nilsson, 2018). Explanations for these low levels of variation include population bottlenecks created during the previous ice age (due to limited refugia), small effective population sizes, and northern pike's ecological role as an apex predator and cannibalistic tendencies (Seeb et al., 1987). Here, we investigate the nature of genetic variation in northern pike with the most comprehensive markers to date by analyzing single nucleotide

polymorphisms (SNPs) generated from whole-genome resequencing data of 47 individuals from across North America.

Questions remain regarding northern pike's colonization of North America and population structure therein. Because of such low levels of variation in northern pike, it has been difficult to define population structure within North America, and the origin and number of the founding populations is unclear (Miller & Senanan, 2003; Senanan & Kapuscinski, 2000; Skog et al., 2014). Hints of genetic distinctions have been observed between northern pike from Alaska (Yukon River drainage) and eastern North America (Hudson's Bay, St. Lawrence, and Mississippi drainages) based on microsatellite and mitochondrial data (Senanan & Kapuscinski, 2000; Skog et al., 2014). The potential for the northern pike to colonize Alaska via Beringia has been recognized (Crossman & Harington, 1970). However, the discovery of an Esocid fossil in central North America dating to the Paleocene (56 – 66 mya) that was more pike-like than pickerel-like led to the suggestion that northern pike in North America may be survivors from this ancient relic (Wilson, 1980). Here, using genome-wide SNPs, we are able to outline population structure within North America and we attempt to clarify northern pike's colonization of North America.

Yet another remaining question is the nature of the sex determination system in North American northern pike. Sex determination is the cue that initiates development towards a male or female phenotype. In contrast to birds and mammals, factors determining sex are diverse in fishes as both genetics and the environment have been shown to exert control, and are not mutually exclusive (Devlin & Nagahama, 2002; Goto-Kazeto et al., 2006). Genetic factors that determine sex are different among and even within fish species. Several master sex determining (MSD) genes have been identified in fish. These genes have been shown to be necessary for the

development of testes or ovaries based on knockout and transgenic experiments; e.g., *dmY* in *Oryzias latipes* and *amhy* in *Odontesthes hatcheri* (Hattori et al., 2012; Kamiya et al., 2012; M. Li et al., 2015; Matsuda et al., 2007; Myosho et al., 2012; Yano et al., 2012). There are other candidate MSD genes that show perfect association with one sex, but have not been proven through knockout and transgenic experiments; e.g. *gsdf* in *Anoplopoma fimbria* (Baroiller et al., 2009; Feron et al., 2019; Kawase et al., 2018; Rondeau et al., 2013; Yano et al., 2013). Although the genes controlling gonadal fate are different in many of the species examined so far, almost all are familiar players in gonadal development, and many have links to the TGF-ß signaling pathway (reviewed in (Devlin & Nagahama, 2002; Kikuchi & Hamaguchi, 2013; Matsuda, 2018; Pandian, 2011)).

A male-specific duplicated gene copy of the anti-Mullerian hormone, *amhby*, has been identified as the MSD gene responsible for male differentiation in northern pike from Europe (Pan et al., 2019). However, previous studies of North American pike (Rondeau et al., 2014) had found no markers associated with sex and early discussions with the authors of Pan et al., (2020) suggested that *amhby* was absent in North America. Pan et al., (2020) did not detect *amhby* in North American populations outside of Alaska using non-individual based (Pool-Seq) and reduced-representation (Rad-Seq) sequencing approaches. Additionally, field and laboratory observations of skewed sex ratios in some North American populations suggest than there may be an environmental influence on sex determination (Carbine, 1942; Clark, 1950; Huffman et al., 2014; Priegel & Krohn, 1975). Northern pike are considered circumpolar and known as a single species (Grande et al., 2004), so the inconsistencies observed in sex determination mechanisms between European and North American lineages are intriguing, and worth detailed investigation.

Using high-resolution resequencing data from male and female North American northern pike, we shed additional light on the loss of *amhby*.

The steady generation of genome assemblies in teleosts is a consequence of rapid advances in DNA sequencing technologies and decreasing cost of generating data. Short-read technologies, in particular, Illumina-based protocols and instruments (Bentley et al., 2008), have revolutionized our understanding of genome-wide genetic variation, but genome assembly contiguity remains a challenge. While short-reads are sufficient within unique regions of the genome, they can fail to assemble correctly in highly variable or repetitive regions, leaving the genome in heavily fragmented pieces; resolving structural variation in such fragmented regions is extremely difficult. Sequencing increasing fragment sizes allows for bridging and ordering of contigs into scaffolds with gaps of estimated sizes. But this strategy often fails to fill in the gaps between contigs, and insert size can often be insufficient to resolve long repetitive elements. Though the majority of coding sequences can be mapped to these new assemblies, new, longer sequencing technologies are required to fully characterize genes, repeat regions and chromosomal structural elements (reviewed in (Chaisson et al., 2015)).

New technologies such as Pacific Bioscience's SMRT (Eid et al., 2009) or ONT's single-molecule nanopore sequencing (Stoddart et al., 2009) allow a slightly different approach to de novo genome assembly. While the higher error rate of both technologies requires a certain level of error-correction with short reads (e.g. (Goodwin et al., 2015; Koren et al., 2012)) or through hierarchical approaches that rely on consensus of multiple long reads (Chin et al., 2013), long-reads produced by these technologies have a significant advantage in that they allow for direct characterization through all but the longest repeat regions while also filling gaps between contiguous sequences. These technologies promise improvements in genomes where repetitive

content is high, and in polyploid species where identifying differences between duplicated copies may require increased length to identify subtle changes between duplicated copies.

While these long-read technologies may increase the contiguity of genome assemblies, some degree of higher-level scaffolding will continue to be required to tackle particularly difficult or highly repetitive genomic regions. New strategies that provide higher orders of connecting scaffolds include chromatin conformation capture (Dekker et al., 2002) such as the in-vitro chromatin proximity ligation technique implemented by the Chicago/HiRise method from Dovetail Genomics (Putnam et al., 2016), Hi-C based proximity ligation techniques (Dudchenko et al., 2017), or the gemcode-based technologies from 10X that partially sequence by short read technologies millions of uniquely barcoded pools of small subsets of the genome and provide linking data that can be used to generate super-scaffolds and long-range haplotypes (Mostovoy et al., 2016).

In this work, we describe improvements to the original reference northern pike genome version 1.0 (v1.0) and present a highly contiguous, de novo assembly in version 4.0 (v4.0) based on PacBio Sequel long reads and scaffolding with 10X chromium and Hi-C libraries. We demonstrate the utility v4.0 through a population genomics analysis that details the nature of genetic variation in North American northern pike, helps clarify population structure and colonization of North America, and sheds light on the intra-specific loss of a sex determining gene.

Materials and Method

Sequencing and long-read assemblies – release 4.0

All sequencing was performed on a new individual, separate from v1.0 - v3.0 (Release 1.0 described in (Rondeau et al., 2014); Release 2.0 and 3.0 described in Supplementary File 1). In September 2017, pike were caught as part of an invasive species removal project in the Canadian portion of the Columbia river, near Castlegar, BC. Following euthanization, the liver tissue from a single, female specimen was removed by dissection and frozen on dry ice for 48 hours before long-term storage at -80°C. (NCBI Biosample: SAMN09690694).

High-molecular weight DNA was extracted from the liver using a modified Dialysis method. 550mg of tissue was ground into a powder using liquid nitrogen and mortar and pestle. The powder was transferred to a 5ml lo-bind Eppendorf tube, along with 3600ul buffer ATL (Qiagen), 400ul proteinase K solution (Qiagen) and 40ul RNAseA solution (Qiagen), followed by digestion at 56 ^{o}C for 3 hours, with rotation at ~4 rpm. This was split equally into two 5ml tubes, where a phenol-chlorform-isoamyl alcohol (25:24:1) purification was performed 3 times, followed by 1 round of chloroform-isoamyl alcohol (24:1). In each stage, 1 volume of organic was mixed with 1 volume of aqueous, inverted slowly for 3 minutes to mix thoroughly, spun for 15 minutes at 5000xg to separate the layers, and aqueous top layer transferred very slowly to a new tube using a 1000ul wide bore tip. 2ul of RNAseA solution (20mg/ml Qiagen) was added and incubated at room temperature for 1hr, followed by 5ul proteinase K (20mg/ml) overnight at 4 ^{o}C. Approximately 750 ul was obtained from each tube and transferred to a Spectra/Por Float-A-Lyzer G2 1000 kD (pink) dialysis device. Dialysis was performed in 1 liter of 10mM Tris-Cl, pH 8.5 at 4 ^{o}C with gentle mixing for one week, changing buffer five times. DNA quantity was

determined by Qubit v2.0 (Life Technologies) and quality by 0.6% agarose gel at 60Volts. Bands greatly exceeded the largest ladder band of 40kb with no visible shearing.

Subsequent library preparation and sequencing steps were performed by the McGill University and Genome Quebec Innovation Centre. PacBio sheared large-insert libraries were constructed following standard protocols, and sequenced across 8 SMRT cells on a PacBio Sequel, generating 76 Gbp of data. A library was constructed for 10X chromium sequencing following standard protocols, and sequenced on 1x lane of Illumina HiSeqX PE150. A library was constructed in-house using the Phase Genomics Proximo Animal Hi-C kit following protocol Phase Genomics protocol 1.0 (adapter barcode N702) and utilizing 0.2g of the aforementioned liver tissue and shipped with remaining samples for sequencing with a single lane of Illumina HiSeq4000 PE100.

PacBio data was assembled using Canu v1.8 (Koren et al., 2017). All subreads were used, and a genome size of 0.95 Gbp was estimated as input. All stages were run with SLURM scheduling on the Compute Canada heterogeneous cluster Cedar (https://docs.computecanada.ca/wiki/Cedar). The option "stageDirectory=\$SLURM_TMPDIR/\$SLURM_JOBID" and gridEngineStageOption="--tmp=150g"was utilized to take advantage off on-node storage during certain heavy I:O stages. Otherwise, default settings were used other than "ovlMerThreshold=2000 corMhapSensitivity=normal correctedErrorRate=0.085 minReadLength=2500" to either reduce runtime, or to use recommendations for Sequel data discussed in the software manual. Following initial assembly, Arrow v 2.2.2 (https://github.com/PacificBiosciences/GenomicConsensus) was utilized in SMRTlink 6.0.0.47841 to polish with PacBio data, using the ArrowGrid wrapper (Koren et al., 2017). The pipeline was run a total of three times, following which it was input

into Pilon v1.22 (Walker et al., 2014) for one round of polishing, using the 10X chromium Illumina-based data as input that had first been processed using Longranger 2.2.2 (10X Genomics) to remove the barcodes.

Scaffolding occurred in three stages. In the first, SALSA2 (github commit 68a65b4) (Ghurye et al., 2017, 2019) was used to scaffold the assembly using Hi-C data. As recommended in the documentation, data was first prepared and aligned following recommendations by Arima Genomics (https://github.com/ArimaGenomics/mapping_pipeline commit: 2e74ea4). R1 and R2 were aligned to the genome using BWA mem 0.7.13-r1126 (H. Li, 2013) separately, followed by sorting, merging and filtering with Samtools 1.8 (H. Li et al., 2009), Picard 2.9.0-1-gf5b9f50-SNAPSHOT (http://broadinstitute.github.io/picard/), Bedtools 2.27.0 (Quinlan & Hall, 2010). SALSA2 was run with "-m yes -e GATC", the post-Pilon reference and the bed file output by the Arima pipeline. In the second scaffolding stage, the Tigmint – arcs -links pipeline was utilized from BC Genome Sciences Centre. Tigmint v1.1.2 (Jackman et al., 2018) was run using the "arcs" pipeline to run all three stages. The Tigmint portion of the pipeline was run with default parameters. Within Arcs v1.0.5 (Yeo et al., 2018) and LINKS v1.8.6 (Warren et al., 2015) portions of the pipeline, parameters were tested for all combinations of l=5-10, a=0.1-0.9. Parameters were optimized for increasing N50, while balancing the number of misjoined scaffolds observed in the third stage of scaffolding with Hi-C. Final parameters reflecting an optimal balance (max 2-3 visible misjoins) were a=0.2 and l=8 with all other parameters remaining default. As gap sizes were no longer meaningful at this stage, and Tigmint introduced some very short fragments, sed was used to remove contigs smaller than 200bp nested within scaffolds, and resize all remaining gaps to 100 Ns. Bioawk fastx (https://github.com/lh3/bioawk) was used to remove remaining scaffolds smaller 1000bp or less (<100 total scaffolds removed).

In the third stage of scaffolding, Hi-C data was aligned to the genome post-Tigmint pipeline using Juicer 1.5.6 (Durand, Shamim, et al., 2016). Parameters used were "-s Sau3AI and -S early" while also including Sau3AI cutsite file with "-y". Resulting "merged_nodups.txt" was input into 3d-dna v. 180922 (Dudchenko et al., 2017) with "-i 50000 -r 0". Juicebox v1.8.8 (Durand, Robinson, et al., 2016) was used to visualize the assembly post-scaffold, where mis-assemblies were identified and split, and linkage groups were identified, ordered and oriented. Linkage group-like groups were oriented such that the greatest density of inter-chromosomal contacts were at the beginning of the linkage group. The Phase Genomics juicebox_assembly_converter.py script (https://github.com/phasegenomics/juicebox_scripts/; commit 7692ad5) was used to generate the NCBI AGP files, and after manual editing to rename linkage groups based on prior assemblies (identified through LastZ (Harris, 2007) alignments in Geneious (Kearse et al., 2012) to prior genome assembly using default parameters) and linkage map (Rondeau et al., 2014) , genome was submitted to NCBI. These scaffold sequences have been uploaded to NCBI under BioProject ID PRJNA221548, accession SAXP0000000; this version is the first version, SAXP00000000.1. It is this version, designated as Eluc_v4, that represents the most recent RefSeq accession for this species, under GCF_004634155.1, and was annotated by the NCBI Eukaryotic Annotation pipeline v8.2 under annotation release 103. Genome completeness was evaluated for genome release 3.0 (Supplementary File 1) and 4.0 via Busco v4.0.2 utilizing both Actinopterygii and Vetebrata odb10 datasets (Seppey et al., 2019). Genome alignments between version 3.0 and 4.0 were conducted using symap v4.2 (Soderlund et al., 2011) using default settings. Repeat analysis utilized the same methods and custom repeat library described in Rondeau et al., 2014.

Population Genomics

Samples

Tissue samples from 47 Northern Pike from across Canada and the northern United States were provided by collaborators and hatcheries as per Table 1 and Figure 1. Samples originated from Chatanika River, Alaska (N=10); Yukon River, Hootalinqua, Yukon Territory (N=5); Palmer Lake, British Columbia (N=4); Charlie Lake, British Columbia (N=1), Columbia River, Castlegar, British Columbia (N=1); Whiteshell Hatchery, Manitoba (N=6); St. Lawrence River, New York (N=11); and Hackettstown Hatchery, Hackettstown, New Jersey (N=9). Fish from the St. Lawrence River were collected and processed following a protocol approved by the SUNY ESF Institutional Animal Care and Use Committee. Other tissues were either archival, opportunistic sampling of fishery harvest or from fish harvested by government agencies for their own purposes (e.g. invasive species control), ethical review was not required by the University of Victoria in accordance with the Canadian Council on Animal Care Guidelines on: the care and use of fish in research, teaching and testing, 2005 4.1.2.2

DNA Extraction and Sequencing

DNA was extracted from a variety of tissues using DNEasy Blood and Tissue Kit (QIAGEN) following the manufacturer's protocols. Extracted DNA was quantified by Nanodrop ND-1000 (Thermo) and Qubit v2.0 (Life Technologies). Samples were sent for sequencing to McGill University and Genome Quebec Innovation Centre, where 35 of the 47 samples underwent PCR-free whole genome shotgun sequencing. Ten samples (the Chatanika River population) were sequenced via PCR shotgun sequencing because the amount of DNA extracted was insufficient for PCR-free libraries. All libraries were sequenced on an Illumina HiSeqX Ten

PE150 with the exception of the two individuals used to build reference genomes v1.0-v3.0 and v4.0 which were sequenced as described above. Samples were pooled such that 5 – 7 samples were sequenced per lane. Lanes were designated for male or female samples exclusively as much as possible in order to reduce the possibility of index switching between sexes.

Alignment and SNP discovery

Read processing and variant calling was based on GATK's best practices, and for all steps involving GATK we used version 3.8-0-ge9d806836 (DePristo et al., 2011; McKenna et al., 2010; Poplin et al., 2018; Van der Auwera et al., 2013). Paired end reads were aligned to the northern pike genome v4.0 (WGS accession SAXP00000000.1, assembly accession GCF_004634155.1) using the Burrows-Wheeler Aligner (BWA) version 0.7.13-r1126) with the "-mem" algorithm (H. Li, 2013). Alignment files were piped to SAMtools version 1.3 (H. Li et al., 2009), converted to binary alignment/map (BAM) format, then sorted and indexed according to position. Information detailing the sequencing platform and multiplexing layout was incorporated and used to mark duplicates with Picard version 2.17.11 (Broad Institute, 2017). Because the reference individuals had read depths 5 – 7 times greater than our samples, we down-sampled their BAM files using the Samtools "view -s" command to obtain files that had similar depth to the rest of our samples. This was necessary to ensure downstream filters worked appropriately and to manage analysis time. Bases in all BAM files were re-calibrated according to GATK's recommendations for non-model organisms.

Variants were called from re-calibrated BAM files independently for each sample using GATK's HaplotypeCaller in GVCF mode and combined as a cohort via GATK's GenotypeGVCF command to produce one VCF file containing 1,910,789 SNPs and InDels for all 47 samples.

SNPs (1,363,731) were extracted and filtered according to the parameters in Table 2. Through GATK we applied a hard filter that removed variants according to the following parameters (thresholds in brackets): quality by depth (2), fisher strand bias (60), root mean square mapping quality (30), mapping quality rank sum test (-12.5), and read position rank sum test (-8.0). We applied further quality control filters through VCFtools version 0.1.15 (Table 2) (Danecek et al., 2011). We removed sites where more than 10 individuals were missing calls in order to ensure our analyses represented the majority of individuals (--max-missing-count 10). We applied a minor allele frequency filter of 1 (--mac 1) to filter sites that were a combination of only homozygous alternate alleles and missing calls. Finally, we applied a filter that required at least one of the calls to be homozygous (variant or reference). The VCF file produced after the last filtration step was the central file for our analyses. Any further filtration steps specific to particular analysis are discussed in their corresponding section.

Using the same methods, we aligned resequencing data to v3.0 and called and filtered SNPs for the purpose of comparing the output to v4.0.

Heterozygosity and variation analysis

The number of variant calls per individual was calculated with Real Time Genomics' RTG Tools version 3.9.1 (*RTG Tools*, 2015/2018) and visualized in R (R Core Team, 2019). Genotype counts per site were obtained through GATK's "VariantsToTable" command for the following call categories: heterozygous, homozygous reference, homozygous variant, no call, total variants called, number of samples called. To obtain per site genotype frequencies, each category was divided by the number of samples called. Our values of observed heterozygosity (H_o) were obtained from this analysis. Using VCFtools, we obtained nucleotide diversity values

per site and Tajima's D in bin sizes of 10,000 (Danecek et al., 2011). This was done for all individuals together and in groups defined by PCA/DAPC analysis.

Phylogeny

A maximum likelihood tree based on genome wide SNPs was generated using SNPhylo v. 20140701 (Lee et al., 2014). Default parameters were used, except to specify to perform 1000 bootstraps. The resulting tree was visualized through Figtree version 1.4.3 (Rambaut, 2016), and rooted by midpoint.

Discriminant analysis of principal components

A DAPC was performed with bi-allelic genome wide SNPs using the R software package Adegenet v. 2.1.1 (Jombart, 2008; Jombart & Ahmed, 2011). Adegenet's "find.clusters" function grouped our samples into 4 clusters based on the lowest Bayesian Information Criterion value when all principal components were kept. We performed the DAPC on the groups identified by the "find.clusters function", and retained 24 principal components and all three discriminant functions. We then used the "snpzip" command with the Ward clustering method to return lists of SNPs that had the greatest contribution to each of the three discriminant axes identified.

Sex Specific Kmer Analysis

This analysis was performed on resequenced populations (Table 1) for which there were a minimum of three males and three females (Chatanika River, Manitoba, New Jersey, New York). Raw reads were concatenated then all possible 31-mers extracted using Jellyfish v2.2.6 (Marçais & Kingsford, 2011) running on Compute Canada, to create a master list of kmers. This master list was used to query each individual's reads with Jellyfish, to create a table of 31mer

counts. Comparing males and females within each population, male-specific kmers were defined as those sequences for which all females had 2 or fewer copies (to allow for sequencing errors), and for which all males had a positive count and the sum of all male counts was >7n; vice versa for female-specific kmers. The resulting sex- and population-specific kmer sets were mapped back to the reference genome using bwa-aln v0.7.13-r1226 (H. Li, 2013). The number of sex-specific kmers in 10kb windows across the genome were counted for each population using bedtools v2.26.0 (Quinlan & Hall, 2010), and graphed in R v3.5.3 (R Core Team, 2019) using ggplot2 (Wickham, 2016).

Genome-Wide Association Study

Using VCF Tools, a filter of --mac 2 was applied to the VCF file and sexed individuals were extracted (21 males and 17 females). A genome-wide association study (GWAS) based on sex was performed using plink v. 1.9b_5.2-x86_64 (Purcell et al., 2007) with the "fisher-midp" option. The -log(p-value) of each SNP was visualized on a manhattan plot created with the R software qqman v. 0.1.4 (Turner, 2018). Significance was assessed via Bonferonni correction at □= 0.05.

Sex-Specific DAPC

We performed a DAPC based on sex among all resequenced northern pike and within populations and defined groups (from PCA/DAPC) with the R software Adegenet (Jombart, 2008; Jombart & Ahmed, 2011). Population and group specific SNPs were extracted for the VCF file and each subset of SNPs was filtered independently using a custom script in R such that any SNP site that had a homozygous alternate allele was removed. This left behind SNP sites that only contained homozygous reference and heterozygous alleles. DAPC was performed on each

subset. We queried the resulting loadings tables for values that were specific to all-male or all-female heterozygosity, and compiled lists of genomic locations where sex-specific SNPs reside in each population. In order to identify regions of dense, sex-specific heterozygosity, we made histograms of sex-specific SNPs on each linkage group for each analysis using ggplot2 (Wickham, 2016).

Sex-specific PCR assays

In-house primers (set 24.5) were manually designed such that both sexes would produce a 500 base amplicon, and males would produce an additional 250 – 300 base amplicon. This smaller amplicon was generated by a nested primer whose sequence was based on male-specific SNPs. We used primers designed by (Pan et al., 2020) to amplify regions of *amhby* (sets SeqAMH1-4 and ConserveAMH1-1). Primer sequences and PCR conditions are detailed in Supplementary File 1. We queried our resequenced individuals with these primer sets and expanded the assay to additional samples that were not used in sequencing, including 2 females and 5 males from Castlegar, BC, and 14 males, 6 females and one fish of undetermined sex from the Minto Flat region of Alaska, USA. Pike from Castlegar were collected and preserved as described above. Fin clips were collected from Minto Flat pike by the Alaska Department of Fish and Game following approved state and departmental regulations and protocols and preserved in 95% ethanol.

Gonadal Tissue Histology

Northern pike gonadal tissue from the Minto Flat population in Alaska were examined histologically at the University of Victoria (Victoria, BC) to confirm phenotypic sex. Samples were saturated with a 1:1 solution of 100% ethanol:LR White resin (hard grade) for 24 hours,

then with pure LR White for 24 hours (resin replaced fresh after 6 hours). Each sample was then placed in separate gelatin capsules with fresh catalyzed LR white and polymerized at 60˚C for 24 hours. Sections were cut (1 micron), stained with Richardson's LM stain, and examined microscopically.

Tissues collected from northern pike in the St. Lawrence river were also processed for histological examination. Gonadal tissues were removed from fish and a piece preserved in Davidson's solution for 48 hours, then transferred to 70% ethanol prior to histological processing. Following processing, tissues were embedded in paraffin wax and 5 micron sections were cut and stained with hematoxylin and eosin (H&E).

Results

Genome Assembly

Rather than continue with incremental improvements on the prior genome (see Additional file X for more information on incremental improvements to the prior assembly), a new approach was taken in a *de novo* assembly. A single female pike, from an invasive population in the Canadian portion of the Columbia River, was sequenced to ~80X depth using 8 SMRT cells on a Pacific Biosciences Sequel instrument. Assembly with Canu (Koren et al., 2017) yielded a primary assembly with total length of 939.0 Mbp, with a contig N50 of 3.94 Mbp and 1,258 total contigs.

Following Arrow (Chin et al., 2013) and Pilon (Walker et al., 2014) polishing, scaffolding occurred across three stages. First, Hi-C sequencing data was utilized with the SALSA2 pipeline (Ghurye et al., 2019) (Supplementary Figure 1) to generate an assembly with

scaffold N50 of 18.8 Mbp across a total of 941 scaffolds. Second, 10X Chromium data was introduced with the Tigmint/ARCS pipeline (Jackman et al., 2018; Yeo et al., 2018), to both break suspect contigs and to further scaffold the genome. Error correction reduced contig N50 to 3.40 Mbp (1,395 contigs) while scaffold N50 increased to 23.3 Mbp. Finally, 3d-DNA (Dudchenko et al., 2017) was utilized to further scaffold using the Hi-C data followed by manual review using Juicebox (Durand, Robinson, et al., 2016; Robinson et al., 2018). This final stage of scaffolding yielded the final scaffold N50 of 37.550 Mbp across 811 total scaffolds, including 25 of chromosome length (Table 3). Chromosome-length scaffolds were assigned to linkage groups using the same linkage map utilized in prior assemblies (Rondeau et al., 2014). As linkage groups were oriented by density of inter-chromosomal contacts, repeat masking (using methods and repeat library in (Rondeau et al., 2014) - results summarized in Supplementary Table 1) reveals the greatest density of repeat elements at the 3' end of the chromosome (Supplementary Figure 2 - Repeat density). Both density of inter-chromosomal contacts and repeat density appear to support 3' orientation of the centromere in the 25 acrocentric chromosomes. The mitochondrial genome was identified by BLAST alignment of NC_025992.1 against the updated genome, with circularization performed manually. This version of the genome was submitted to NCBI through WGS accession SAXP01000000.1 and assembly accession GCA_004634155.1, with RefSeq assembly accession GCF_004634155.1 curating the assembly in the Genome database.

Annotation of genomes described in this work was performed by the NCBI eukaryotic genome annotation pipeline using previously utilized (Rondeau et al., 2014) and recently released (Pan et al., 2019) RNA-seq and EST data (Leong et al., 2010). In version 4, this encompasses 24,843 protein-coding genes across the genome. Busco analyses using both

Actinopterygii and Vertebrate datasets were performed, with results in Table 4. Both version 3 (hybrid) and version 4 (long-read) demonstrate near-identical counts of complete and missing genes, implying both assemblies capture and resolve the majority of genic regions. Chromosome-level alignments do demonstrate some re-orientation along individual regions within chromosomes (Supplementary Figure 3); however, this is most likely due to how scaffolds were assigned and oriented to chromosomes, with version 3 placing relatively low contiguity scaffolds using a low-density genetic map and conserved synteny, while version 4 began with a large scaffolds in a highly-contiguous assembly and relied on HiC chromosomal intra-chromosomal contacts. This is further supported by increased re/mis-orientation at the ends of the chromosomes in V3.0 where repetitive content is higher thus leading to increased assembly difficulties. It can also be observed the increase in assignment to chromosomes occurs primarily on the ends of the chromosome, implying increased contiguity from the long-read assembly and the utilization of Hi-C allows greatest improvement in chromosome representation in regions of high-repetitiveness.

Genetic Variation and Population Genomics

We obtained whole genome resequencing data from northern pike collected across North America (Figure 1 ; Table 1) and is available through NCBI under BioProject ID PRJNA512507. The greater contiguity and repeat resolution of the northern pike reference genome v4.0 allowed resequencing data to be aligned with greater accuracy, resulting in increased mapping quality as compared to alignment against v3.0 (Table 2). This is illustrated by the number of raw SNPs called by GATK and the number remaining after pruning. With the same resequencing data and the same filtering parameters (see methods), 1,277,574 SNPs were pruned when called against v3.0 of the genome, leaving 944,574 SNPs for analysis. Just 235,788 SNPs were pruned when

data was aligned to v4.0, leaving 1,127,943 SNPs for analysis. This was our final variant set. The mean depth per individual at variant sites ranged between 15 - 30, with an average of 20 and a standard deviation of 3.3.

Across our sampling range, variation in northern pike was harbored differently east and west of the NACD. Populations along the Yukon River drainage basin held two to three times more polymorphic sites than populations east of the NACD (Figure 2). This represents an average of one heterozygous SNP every 6,250 bases in Yukon River drainage populations, and one every 16,500 bases in all populations east of the NACD. Exceptions to this trend are Palmer Lake and the individual from Castlegar, BC. Northern pike from Palmer Lake cluster with western populations but have similar numbers of polymorphic sites (mean = 58,337) as populations east of the NACD (Figure 2). Castlegar, BC, is located west of the NACD. The individual collected here had the fewest number of heterozygous SNPs of all individuals - just 14,565 across the genome.

Genetic variation in northern pike is not evenly distributed across the genome. Rather, long swaths of low SNP density are separated by peaks of high SNP density of up to 400 SNPs per 10kb (Figure 3). The most extensive regions of elevated SNP densities contain genes related to immunity and multiple copy number variants. For example, on linkage group 9 between 5 and 10 MB is a region of elevated SNP density which contains sections rich with zinc-finger genes, immune-associated nucleotide-binding protein genes (GIMAPs), as well as multiple genes of the major histocompatibility complex (MHC) class II type. It is beyond the scope of this paper to describe all of the regions of the northern pike genome with elevated SNP density, but the four most extended regions of increased SNP density are highlighted in Figure 3 and all contain immune-related genes (Table 5).

Clustering analysis and cladogram construction show that the greatest genetic distance lies between western and eastern northern pike populations separated by the NACD (Figure 4A). Bootstrap values between 90-100% delineate most clades with the exception of the Hootalinqua (Yukon River) and Chatanika River populations which are linked by one individual, indicating genetic continuity exists between these two populations. This may reflect migration between these two locations, as may be expected as they are part of the same drainage system. In DAPC analysis, populations east of the NACD cluster very tightly together, reflective of the reduced number of heterozygous SNPs and depressed variation statistics (Figure 4B, Table 6). While DAPC places populations from this vast geographical range into one group, populations west of the divide are distinct enough to comprise their own clusters. Again, the exception to this is the Castlegar pike, which groups tightly with others from eastern North America and belongs to a population that was translocated from east of the NACD . The majority of the variation in the data is accounted for by the first discriminant axis in the DAPC. This is attributable to 2,115 genome-wide SNPs that are fixed at alternate alleles in populations east and west of the NACD (Figure 4C).

Genome-wide analysis of observed heterozygosity (H_o), nucleotide diversity (π), and Tajima's D among and within populations revealed that these indices are depressed in populations east of the NACD (Table 6). An exception being that the lowest H_o is obtained when all pike are analyzed together and this is likely due to the increased number of unique SNP sites and greater sample size when all populations are pooled.

Sex Determination

After filtration and extraction of sexed individuals, 672,565 SNPs across 21 males and 17 females remained for GWAS. No significant or even suggestive signals were detected, indicating

that the described northern pike MSD gene, *amhby*, was not ubiquitous across North America (Supplementary Figure 4). This result was corroborated by sex-specific DAPC analysis and k-mer analysis, which also failed to identify sex-specific polymorphism in northern pike across North America on a continental scale. K-mer analysis is performed on raw sequence reads, therefore eliminating the possibility that failure to detect sex-specific signals could be due to the possibility of their absence from the reference genome.

In alignment with Pan et al., 2020, , we were able to detect a sex-specific signal within our population from Chatanika River, Alaska, which was comprised of 5 females and 5 males, sexed by visual observation of ovaries (female) or testes (male). In agreement with (Pan et al., 2019), we detected male-specific heterozygosity on linkage group 24 through both DAPC and k-mer analysis (Figure 5A, 5E). This signal consisted of 3,552 male specific SNPs spanning a 500 kb region between 650 - 1150 kb on linkage group 24. An additional 1,137 male-specific and 39 female-specific SNPs were identified through DAPC analysis, and were located in various regions across the genome (Supplementary Table 2). This signal on LG24 and the additional male-specific SNPs were also present in one fish from the Yukon River, but were absent from all other sexed populations (Figure 5B, 5C, and 5D). Because the northern pike reference individual is female (v4.0), male-specific sequence is not present and therefore we did not observe *amhby* directly in our resequencing data. To detect *amhby*, we used *amhby*-specific primers designed by (Pan et al., 2020) and confirmed that *amhby* was present in males from Chatanika River, Alaska. Using these primers, we detected the presence of *amhby* in one of our pike from Hootalinqua, Yukon River. However, pike from this population were not phenotypically sexed so association with phenotype could not be confirmed. These primers failed to produce a band in all northern

pike east of the NACD, corroborating results from our continent-wide GWAS and sex-specific DAPC. Results are summarized in Table 7.

We tested additional samples (not used in resequencing analysis) for the presence of *amhby* using the three primer sets discussed above. Five additional northern pike from Castlegar, B.C. (2 female, 3 male) were screened. *Amhby* was present in all three males and absent in all females. This is an intriguing observation because this population is invasive, and the source population is from east of the NACD. Unexpectedly, PCR assays on individuals from the Minto Flat region (Alaska) did not confirm a consistent association of *amhby* with maleness. While all females were *amhby*-negative, only 5 of the 14 males tested positive for *amhby*. As such, we found *amhby*-negative and *amhby*-positive males. Histological examination of gonadal tissue from *amhby*-positive and *amhby*-negative phenotypic males confirms that the development of testes is possible without the presence of *amhby* (Figure 6). This observation shows that *amhby* does not possess the sole sex-determining role in this population.

Discussion

Genome assembly

Even in the relatively short time since sequencing was generated for v3.0 of the assembly, significant advances were made in both sequencing and assembly technologies. Rather than continue with the tiered approach, a de novo assembly was begun with a new individual. The use of a new specimen was due primarily to two factors. First, the material obtained and utilized in the first 3 iterations of the reference genome had been utilized often enough that noticeable degradation was observed as a smear in high-molecular weight DNA extractions. This led to relatively low PacBio read lengths with libraries generated in improvements to the prior

assembly. While this served a purpose in filling gaps within the relatively fragmented prior assembly, it indicated that a new material source would be required to improve on read-lengths. Second, the amount of material required for building libraries for iterative improvements to the original genome assembly had exhausted much of what had been a small sample. Hi-C and PacBio methods both utilized more genetic material than was available, requiring the sourcing of a new specimen.

While the long-read assembly presented in this work represents a significant improvement over the prior hybrid assembly, continued advancements in genomic technologies and especially read-lengths will likely lead to further improvements in overall contiguity. As prices decrease, we are also likely to see the introduction of assemblies with specific uses rather than a single reference genome. Indeed, the work of Pan et al (Pan et al., 2019) produced a long-read assembly of a European male northern pike, with the aim of characterizing the region of the sex-determination gene within the genome. As such works are assembled and released, it is likely that the genome (or perhaps genomes, given the two biological sources) presented here will be North American representatives incorporated into a pan-genome for the species, with additional assemblies from European and Asian lineages being added to represent the full diversity of the Holarctic species. It is likely in the future that selection of a reference may be based on genetic similarity to a study site or strain, such that the reference utilized best captures the genomic sequence within the locality, improving read alignment and variant detection.

Alignment of our resequencing data to v3.0 and v4.0 of the northern pike reference genome demonstrated that reads align with greater accuracy to a more contiguous genome. This has implications on downstream analyses in terms of making interpretations and conclusions that are representative of true biological phenomena and not artifacts of misalignments.

Population genomics

Our SNP data revealed northern pike are an incredibly homogeneous species with extremely low levels of variation genome wide. On average, northern pike in our study were 0.0071% - 0.035% different from the reference genome (includes homozygous alternate and heterozygous sites). For comparison, in a 2016 study using whole genome data, humans were reported to be between 0.11% - 0.13% different from the reference genome (The 1000 Genomes Project Consortium, 2015), a difference of at least 3-fold and at most 20-fold. We observed an average of one heterozygous SNP every 6,250 bases in Yukon River drainage populations, and one every 16,500 bases in all populations east of the NACD. For comparison, an average of one heterozygous SNP per 309 bases was reported in herring (Martinez Barrio et al., 2016), one per 500 in Atlantic cod (Star et al., 2011), and one per 750 bases in rainbow trout (Gao et al., 2018). Individuals sampled from Palmer Lake had similar numbers of heterozygous SNPs to populations east of the NACD (Figure 2). Although located very close to the headwaters of the Yukon River, Palmer Lake may be isolated from other water bodies. We believe isolation and genetic drift caused homogenization in this population. The individual from Castlegar was collected from west of the NACD but had the fewest number of heterozygous SNPs (14,565). This fish, which is the basis for the reference genome v4.0, is from an invasive population in the Columbia River whose source can be traced back to the east of the NACD (Carim, 2018). This translocation history supports the low number of heterozygous SNPs observed, despite being collected west of the NACD. However, because sequence data for this fish was generated by a different method (10X Chromium library + HiSeq X PE150) than other pike (shotgun library + HiSeq X Ten PE150), we cannot exclude the possibility that alternate library and sequencing platforms have contributed to the lower SNP count in some way.

Northern pike joins cheetahs (*Acinonyx jubatus*), albatrosses (*Diomedea exuland* and *Diomedea amsterdamensis*), northern elephant seals (*Mirounga angustirostris*), and channel island foxes (*Urocyon littoralis*) as a species surviving with low levels of genetic variation (Abadía-Cardoso et al., 2017; Merola, 1994; Milot et al., 2007; J. A. Robinson et al., 2016), challenging the concept that greater levels of variation are associated with the ability to adapt and survive. Increased SNP density in the northern pike occurs in immune-related regions of the genome including the MHC. Indeed, the MHC is known to be highly variable in vertebrates, and this polymorphism is valuable in terms of antigen diversity and the ability to effectively recognize pathogens (Piertney & Oliver, 2006; Unanue et al., 2016) SNP density is greatest in olfactory receptor genes in channel island foxes (J. A. Robinson et al., 2016); genes that control bony plate number and morphology in stickleback (T. C. Nelson et al., 2019); and genes with functions in sensory perception and neurophysiology in humans (Redon et al., 2006). These observations suggest a pattern: regions of high variation in the genome are associated with phenotypic variation whose sensitivity may be instrumental to the survival of the species. This may mean that the maintenance of genetic variation is most crucial in genetic regions that affect a species' ability to detect/respond to stimuli that are both fundamental to survival and in flux in their habitat.

Previous studies have been unable to distinguish different populations of northern pike in eastern North America (Miller & Senanan, 2003; Senanan & Kapuscinski, 2000; Skog et al., 2014), but in agreement with (Ouellet-Cauchon et al., 2014) our clustering analyses indicate that population-level variation does exist despite the overall low level of variation in the genome, and could be used for population management in a similar manner to recent work on *Esox masquinongy* (Rougemont et al., 2019).

The two to three-fold excess of heterozygous loci, three-fold increase of H_o, and three-fold increase in π in northwestern northern pike populations indicate that these are the oldest in North America. Bottlenecks, founder effects, and inbreeding can all lead to a loss of polymorphic loci (i.e. increase in the number of homozygous loci), reduced H_o, and reduced nucleotide diversity. Together, these results strongly indicate that at least one, and likely all, of these processes affected the northern pike east of the NACD. The depression of Tajima's D from -0.47 in Alaska to -1.61 east of the NACD (Table 5) suggests that northern pike populations in eastern North America have recently undergone population expansion. The drastic difference in these diversity statistics between northwestern and eastern populations of northern pike suggests that pike east of the NACD colonized from a small genetically homogenous population (or populations), and are younger than their northwestern counterparts.

Sex Determination

Initial investigations on sex-determination in northern pike began as a QTL and trait mapping exercise. Northern pike were generated from 5 family crosses from wild-caught parents in the thousand islands region of the St. Lawrence region with sex determined by histological examination of gonadal tissue. Difficulties in resulting analyses were encountered for a few reasons. First, analyses indicated families were strongly sex-biased, with three families almost exclusively male (male:female ratio of: 86:12; 44:14; 55:5), with two families still male biased but with more equal sex-ratios (male:female ratio of = 39:21; 33:26). Second, while 1,611 designed microsatellite primer pairs were available for map construction, the lack of polymorphism left between 125-250 polymorphic markers per parent pair, with linkage groups only thinly, or in some cases not, covered by genotyped markers. Resulting analyses with what markers were successfully genotyped did not yield any statistically significant association

through direct linkage mapping or QTL analysis (unpublished results). Early tests in collaboration with the authors of Pan et al., 2020 failed to detect *amhby* (unpublished). Thus, a change in approach was taken by utilizing the family parents amongst the re-sequenced individuals. Along with parents collected for a similar attempt the year prior (whereby too few offspring attained sufficient size for histology), these comprised the St. Lawrence (New York) individuals used in resequencing.

Sex specific analyses on our resequenced individuals revealed that the male-specific sex determining region on LG24 described by Pan et al. (2019) was present and associated with males in individuals from Chatanika River, Alaska, but absent from all males and females east of the NACD. This supports the finding of Pan et al. (2020) through an individual-based, high sequence resolution methodology. PCR assays targeting *amhby* corroborated these results, and surprisingly revealed the presence of *amhby* in males from Castlegar, BC. Our resequenced individual from Castlegar clusters tightly with eastern populations, but the presence of *amhby* in males could indicate an additional long-distance introduction of northern pike from Alaska, or could indicate that pike from the southwestern range of North America are descendents from a separate population where *amhby* was maintained.

PCR assays revealed the presence of amhby in some, but not all males in the Minto Flats of Alaska. This observation proves that *amhby* is not the sole determiner of sex in this population of northern pike, and supports a possible route by which amhby could have been lost in eastern North America. Given that *amhby* is not necessary for the development of testes in the Minto Flat (Alaska) population, and that both *amhby* and all male-specific heterozygosity are absent in northern pike east of the NACD, our results support a scenario in which the population that founded east of the NACD was one that recently lost *amhby* through female to male sex

reversals. In such a scenario, progeny would not only lack *amhby*, but also be void of all male-specific genetic variation genome-wide, consistent with our observations. The loss of *amhby* and lack of sex-specific genetic variation east of the NACD, begs the question: what took over the role of sex determination? Epigenetic factors or environmental influences could be at play. Additionally, population-specific systems may be evolving.

Conclusions

Here we present a long-read assembly for northern pike with an N50 of 37.5 Mbp and where 97.5% of sequence data is anchored on chromosomes. The contiguity of this genome surpasses previous versions constructed with an abundance of shorter length read data from multiple platforms. We demonstrate that this long read, contiguous genome allowed resequencing data to be aligned with greater accuracy and quality than less contiguous versions. Our resequencing analysis reveals a remarkable lack of genetic variation in the northern pike genome and illustrates that polymorphism is clustered in regions that are largely associated with immunity. Our results are consistent with scenarios and timelines put forth by (Pan et al., 2020; Skog et al., 2014)t, suggesting that northern pike originally colonized North America through Beringia and a small population that lost the sex-determining gene (*amhby*) left the Alaska/Yukon region around 80 - 120 thousand years ago, possibly during a glacial retreat, and was able to colonize North America east of the NACD. Glacial advances pushed this population(s) south, and they became the founders for re-colonization east of the North American NACD after the final retreat of the ice sheets some 12,000 years ago. Alaska was known to be ice-free during the last ice age, and the source population for eastern North American populations likely persisted there throughout the ice age and through the present day. We uncover a population in the Minto Flats of Alaska that is actively undergoing the loss of

amhby through development of males lacking this gene. Further testing of northern pike populations from the northern and southern extent of their range in North America are needed to delineate specific recolonization patterns and determine the prevalence of the sex-determining gene *amhby*. The nature of sex determination in northern pike populations lacking amhby requires further investigation.

Acknowledgements

We are very grateful to the Charles O. Hayford State Fish Hatchery (Hackettstown, NJ, USA), the Alaska Department of Fish and Game, the Manitoba Fisheries Branch, and Whiteshell fish Hatchery (West Hawk Lake, Manitoba), Jeremy Baxter (Mountain Water Research) and Marco Marrello (Terraquatic Resource Management) for sample collection and provision. We thank Qiaowei Pan and Yann Guiguen for their collaboration and insight that contributed to the intiation of our sex-specific analyses. This work was supported by NSERC (RGPIN/3888-2017) and the New York Environmental Protection Fund AM-10165 administered by the NYS Department of Environmental Conservation.

References

Baroiller, J. F., D'Cotta, H., Bezault, E., Wessels, S., & Hoerstgen-Schwark, G. (2009). Tilapia sex determination: Where temperature and genetics meet. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *153*(1), 30–38. https://doi.org/10.1016/j.cbpa.2008.11.018

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44. https://doi.org/10.1016/j.tree.2007.09.008

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bosworth, A., & Farrell, J. M. (2006). Genetic Divergence among Northern Pike from Spawning Locations in the Upper St. Lawrence River. *North American Journal of Fisheries Management*, *26*(3), 676–684. https://doi.org/10.1577/M05-060.1

Carbine, W. F. (1942). Observations on the Life History of the Northern Pike, *Esox Lucius* L., in Houghton Lake, Michigan. *Transactions of the American Fisheries Society*, *71*(1), 149–164. https://doi.org/10.1577/1548-8659(1941)71[149:OOTLHO]2.0.CO;2

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews. Genetics*, *16*(11), 627–640. https://doi.org/10.1038/nrg3933

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/nmeth.2474

Clark, C. F. (1950). Observations on the Spawning Habits of the Northern Pike, *Esox lucius* , in Northwestern Ohio. *Copeia*, *1950*(4), 285–288. JSTOR. https://doi.org/10.2307/1437909

Craig, J. F. (2008). A short review of pike ecology. *Hydrobiologia*, *601*(1), 5–16. https://doi.org/10.1007/s10750-007-9262-3

Crossman, E. J., & Harington, C. R. (1970). Pleistocene Pike, Esox lucius, and Esox sp., from the Yukon Territory and Ontario. *Canadian Journal of Earth Sciences*, *7*(4), 1130–1138. https://doi.org/10.1139/e70-107

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Devlin, R. H., & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: An overview of genetic, physiological, and environmental influences. *Aquaculture*, *208*(3), 191–364. https://doi.org/10.1016/S0044-8486(02)00057-1

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps

with Unlimited Zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Feron, R., Zahm, M., Cabau, C., Klopp, C., Roques, C., Bouchez, O., Eché, C., Valière, S., Donnadieu, C., Haffray, P., Bestin, A., Morvezen, R., Acloque, H., Euclide, P. T., Wen, M., Jouano, E., Schartl, M., Postlethwait, J. H., Schraidt, C., … Guiguen, Y. (2019). Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, Perca flavescens. *BioRxiv*, 717397. https://doi.org/10.1101/717397

Forsman, A., Tibblin, P., Berggren, H., Nordahl, O., Koch-Schmidt, P., & Larsson, P. (2015). Pike *Esox lucius* as an emerging model organism for studies in ecology and evolutionary biology: A review. *Journal of Fish Biology*, *87*(2), 472–479. https://doi.org/10.1111/jfb.12712

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), 527. https://doi.org/10.1186/s12864-017-3879-z

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*. https://doi.org/10.1101/gr.191395.115

Goto-Kazeto, R., Abe, Y., Masai, K., Yamaha, E., Adachi, S., & Yamauchi, K. (2006). Temperature-dependent sex differentiation in goldfish: Establishing the temperature-sensitive period and effect of constant and fluctuating water temperatures. *Aquaculture*, *254*(1), 617–624. https://doi.org/10.1016/j.aquaculture.2005.10.009

Government of Canada, F. and O. S. S. (2016, October 6). *2010 Survey of Recreational Fishing in Canada | Fisheries and Oceans Canada*. http://www.dfo-mpo.gc.ca/stats/rec/can/2010/index-eng.htm

Grande, L. (1999). The First Esox (Esocidae: Teleostei) from the Eocene Green River Formation, and a Brief Review of Esocid Fishes. *Journal of Vertebrate Paleontology*, *19*(2), 271–292. JSTOR.

Grande, Laten, H., López, J. A., & Quattro, J. M. (2004). Phylogenetic Relationships of Extant Esocid Species (Teleostei: Salmoniformes) Based on Morphological and Molecular Characters. *Copeia*, *2004*(4), 743–757. https://doi.org/10.1643/CG-04-007R1

Hattori, R. S., Murai, Y., Oura, M., Masuda, S., Majhi, S. K., Sakamoto, T., Fernandino, J. I., Somoza, G. M., Yokota, M., & Strüssmann, C. A. (2012). A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proceedings of the*

National Academy of Sciences, *109*(8), 2955–2959. https://doi.org/10.1073/pnas.1018392109

Höglund, J. (2009). *Evolutionary Conservation Genetics*. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199214211.001.0001/acprof-9780199214211

Huffman, K., Farrell, J. M., & Whipps, C. M. (2014). *Environmental Determinants of Sex Ratio in St. Lawrence River Northern Pike: Development of a Molecular Sex Identification Tool and Experimentation with Physical and Chemical Variables*. *144th Annual Meeting*. https://www.researchgate.net/publication/267898945_Environmental_Determinants_of_Sex_Ratio_in_St_Lawrence_River_Northern_Pike_Development_of_a_Molecular_Sex_Identification_Tool_and_Experimentation_with_Physical_and_Chemical_Variables

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., Xue, Z., Mohamadi, H., Bohlmann, J., Jones, S. J. M., & Birol, I. (2018). Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, *19*(1), 393. https://doi.org/10.1186/s12859-018-2425-6

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., Tohari, S., Brenner, S., Miyadai, T., Venkatesh, B., Suzuki, Y., &

Kikuchi, K. (2012). A Trans-Species Missense SNP in *Amhr2* Is Associated with Sex Determination in the Tiger Pufferfish, *Takifugu rubripes* (Fugu). *PLOS Genetics*, *8*(7), e1002798. https://doi.org/10.1371/journal.pgen.1002798

Kawase, J., Aoki, J., Hamada, K., Ozaki, A., & Araki, K. (2018). Identification of Sex-associated SNPs of Greater Amberjack (Seriola dumerili). *Journal of Genomics*, *6*, 53–62. https://doi.org/10.7150/jgen.24788

Kikuchi, K., & Hamaguchi, S. (2013). Novel sex-determining genes in fish and sex chromosome evolution. *Developmental Dynamics*, *242*(4), 339–353. https://doi.org/10.1002/dvdy.23927

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

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. https://doi.org/10.1101/gr.215087.116

Lee, T.-H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, *15*(1), 162. https://doi.org/10.1186/1471-2164-15-162

Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., Munro, S., Moore, R., Holt, R. A., Jones, S. J. M., Davidson, W. S., & Koop, B. F. (2010). *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in

evolutionary pressures on a post-tetraploidization genome. *BMC Genomics*, *11*, 279. https://doi.org/10.1186/1471-2164-11-279

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, M., Sun, Y., Zhao, J., Shi, H., Zeng, S., Ye, K., Jiang, D., Zhou, L., Sun, L., Tao, W., Nagahama, Y., Kocher, T. D., & Wang, D. (2015). A Tandem Duplicate of Anti-Müllerian Hormone with a Missense SNP on the Y Chromosome Is Essential for Male Sex Determination in Nile Tilapia, *Oreochromis niloticus*. *PLOS Genetics*, *11*(11), e1005678. https://doi.org/10.1371/journal.pgen.1005678

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Matsuda, M. (2018). Genetic Control of Sex Determination and Differentiation in Fish. In K. Kobayashi, T. Kitano, Y. Iwao, & M. Kondo (Eds.), *Reproductive and Developmental Strategies: The Continuity of Life* (pp. 289–306). Springer Japan. https://doi.org/10.1007/978-4-431-56609-0_14

Matsuda, M., Shinomiya, A., Kinoshita, M., Suzuki, A., Kobayashi, T., Paul-Prasanth, B., Lau, E., Hamaguchi, S., Sakaizumi, M., & Nagahama, Y. (2007). DMY gene induces male

development in genetically female (XX) medaka fish. *Proceedings of the National Academy of Sciences*, *104*(10), 3865–3870. https://doi.org/10.1073/pnas.0611707104

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Miller, L. M., & Kapuscinski, A. R. (1996). Notes: Microsatellite DNA Markers Reveal New Levels of Genetic Variation in Northern Pike. *Transactions of the American Fisheries Society*, *125*(6), 971–977. https://doi.org/10.1577/1548-8659(1996)125<0971:NMDMRN>2.3.CO;2

Miller, L. M., & Kapuscinski, A. R. (1997). Historical Analysis of Genetic Variation Reveals Low Effective Population Size in a Northern Pike (*Esox lucius*) Population. *Genetics*, *147*(3), 1249–1258.

Miller, L. M., & Senanan, W. (2003). A Review of Northern Pike Population Genetics Research andIts Implications for Management. *North American Journal of Fisheries Management*, *23*(1), 297–306. https://doi.org/10.1577/1548-8675(2003)023<0297:ARONPP>2.0.CO;2

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., Cao, H., Schlebusch, S. A., Giorda, K., Schnall-Levin, M., Wall, J. D., & Kwok, P.-Y. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature Methods*, *13*(7), 587–590. https://doi.org/10.1038/nmeth.3865

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., & Sakaizumi, M. (2012). Tracing the Emergence of a Novel Sex-

Determining Gene in Medaka, *Oryzias luzonensis*. *Genetics*, *191*(1), 163–170. https://doi.org/10.1534/genetics.111.137497

Nelson, J. S. (2006). *Fishes of the World* (4 edition). Wiley.

Pan, Q., Feron, R., Jouanno, E., Darras, H., Herpin, A., Koop, B., Rondeau, E., Goetz, F. W., Larson, W. A., Bernatchez, L., Tringali, M., Curran, S. S., Saillant, E., Denys, G. P. J., Hippel, F. A. von, Chen, S., López, J. A., Verreycken, H., Ocalewicz, K., … Guiguen, Y. (2020). The rise and fall of the ancient northern pike master sex determining gene. *BioRxiv*, 2020.05.31.125336. https://doi.org/10.1101/2020.05.31.125336

Pan, Q., Feron, R., Yano, A., Guyomard, R., Jouanno, E., Vigouroux, E., Wen, M., Busnel, J.-M., Bobe, J., Concordet, J.-P., Parrinello, H., Journot, L., Klopp, C., Lluch, J., Roques, C., Postlethwait, J., Schartl, M., Herpin, A., & Guiguen, Y. (2019). Identification of the master sex determining gene in Northern pike (Esox lucius) reveals restricted sex chromosome differentiation. *PLOS Genetics*, *15*(8), e1008013. https://doi.org/10.1371/journal.pgen.1008013

Pandian, T. J. (2011). *Sex Determination in Fish*. Taylor & Francis.

Piertney, S. B., & Oliver, M. K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity*, *96*(1), 7–21. https://doi.org/10.1038/sj.hdy.6800724

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. V. der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178. https://doi.org/10.1101/201178

Priegel, G. R., & Krohn, D. C. (1975). *Characteristics of a Northern Pike Spawning Population* (No. 86). Wisconsin Department of Natural Resources. https://dnr.wi.gov/files/PDF/pubs/ss/SS0086.pdf

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., & Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, *26*(3), 342–350. https://doi.org/10.1101/gr.193474.115

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rambaut, A. (2016). *Figtree* (Version 1.4.3) [Computer software]. http://tree.bio.ed.ac.uk/software/figtree/

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Systems*, *6*(2), 256-258.e1. https://doi.org/10.1016/j.cels.2018.01.001

Rondeau, E. B., Messmer, A. M., Sanderson, D. S., Jantzen, S. G., von Schalburg, K. R., Minkley, D. R., Leong, J. S., Macdonald, G. M., Davidsen, A. E., Parker, W. A.,

Mazzola, R. S., Campbell, B., & Koop, B. F. (2013). Genomics of sablefish (Anoplopoma fimbria): Expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. *BMC Genomics*, *14*, 452. https://doi.org/10.1186/1471-2164-14-452

Rondeau, E. B., Minkley, D. R., Leong, J. S., Messmer, A. M., Jantzen, J. R., Schalburg, K. R. von, Lemon, C., Bird, N. H., & Koop, B. F. (2014). The Genome and Linkage Map of the Northern Pike (*Esox lucius*): Conserved Synteny Revealed between the Salmonid Sister Group and the Neoteleostei. *PLOS ONE*, *9*(7), e102089. https://doi.org/10.1371/journal.pone.0102089

Seeb, J. E., Seeb, L. W., Oates, D. W., & Utter, F. M. (1987). Genetic Variation and Postglacial Dispersal of Populations of Northern Pike (*Esox lucius*) in North America. *Canadian Journal of Fisheries and Aquatic Sciences*, *44*(3), 556–561. https://doi.org/10.1139/f87-068

Senanan, W., & Kapuscinski, A. R. (2000). Genetic relationships among populations of northern pike (*Esox lucius*). *Canadian Journal of Fisheries and Aquatic Sciences*, *57*(2), 391–404. https://doi.org/10.1139/f99-261

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227–245). Springer. https://doi.org/10.1007/978-1-4939-9173-0_14

Skog, A., Vøllestad, L. A., Stenseth, N. C., Kasumyan, A., & Jakobsen, K. S. (2014). Circumpolar phylogeography of the northern pike (Esox lucius) and its relationship to the Amur pike (E. reichertii). *Frontiers in Zoology*, *11*(1), 67. https://doi.org/10.1186/s12983-014-0067-8

Skov, C., & Nilsson, P. A. (2018). *Biology and Ecology of Pike* (1st ed.). CRC Press. https://www.taylorfrancis.com/books/e/9781315119076

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. https://doi.org/10.1093/nar/gkr123

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(19), 7702–7707. https://doi.org/10.1073/pnas.0901054106

Turner, S. (2018). qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731. https://doi.org/10.21105/joss.00731

Unanue, E. R., Turk, V., & Neefjes, J. (2016). Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annual Review of Immunology*, *34*, 265–297. https://doi.org/10.1146/annurev-immunol-041015-055420

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1-33. https://doi.org/10.1002/0471250953.bi1110s43

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for

Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1), 35. https://doi.org/10.1186/s13742-015-0076-3

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wilson, M. V. H. (1980). Oldest known Esox (Pisces: Esocidae), part of a new Paleocene teleost fauna from western Canada. *Canadian Journal of Earth Sciences*, *17*(3), 307–312. https://doi.org/10.1139/e80-030

Yano, A., Guyomard, R., Nicol, B., Jouanno, E., Quillet, E., Klopp, C., Cabau, C., Bouchez, O., Fostier, A., & Guiguen, Y. (2012). An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, *Oncorhynchus mykiss*. *Current Biology*, *22*(15), 1423–1428. https://doi.org/10.1016/j.cub.2012.05.045

Yano, A., Nicol, B., Jouanno, E., Quillet, E., Fostier, A., Guyomard, R., & Guiguen, Y. (2013). The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evolutionary Applications*, *6*(3), 486–496. https://doi.org/10.1111/eva.12032

Yeo, S., Coombe, L., Warren, R. L., Chu, J., & Birol, I. (2018). ARCS: Scaffolding genome drafts with linked reads. *Bioinformatics*, *34*(5), 725–731. https://doi.org/10.1093/bioinformatics/btx675

Data Accessibility

-Reference Genome v2.0: NCBI WGS accession AZJR03000000.2 and assembly accession GCA_000721915.2.

-Reference Genome v3.0: WGS accession JAAIYR000000000.1 and assembly accession GCA_011004845.1.

-Reference Genome v4.0: NCBI WGS accession SAXP01000000.1 and assembly accession GCA_004634155.1, with RefSeq assembly accession GCF_004634155.1.

-Whole genome resequencing data: Bioproject Accession PRJNA512507 and sequence read archive accession numbers SAMN10685075 – SAMN10685119.

- SNP set from: Population genomics of North American northern pike: variation and sex-specific signals from a chromosome-level, long read genome assembly. Filtered VCF file aligned to reference genome v4.0. https://doi.org/10.5061/dryad.rfj6q577h.

Author Contributions

Designed Research: Ben F Koop, Eric B Rondeau

Performed Research: Eric B Rondeau, Hollie A Johnson, Joanne Whitehead, Cody A Despins, Brent E Gowen, Christopher M Whipps, John M Farrell, Brian J Collyard.

Analyzed Data: Hollie A Johnson, Eric B Rondeau, David R Minkley, Jong S Leong, Joanne Whitehead.

Wrote the Paper: Hollie A Johnson, Eric B Rondeau

Figures

Figure 1. Sampling locations of resequenced pike. Colour scheme is carried throughout figures. CR – Chatanika River, Alaska. YR – Yukon River at Hootalinqua, Yukon Territory. PL – Palmer Lake, British Columbia. CL – Charlie Lake, British Columbia. CaG – Columbia River, Castlegar, British Columbia. Mb – Whiteshell Hatchery, Manitoba. NY – St. Lawrence River, New York. NJ – Hackettstown Hatchery, New Jersey.

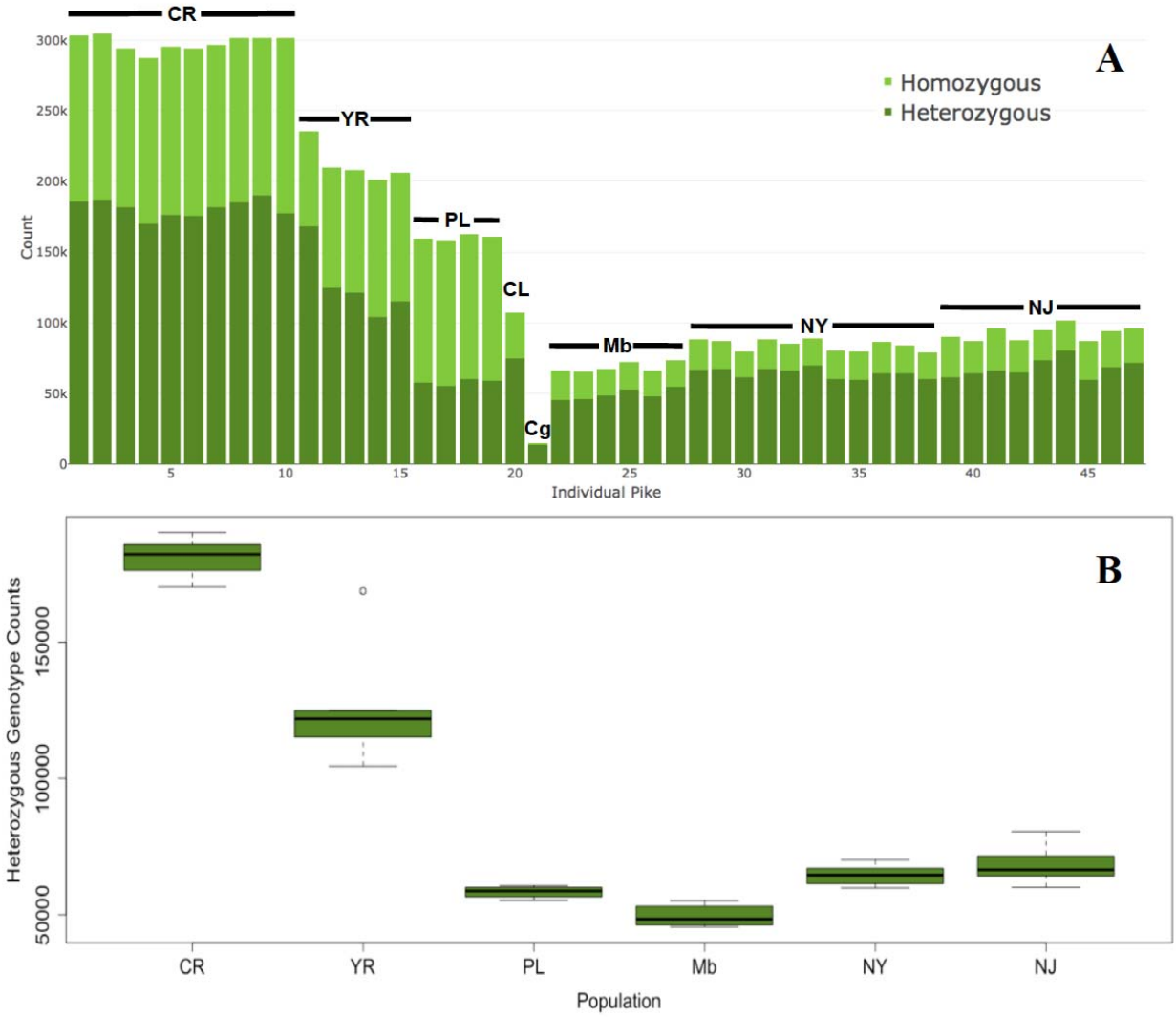

Figure 2. Genotype counts. In (A), homozygous alternate (light green) and heterozygous (dark green) genotype counts per resequenced northern pike genome are plotted. In (B), box plots of heterozygous genotype counts are plotted for each population. Note the northern pike reference genome is 941 Mb. In both A and B, populations are indicated by the following abbreviations: CR – Chatanika River, Alaska. YR – Yukon River at Hootalinqua, Yukon Territory. PL – Palmer Lake, British Columbia. CL – Charlie Lake, British Columbia. Cg – Columbia River at Castlegar, British Columbia. Mb – Whiteshell Hatchery, Manitoba. NY – St. Lawrence River, New York. NJ – Hackettstown Hatchery, New Jersey.

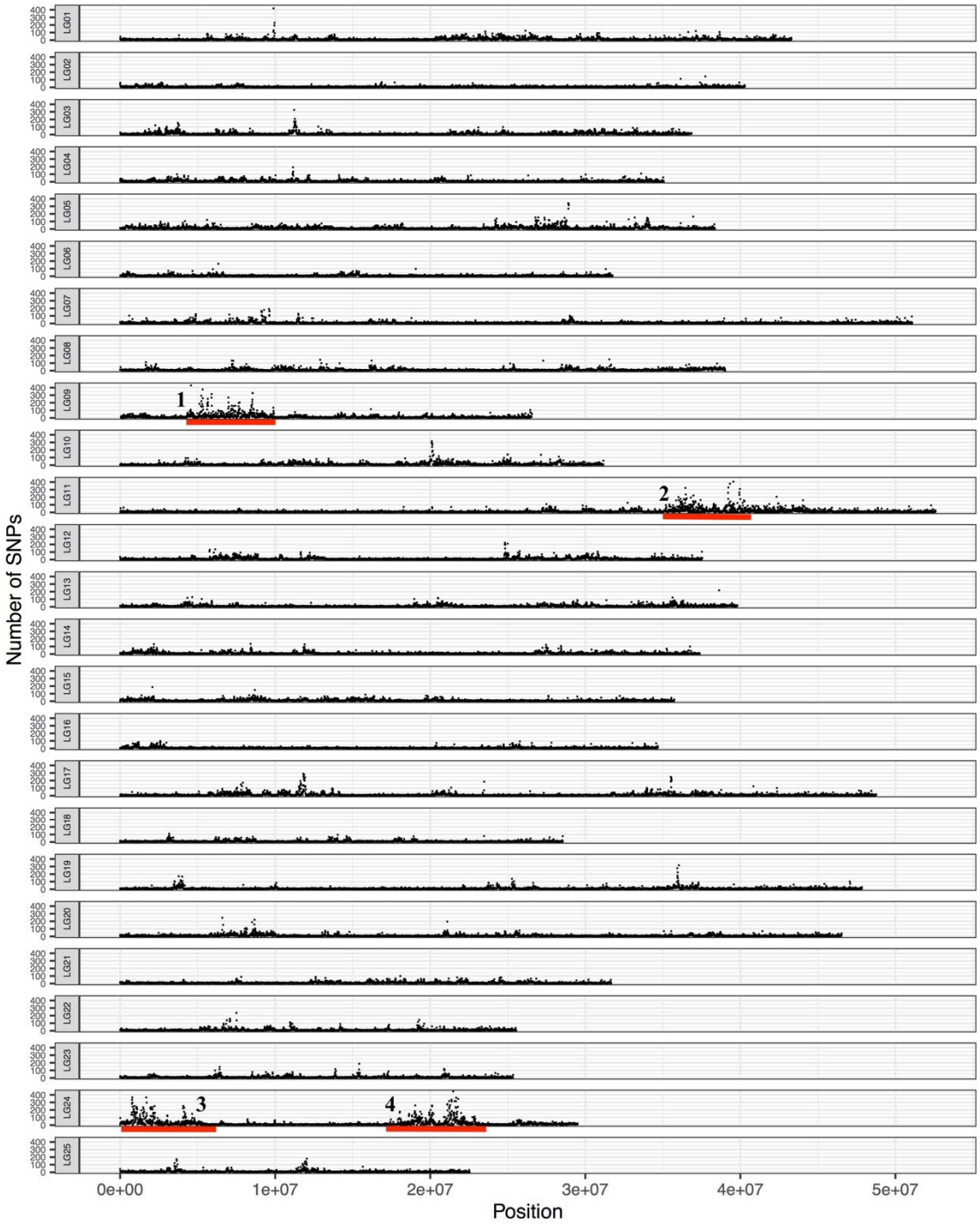

Figure 3. Genome-wide SNP density by chromosome. Chromosome-panels with SNP count per 10,000 base pairs. All resequenced individuals were pooled for this analysis. Red lines indicate regions where the heterozygous variant count per 10kb is elevated for swaths of greater than 1MB, and are described in Table 5.

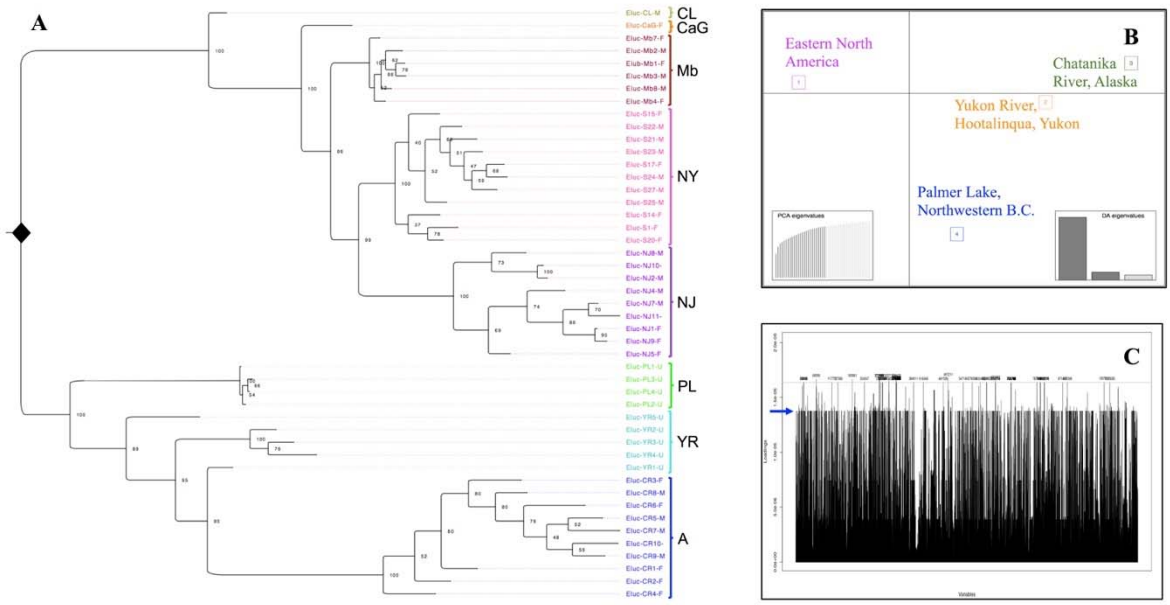

Figure 4. Population structure of North American northern pike. In (A), a dendrogram with 1000 bootstraps (values in percent), rooted by midpoint, showing that individuals group by geographic proximity, with most major groupings supported by bootstrap values of greater than 90%. CL - Charlie Lake, CaG - Castlegar, Mb - Manitoba, NY - New York, NJ - New Jersey, PL - Palmer Lake, YR - Yukon River, A – Chatanika River, Alaska. DAPC is plotted in (B) and shows the majority of variation in the data is explained by eigenvalue 1, which separates groups east and west of the continental divide. Eastern North America includes CL, CaG, Mb, NY, and NJ. A loading plot of the first discriminate axis is show in (C). The blue arrow points to genome-wide loadings of SNPs that are homozygous reference in eastern North American populations, and homozygous alternate in Chatanika River, Yukon River, and Palmer Lake.

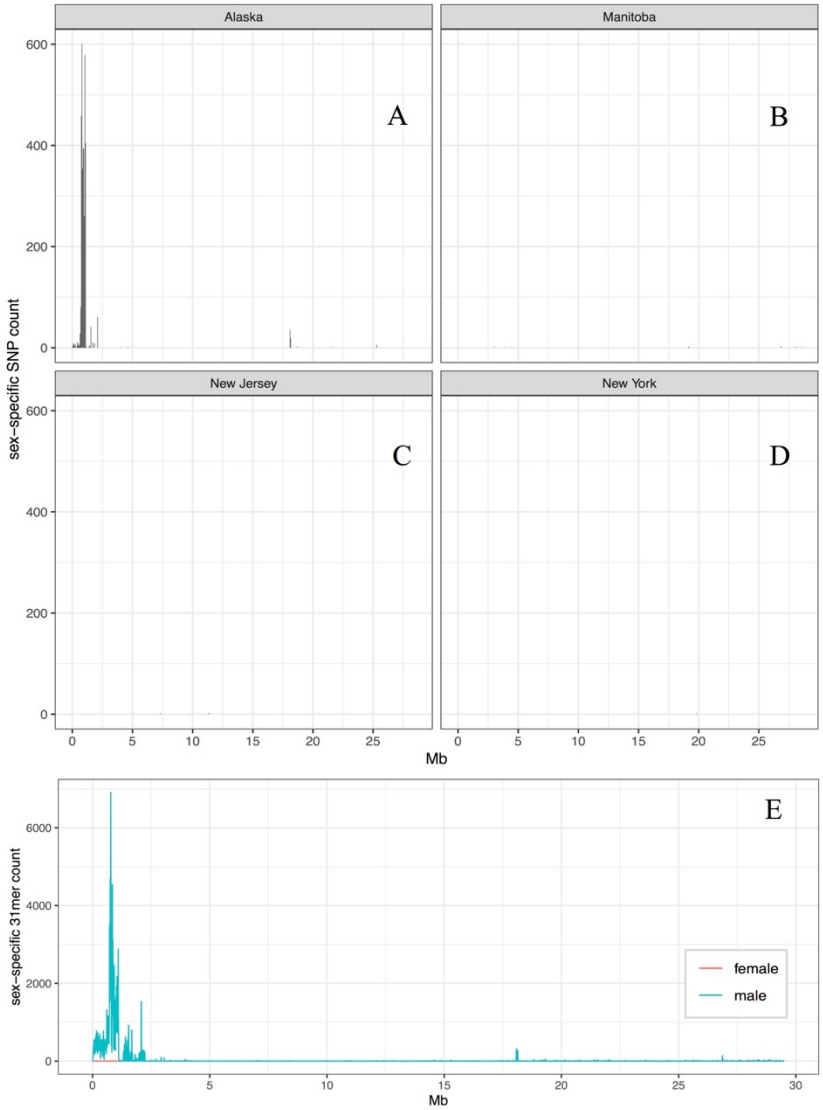

Figure 5. Sex-specific SNP and k-mer counts. Panels A-D show sex-specific SNP counts on LG24 for Chatanika River (Alaska), Manitoba, New Jersey, and New York populations, highlighting the absence of sex-specific signals in populations east of the NACD. In panel E, k-mer counts on LG24 of the Chatanika River population are shown.

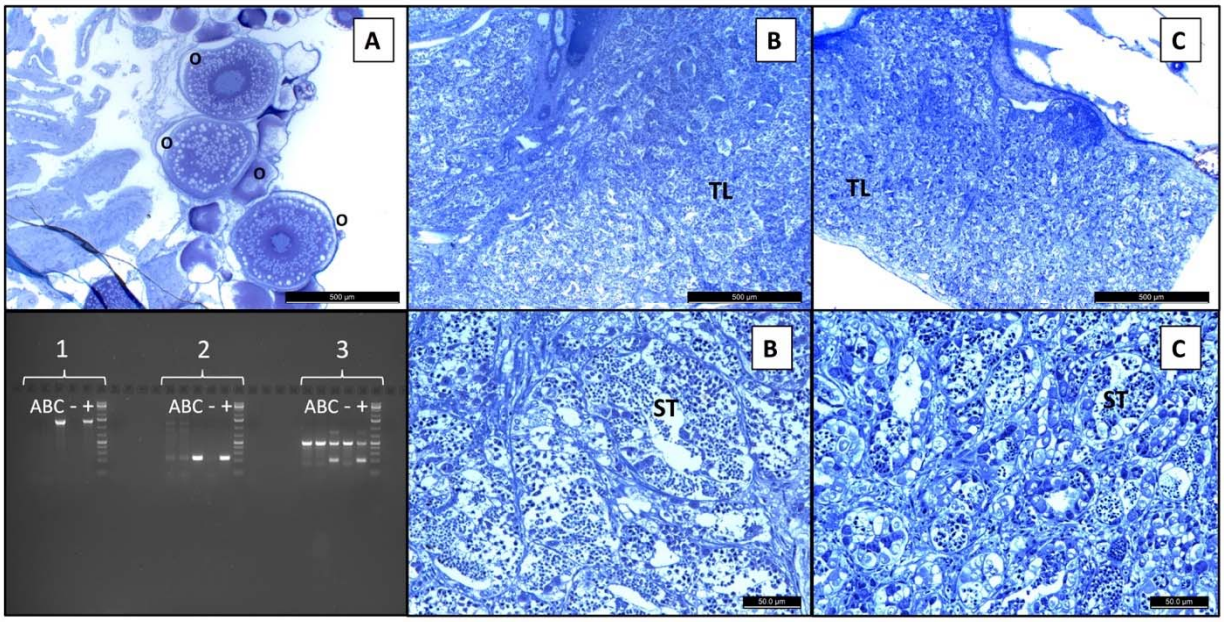

Figure 6. Northern pike of Minto Flats, Alaska: gonadal histology and *amhby*. O = oocytes at various stages of vitellogenesis. TL = testis lobule. ST = seminiferous tubules with gametes at varying states of spermatogenesis. Slide A – phenotypic female and *amhby* negative. B - phenotypic male and *amhby* negative. C – phenotypic male and *amhby* positive. PCR1 was performed with conserveAMH1-1 primers from Pan et al., 2020. PCR2 was performed with SeqAMH1-4 primers also from Pan et al., 2020. PCR3 performed with primer set 24.5 from this paper. Negative control is a female – *amhby* negative pike from Chatanika River. Positive control is male – *amhby* positive from Chatanika River. Both controls were resequenced.

Tables

Population	Latitude	Longitude	Males	Females	Total
Chatanika River, Alaska, U.S.A.	64.98396	-148.86032	5	5	10
Hootalinqua, Yukon Territory, CANADA	61.51156	-135.13209	?	?	5
Palmer Lake, British Columbia, CANADA	59.43708	-133.57592	?	?	4
Charlie Lake, British Columbia, CANADA†	56.32853	-120.97835	1	-	1
Castlegar, British Columbia, CANADA ‡	49.31538	-117.65344	-	1	1
Whiteshell Hatchery, Manitoba, CANADA	49.80051	-95.17243	3	3	6
St. Lawrence Waterway, New York, U.S.A.	44.24787	-76.09785	6	5	11
Hackettstown Hatchery, New Jersey, U.S.A.	40.84155	-74.83359	6	3	9
Total			21+?	17+?	47

Table 1: Origin of northern pike used in resequencing analysis. (?) Indicates that sex was unknown. (-) Indicates that no samples of this sex were collected. †Indicates northern pike used to build reference genome v3.0. ‡Indicates northern pike used to build the reference genome v4.0.

Filter Description	Program	Number of Variants Remaining (v4.0)	Number of Variants Remaining (v3.0)
Raw Genotype Calls†	-	1,910,789	3,606,293
Extract SNPs	GATK	1,363,731	2,222,114
GATK Hard Filter	GATK	1,189,068	1,604,021
Minimum Quality 20	VCFtools	1,186,793	1,600,847
Minimum Mean Depth 10	VCFtools	1,153,668	1,480,998
Maximum Mean Depth 60	VCFtools	1,152,122	1,440,437
Max Missing Count 10	VCFtools	1,129,884	1,048,979
Minor Allele Count 1	VCFtools	1,129,701	1,023,826
All Heterozygote Filter	R/ VCFtools	1,127,943	944,574

Table 2. Filtering parameters used to prune SNPs from VCF files after alignment to the northern pike reference genome versions 3 and 4. †Raw genotype calls include SNPs and InDels.

		Number	Bases (Mb)	N50 (Sequence length)	N50 (number of sequences)	Maximum Length (bp)
Version 1†	**Contigs**	94,266	824	16,910	13,483	232,364
	Scaffolds	5,688	878	700,535	318	5,140,982
Version 2	**Contigs**	18,738	892	126,056	1,985	1,089,236
	Scaffolds	1,708	904	4,231,106	45	23,611,278
Version 3	**Contigs**	18,750	890	125,635	1,991	1,089,236
	Scaffolds	1,211	904	7,945,253	29	25,961,733
Mapped	**Unplaced Scaffolds**	993	108	423,854	68	2,378,070
	Chromosomes	25	796	-	-	-
Version 4	**Contigs**	1,395	941	3,396,779	71	18,262,732
	Scaffolds	811	941	37,550,661	11	52,601,242
Mapped	**Unplaced/Unlocalized Scaffolds**	785	23	33,367	183	510,802
	Chromosomes	25 + mito	918	-	-	-

Table 3. Summary genome statistics for the iterative hybrid assembly releases (RefSeq Versions 1-3) and the long-read assembly (Version 4). †Rondeau et al, 2014

A. V4.0 - Castlegar PacBio		
Database	Actinopterygii odb10	Vertebrate odb10
Complete	3474 (95.4%)	3254 (97.1%)
Single Copy	3426 (94.1%)	3212 (95.8%)
Duplicated	48 (1.3%)	42 (1.3%)
Fragmented	46 (1.3%)	54 (1.6%)
Missing	120 (3.3%)	46 (1.3%)
Total Searched	3640	3354
B. V3.0 - Charlie Lake Hybrid		
Database	Actinopterygii odb10	Vertebrate odb10
Complete	3468 (95.3%)	3232 (96.3%)
Single Copy	3433 (94.3%)	3198 (95.3%)
Duplicated	35 (1.0%)	34 (1.0%)
Fragmented	57 (1.6%)	65 (1.9%)
Missing	115 (3.1%)	57 (1.8%)
Total Searched	3640	3354

Table 4. Description of Busco v4.0.2 results run with both release A) 3.0 (Charlie Lake Hybrid) and B) release 4.0 (Castlegar long-read) northern pike assemblies. For each assembly, Busco was run twice with Actinopterygii and Vertebrate odb10 databases.

#	LG : Region (MB)	Region Description
1	LG09: 4.5 – 10	Gene-rich region. Genes in this region are largely immune related. The major histocompatibility complex is located in this region.
2	LG11: 35 – 40	Gene-rich region. Immune and inflammation related genes (e.g. multiple copies of immunoglobulin-lamda light chain- like genes, NRLC3-like, and toll-like receptor 13).
3	LG24: 0.6 – 2.4	Male specific SNPs in Chatanika River. Immune related genes (e.g. multiple copies of Fc receptor- like genes and butyrophilin subfamily genes).
4	LG24: 21 – 22	SNPs density is mainly in Chatanika River and Hootalinqua populations. Genes affected involved multiple copies of Fc receptor-like genes involved in immunity, as well as multiple copies of ferric-chelate reductase and serine/threonine protein kinases.

Table 5. Regions of the northern pike genome where the heterozygous SNP count per 10kb is elevated over a swath of at least 1MB.

Group	Mean no. of heterozygous loci per fish	Observed Heterozygosity (H_o)	Nucleotide Diversity (π)	Tajima's D		N	
All Pike	93,429	0.08350	0.1250	-1.44		47	
Chatanika River	181,268	0.3228	0.2906	-0.228		10	
Yukon River, Hootalinqua	127,012	0.3171	0.3288	-0.34	-0.47†	5	19†
Palmer Lake	58,337	0.2849	0.2135	-0.09		4	
Eastern North America	61,073	0.0922	0.0915	-1.609		28	

Table 6. Summary of variation statistics. †Indicates values when Chatanika River, Yukon River, and Palmer Lake populations are pooled.

Primer Pair name	Region amplified	Chatanika River		Minto Flats, Alaska			Hootalinqua	Palmer Lake	Castlegar		Eastern NA	
		Males	Females	Males	Females	Unassigned	Not sexed	Not sexed	Males	Females	Males	Females
Amhby_conserve F1 & R1[†]	*amhby*: partial exon 2	5/5	0/5	6/14	0/6	1/1	1/5	0/4	3/3	0/3	0/16	0/11
SeqAMH_1 F4 & R4[†]	*amhby*: partial exon 7	5/5	0/5	6/14	0/6	1/1	1/5	0/4	3/3	0/3	0/16	0/11
24.5	LG24: 996,878 - 997,339	5/5	0/5	6/14	0/6	1/1	1/5	0/4	3/3	0/3	0/16	0/11

Table 7. Positive detections of *amhby* and the LG24 sex determining region. Numerators denote the number of pike in which the respective signal was detected and denominators denote the total number of pike of the noted sex. [†]Indicates primers developed by Pan (2020). Eastern NA includes pike from Charlie Lake (1M), Manitoba (3M, 3F), New York (6M, 5F), and New Jersey (6M, 3F).

Supplementary Documents

Supplementary Figure 1: Hi-C contact map as visualized in Juicebox v 1.8.8 for the long-read assembly v4.0 GCF_004634155.1.

Supplementary Figure 2: Distribution of repeats along the genome release v4.0 GCF_004634155.1, displayed in bins of 100,000 bp along the chromosome, with counts of masked bases per bin.

Supplementary Figure 3: Chromosome-level alignments between release 3.0 (GCF_011004845.1) and 4.0 (GCF_004634155.1) performed in Symap v4.2. In panel A, all 25 chromosomes are displayed. In panel B, a representative example from Eluc17 is shown to demonstrate improvements to scaffold orientation and chromosome completeness in v4.0.

Supplementary Figure 4. GWAS of all resequenced sexed pike. In total, 672,565 SNPs were analyzed across 21 males and 17 females. The red line indicates significance as assessed by Bonferonni correction. The blue line is the suggestive significance line, set by default to -log10(1e-5).

Supplementary Table 1: RepeatMasker summary table for version 3.0 (GCF_011004845.1) and version 4.0 (GCF_004634155.1) of the assemblies. Description of custom repeat library generation and masking methods described in Rondeau et al, 2014.

Supplementary Table 2: Genetic co-ordinates of sex-specific heterozygosity and associated genotypes in northern pike from Chatanika River, Alaksa. “0” represents homozygous reference and “1” represents heterozygous. Female-specific heterozygosity is highlighted in pink.

Supplementary File 1: Description of methods and results of versions 2.0 and 3.0 of the hybrid northern pike assembly, and description of PCR assays and primers used for evaluating sex-gene presence and absence.