# Loss of genetic variation and sex determination system in North American northern pike characterized by whole-genome resequencing

Hollie A. Johnson[1*], Eric B. Rondeau[1*], Ben J. G. Sutherland[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]Sutherland Bioinformatics, Lantzville, British Columbia, Canada V0R 2H0

[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 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

*Authors contributed equally

Author for correspondence: BFK

Email: bkoop@uvic.ca

Running title: Population genomics of North American pike

## Abstract

The northern pike *Esox lucius* is a freshwater fish renowned for having low genetic diversity but ecological success throughout the Northern Hemisphere. Here we generate an annotated chromosome-level genome assembly of 941 Mbp in length with 25 chromosome-length scaffolds using long-reads and chromatin capture technology. We then align whole-genome resequencing data against this reference to genotype northern pike from Alaska through New Jersey (n = 47). A striking decrease in genetic diversity occurs along the sampling range, whereby samples to the west of the North American Continental Divide have substantially higher diversity than populations to the east. As an example, individuals from Interior Alaska in the west and St. Lawrence River in the east have on average 181K and 64K heterozygous SNPs per individual, respectively (i.e., a SNP variant every 3.2 kbp and 11.2 kbp, respectively). Even with such low diversity, individuals clustered with strong support within each population, and this may be related to numerous private alleles in each population. Evidence for recent population expansion was observed for a Manitoba hatchery and the St. Lawrence population (Tajima's D = -1.07 and -1.30, respectively). Non-uniform patterns of diversity were observed across the genome, with large regions showing elevated diversity in several chromosomes, including LG24. In populations with the master sex determining gene *amhby* still present in the genome, *amhby* is in LG24. As expected, *amhby* was largely male-specific in Alaska and the Yukon and absent southeast to these populations, but we also document some *amhby*(-) males in Alaska and *amhby*(+) males in the Columbia River. This indicates that rather than a discrete boundary after which *amhby* was lost in North America, there is a patchwork of presence of this system in the western region. These results support the theory that northern pike recolonized North America from refugia in Alaska and expanded following deglaciation from west to east, with probable founder effects resulting in loss of both neutral and functional diversity including the loss of the sex determination system.


**Keywords**: *Esox lucius*; genetic variation; genomics; long-read assembly; northern pike; population genomics; sex determination; whole-genome resequencing

# Introduction

Historical population dynamics and contemporary genetic diversity and structure of freshwater fishes in the Northern Hemisphere have been significantly shaped by Pleistocene glacial cycles of expansion and retreat that occurred up to around 14,000 years ago (Bernatchez & Wilson, 1998; Skog et al., 2014; Wooller et al., 2015). Glacial advances resulted in large mortalities, range compressions to glacier edges, and isolation of populations into refugia. Glacial retreats resulted in formations of large postglacial lakes allowing rapid range recolonization across large geographic scales (see Bernatchez & Wilson, 1998). Northern pike *Esox lucius* (Order: Esociformes), with a distribution across much of the Northern Hemisphere in fresh and brackish water (Craig, 2008), was significantly impacted by these glacial cycles (Wooller et al., 2015). In eastern Eurasia, northern pike survived in a Siberian refugium and expanded into Beringia (Bachevskaja et al., 2019), which remained unglaciated during the Pleistocene and was a refugium for many freshwater species (reviewed by Wooller et al., 2015). Following glacial retreat, the Beringia refugium was the likely source for northern pike recolonization of Alaska (Crossman & Harington, 1970). However, Esocid ancestors of northern pike were in North America during the Paleocene (i.e., 56-66 mya) (Wilson, 1980), and therefore had a long history in the area prior to the glacial impacts (Wilson et al., 1992).

Genetic variation is considered pivotal for adaptation (Barrett & Schluter, 2008; Höglund, 2009), but northern pike throughout their range have low genetic diversity (Skov & Nilsson, 2018). Genetic diversity of northern pike is particularly low in central North America (Bosworth & Farrell, 2006; Miller & Kapuscinski, 1996; Rondeau et al., 2014; Senanan & Kapuscinski, 2000), in concordance with expectations due to glacial impacts on genomic diversity (Bernatchez & Wilson, 1998) resulting in a 'younger' and more recently recolonized population (reviewed by Skog et al., 2014). Although northern pike also have low diversity elsewhere in their range, including in Europe (Nicod et al., 2004), higher levels than those observed in North America occur in Sweden (Sunde et al., 2022), China (Luan et al., 2021), and Siberia (Senanan & Kapuscinski, 2000). Globally, the generally low diversity is likely due to ice age population bottlenecks and founder effects, small effective population sizes, restricted gene flow between populations, and the role of the pike as an apex predator, including with cannibalistic tendencies (Seeb et al., 1987; Skog et al., 2014). This low genetic diversity and structure challenges fine-scale population structure characterization in northern pike in North America (Miller & Senanan, 2003; Senanan & Kapuscinski, 2000; Skog et al., 2014). In general, currently three main northern pike haplogroups exist across the Northern Hemisphere (Skog et al., 2014). One of these haplogroups is Holarctic, and is present across North America as the likely

expansion from an Alaskan, postglacial source described above (Wooller et al., 2015). Even with low genetic diversity, and involving only a few individuals, northern pike excel at accessing and colonizing new regions (Luan et al., 2021). Considering this historical context, and in addition to its value in sports fishing (DFO, 2012) and as a model for physiological, toxicological, and ecological studies (Forsman et al., 2015), the northern pike is a good model system to understand adaptability and resilience with low genetic diversity.

Northern pike has also been the focus of studies on mechanisms of sex determination. Pan and co-workers (2021) characterize global dynamics of the master sex determination (MSD) system in northern pike, where a male-specific duplicate of *anti-mullerian hormone*, termed *amhby*, was identified as the MSD gene in Europe, but has been lost in parts of North America. Furthermore, no replacement MSD system for *amhby*(-) populations in North America has been identified (Pan et al., 2021), which explains previous challenges in mapping sex in northern pike populations of eastern North America (see Rondeau et al., 2014). This MSD system variation is notable given that it is considered a single, circumpolar species (Grande et al., 2004). Teleosts are known to have high diversity and turnover of sex determination systems (Pan et al., 2019), and their sex determination can involve environmental factors (Devlin & Nagahama, 2002; Goto-Kazeto et al., 2006). As *amhby* is the MSD gene in Alaska, and is missing in other North America populations, this may indicate that it was lost during range expansion through bottleneck or founder effects from Alaska eastward (Pan et al., 2021). This evolving MSD system in northern pike adds to its value as a model species. Field and laboratory observations of skewed sex ratios in some North American populations do suggest an environmental effect on sex determination (Carbine, 1942; Clark, 1950; Huffman et al., 2014; Priegel & Krohn, 1975) although field-based estimates of sex ratio may contain biases (Casselman, 1975). Finer-scale resolution of populations in North America will benefit the understanding of *amhby* loss and may provide more information to understand current or evolving mechanisms underlying sex determination in *amhby*(-) populations.

Genetic resources are available for northern pike including reference genome assemblies, a linkage map, and expressed sequence tag libraries (Leong et al., 2010; Pan et al., 2019; Rondeau et al., 2014), as well as genome assemblies of other *Esox*, *Dallia*, *Novumbra*, and *Umbra* spp. (Pan et al., 2021), further adding to the value of northern pike as a model. Short-read technology has provided much insight in ecology and evolution, but long-read technology such as PacBio and Oxford Nanopore (Eid et al., 2009; Stoddart et al., 2009) can provide a more contiguous assembly that increases the potential to fully characterize genes, repeat regions, and chromosomal structural elements (reviewed by Chaisson et al., 2015). This is particularly relevant for areas of high repeat content (e.g., Bongartz,

2019) or for polyploid species in order to resolve haplotypes (e.g., Aury et al., 2022; Sun et al., 2022; Yuan et al., 2021). Long reads require error-correction with short reads (e.g., Goodwin et al., 2015; Koren et al., 2012) or consensus building with sequencing depth (Chin et al., 2013). Furthermore, new advances in scaffolding can facilitate further improvements in contiguity for super scaffolds and long-range haplotypes including chromatin conformation capture (Dekker et al., 2002) as applied in chromatin proximity ligation methods (Dudchenko et al., 2017; Mostovoy et al., 2016; Putnam et al., 2016).

In the present study we provide a new genome assembly for northern pike using long-read technology and chromatin conformation capture methods, using naming conventions consistent with the original assembly (Rondeau et al., 2014). We then use the improved assembly to characterize the genomic diversity of northern pike throughout North America using whole-genome resequencing with a particular focus on the region from Alaska through British Columbia due to its importance regarding the loss of *amhby* (Pan et al., 2021). Furthermore, we characterize the intrachromosomal variation in polymorphic and repetitive element content. Alongside this characterization of genomic variation, we use recently developed (Pan et al., 2021) and newly developed sex markers combined with histology to identify which populations have *amhby*(+) males. Collectively, this work improves our understanding of the loss of genomic variation and MSD function in a west-to-east pattern in North America, likely related to the expansion and recolonization of northern pike into North America.

## Methods

*Reference genome sampling, sequencing, and assembly*

Northern pike were sampled near Castlegar, BC in September 2017 as part of an invasive species removal project in the Canadian portion of the Columbia River. Following euthanization, the liver tissue from a female individual was removed by dissection and frozen on dry ice for 48 hours before long-term storage at -80ºC.

High-molecular weight DNA was extracted from the liver using a modified dialysis method as follows. Approximately 550 mg of frozen tissue was ground into a powder using liquid nitrogen and mortar and pestle. The powder was transferred to a 5 ml lo-bind tube (Eppendorf), along with 3600 µl buffer ATL, 400 µl proteinase K solution and 40 µl RNAse A solution (QIAGEN), followed by digestion at 56ºC for 3 hours, with rotation at approximately 4 RPM. The homogenate was split equally into two 5 ml tubes, where a phenol-chlorform-isoamyl alcohol (25:24:1) purification was performed three times per tube, followed by a chloroform-isoamyl alcohol (24:1) purification. In each stage, one

146 volume of organic solvent was mixed with one volume of aqueous solution, mixed by slow inversion
147 for three minutes, then spun for 15 minutes at 5000xg for phase separation. The aqueous top layer was
148 then transferred slowly to a new tube using a 1000 ul wide-bore pipette tip. Subsequently, 2 µl of
149 RNAse A (20 mg/ml; QIAGEN) was added to the aqueous solution and incubated for one hour at room
150 temperature, then 5 µl of proteinase K (20 mg/ml) was added and incubated overnight at 4ºC. Finally,
151 approximately 750 µl was obtained from each tube (i.e., 1,500 µl total) and transferred to a Spectra/Por
152 Float-A-Lyzer G2 1000 kD (pink) dialysis device (Spectra/Por). Dialysis was performed in 1 L of 10
153 mM Tris-Cl (pH 8.5) at 4ºC with gentle mixing for one week, changing the buffer five times throughout
154 the process. DNA quantity was determined by Qubit v2.0 (Life Technologies) and quality by 0.6%
155 agarose gel electrophoresis at 60 Volts. Bands greatly exceeded the largest ladder band of 40 kb with
156 no visible shearing.

157 Subsequent library preparation and sequencing was performed by the McGill University and
158 Genome Quebec Innovation Centre. In brief, PacBio sheared large-insert libraries were constructed
159 following standard protocols and sequenced across eight SMRT cells on a PacBio Sequel, generating
160 76 Gbp of data. An additional library was constructed using the same input genomic DNA for 10X
161 chromium sequencing following standard protocols and sequenced on one lane of Illumina HiSeqX
162 PE150. A third library preparation method was conducted at the University of Victoria using the
163 Proximo Animal Hi-C kit (Phase Genomics) following the Phase Genomics protocol 1.0 with adapter
164 barcode N702 using 0.2 g of the liver tissue and sequenced within a single lane of Illumina HiSeq4000
165 PE100.

166 PacBio data was assembled using Canu v1.8 (Koren et al., 2017). All subreads were used, and
167 a genome size of 0.95 Gbp was estimated as input. All stages were run with SLURM scheduling on
168 the Compute Canada heterogeneous cluster (Cedar). Options *stageDirectory* and
169 *gridEngineStageOption* were used for optimal on-node storage at heavy input/output stages. Default
170 settings were used except that *ovlMerThreshold* was set to 2000, *corMhapSensitivity* was set to normal,
171 *correctedErrorRate* to 0.085, and *minReadLength* to 2500 to reduce runtime and/or to use
172 recommendations for Sequel data as per software guidelines. Following the initial assembly, Arrow
173 v2.2.2 was used in SMRTlink (6.0.0.47841) to polish with PacBio data, using the ArrowGrid wrapper
174 (Koren et al., 2017). The ArrowGrid pipeline was run a total of three times, then the output was
175 subjected to a round of polishing using Pilon v1.22 (Walker et al., 2014) using the 10X chromium data
176 that had been pre-processed using Longranger v2.2.2 (10X Genomics) to remove barcodes.

Scaffolding occurred in three stages. In the first stage, Hi-C data was used to scaffold the assembly with SALSA2 (Ghurye et al., 2017; Ghurye et al., 2019). Hi-C data was first prepared and aligned to the input genome following recommendations by Arima Genomics (Arima Genomics, 2023). In brief, paired-end reads from the Hi-C library were separately aligned to the genome using BWA mem 0.7.13-r1126 (Li, 2013), then sorted, merged and filtered with SAMtools v.1.8 (Li et al., 2009), Picard v.2.9.0-1-gf5b9f50 (Broad Institute, 2023), and bedtools v.2.27.0 (Quinlan & Hall, 2010). SALSA2 was then run with options *-m* yes and *-e* GATC to identify misassemblies in the input assembly and to indicate the applied restriction enzyme, respectively, and indicating the post-Pilon reference and the bed file output by the Arima pipeline.

In the second scaffolding stage, the Tigmint-ARCS-LINKS pipeline was applied. In brief, 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. For ARCS v1.0.5 (Yeo et al., 2017) and LINKS v1.8.6 (Warren et al., 2015), parameters were used as default except *-l* (i.e., the minimum number of links, or k-mer pairs, required to compute a scaffold; default = 5), and *-a* (i.e., the maximum link ratio between the two best contig pairs; default = 0.3), which were tested for all combinations of *l*=5-10, *a*=0.1-0.9. Optimization focused on increasing N50 but also balancing the number of misjoined scaffolds observed in the third stage of scaffolding with Hi-C (*see below*). The selected parameters (i.e., $l = 8$; $a = 0.2$) had a maximum of two or three visible misjoins. This stage introduced some short contigs (< 200 bp) assembled within scaffolds without meaningful gap sizes, and these were removed using *sed*, with all remaining gaps resized to strings of 100 Ns. The small number of scaffolds that were smaller than 1 kbp (i.e., fewer than 100 scaffolds) were removed using *fastx* of Bioawk (Li, 2017).

In the third scaffolding stage, Hi-C data was re-aligned to the genome following the scaffolding stages above using Juicer v.1.5.6 (Durand, Shamim, et al., 2016) with flags and arguments *-s* Sau3AI (i.e., the restriction enzyme applied) and *-y* to include the restriction site file, as well as *-S* early to indicate early exit of the program. The resultant output (i.e., merged_nodups.txt) was used with 3d-dna v.180922 (Dudchenko et al., 2017) using parameters *-i* 50000 (i.e., the minimum size of contigs to scaffold) and *-r* 0 (the number of iterative rounds for misjoin correction). The assembly was visualized after scaffolding using Juicebox v1.8.8 (Durand, Robinson, et al., 2016) to identify and split mis-assemblies, and to identify, order, and orient linkage groups (LGs). LG-like groups were oriented such that the greatest density of inter-chromosomal contacts was at the 5' end of each LG. Using *juicebox_assembly_converter.py* (Phase Genomics), NCBI AGP files were generated. The newly generated LGs were compared to the northern pike v.1.0 assembly (Rondeau et al., 2014) using LastZ (Harris, 2007) visualized in Geneious (Kearse et al., 2012) using default parameters, and the AGP file

210 was manually edited to keep naming consistent to the v.1.0 assembly. The genome was then submitted
211 to NCBI (see *Data Availability*).
212 Genome annotation was conducted using the NCBI Eukaryotic Annotation pipeline v.8.2 under
213 annotation release 103. Genome completeness was evaluated using BUSCO v.4.0.2 with both
214 Actinopterygii and Vertebrata odb10 datasets (Seppey et al., 2019). Analysis of repeat content used
215 the same methods and custom repeat library as previously described (Rondeau et al., 2014).
216
217 *Population genomics: sampling, sequencing, and analysis*
218 Northern pike tissue samples (n = 47) were obtained from across Canada and the northern United States
219 as provided by collaborators and hatcheries (see Table 1; Figure 1). Fish from the upper St. Lawrence
220 River were collected and processed following a protocol approved by the SUNY ESF Institutional
221 Animal Care and Use Committee. Alaskan Minto Flats samples used for sex marker validation (*see*
222 *below*) were provided as fin clips in 95% ethanol by the Alaska Department of Fish and Game
223 following approved state and departmental regulations and protocols. All other tissues were either
224 archival, opportunistic sampling of fishery harvest or government purposes (e.g., invasive species
225 control), and therefore did not require ethical review by the University of Victoria in accordance with
226 the Canadian Council on Animal Care Guidelines on the care and use of fish in research, teaching and
227 testing (s4.1.2.2; CCAC, 2005).
228 DNA was extracted from a variety of tissues using DNEasy Blood and Tissue Kit (QIAGEN)
229 following the manufacturer's protocols, then quantified by Nanodrop ND-1000 (ThermoFisher) and
230 Qubit v2.0 (Life Technologies). Samples were sent to McGill University and Genome Quebec
231 Innovation Centre for library preparation and sequencing. Most of the samples (i.e., 35 of 47)
232 underwent PCR-free whole genome shotgun sequencing. The ten samples from the Chatanika River
233 had insufficient quantities for PCR-free libraries and therefore were sequenced via PCR shotgun
234 sequencing. All 45 of these libraries were sequenced on an Illumina HiSeqX Ten in paired-end 150 bp
235 reads with between five and seven samples sequenced per lane. The two remaining samples (i.e.,
236 Castlegar and Charlie Lake, BC) were used for reference genome assembly, with the Castlegar
237 individual being the focus for the final v.4.0 assembly. Due to the focus on detecting sex determination
238 loci, lanes were designated as sex specific as much as possible to reduce the impact of potential index
239 switching between sexes.
240 Read processing and variant calling was based on GATK's best practices (GATK v.3.8-0-
241 ge9d806836; DePristo et al., 2011; McKenna et al., 2010; Poplin et al., 2018; Van der Auwera et al.,

2013). In brief, paired-end reads were aligned to the latest northern pike genome (*described above*; GCF_004634155.1) using bwa-mem v.0.7.13-r1126 (Li, 2013). Alignment files were converted to bam format, sorted, and indexed by position using SAMtools v.1.3. Sequencing platform and multiplexing layout metadata was used to mark duplicates using Picard v.2.17.11 to flag for downstream genotyping. The two samples used for reference genome assemblies (i.e., Castlegar and Charlie Lake, BC) had 5-7x deeper coverage than the rest of the samples, and so they were down-sampled (i.e., reads were randomly subsampled to the targeted coverage of 25X for Charlie Lake and 20X for Castlegar) using SAMtools view with the *-s* flag. Nucleotides 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 per sample using the *HaplotypeCaller* of GATK in GVCF mode, then combined as a cohort using the *GenotypeGVCF* function of GATK to produce a VCF file containing 1,910,789 SNPs and insertion/deletions (indels) for all 47 samples. The 1,363,731 SNPs within this VCF were then extracted and filtered (see Table 2). A hard filter was applied using GATK to remove variants according to the following cutoffs or parameters: 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. Additional quality control filters were applied using VCFtools v.0.1.15 (Danecek et al., 2011) including removing sites where more than 10 individuals had missing data (*--max-missing-count* 10) or those with no minor alleles observed (*--mac* 1). Sites were further required to have at least one homozygous call, either reference or alternate. The resultant VCF file with 1,127,943 variants was used for the downstream analysis, although further filtration was conducted for specific analyses as discussed in individual sections below.

The numbers of variant calls per individual were summarized and visualized in R (R Core Team, 2023) using vcfR v.1.14.0 (Knaus & Grünwald, 2017) and dplyr v.1.1.4 (Wickham et al., 2023). Genotype counts per site were obtained using the *VariantsToTable* function of GATK for heterozygotes, homozygous reference, homozygous variant, and no call, as well as the total number of variants and samples called per site. To obtain per site genotype frequencies, each category was divided by the number of samples called. Observed heterozygosity ($H_{OBS}$) was calculated by reading in the VCF into R, converting it to genind format using vcfR then to genlight format to calculate $H_{OBS}$ with the function *gl.report.heterozygosity* using dartR v.2.9.7 (Gruber et al., 2018). Private alleles per population were calculated using the *private_alleles* function of poppr v.2.9.4 (Kamvar et al., 2014). Tajima's D in bin sizes of 10,000 were calculated using bcftools (Danecek et al., 2021).

A maximum likelihood tree based on genome-wide SNPs was generated using SNPhylo v.20140701 (Lee et al., 2014) using default parameters and with 1,000 bootstraps. The resulting tree was visualized using FigTree version 1.4.3 (Rambaut, 2016), and rooted by midpoint. A discriminant analysis of principal components (DAPC) was performed with bi-allelic genome-wide SNPs using adegenet v.2.1.1 (Jombart, 2008; Jombart & Ahmed, 2011) in R. The *find.clusters* function of adegenet was used to determine the appropriate number of clusters based on the lowest Bayesian Information Criterion value when all principal components were kept, and DAPC was conducted on these groups, retaining 24 principal components and three discriminant functions. The *snpzip* function with the Ward clustering method was used to list SNPs with the greatest contribution to each of the three discriminant axes identified.

*Characterizing diversity across chromosomes*

The genome assembly was unwrapped using custom python script *fasta_unwrap.py* (E. Normandeau, Scripts; see *Data Availability*) and subset to only contain chromosomes. Chromosome lengths were calculated using *fasta_lengths.py* (E. Normandeau, Scripts). The genome was then indexed using samtools, and a bed file containing 1 Mbp windows was prepared using the *makewindows* function of bedtools. The number of variants per window for each chromosome was then calculated using the *coverage* function of bedtools. The distribution of variants per window for all windows was then used to determine the minimum number that would be considered an outlier (i.e., third quartile + 1.5x the interquartile range) using the *boxplot.stats* function of R, and the number of variants per window and the mean number of variants per kbp per chromosome were plotted using custom scripts in R (see *Data Availability*).

*Empirical sex-linked variants and genotypes in resequencing data*

The identification of k-mers associated with sex was performed on the resequenced populations that had at least three females and three males (i.e., Chatanika River (CHT), Manitoba (WHI), New Jersey (HCK), and New York (SLA); see Table 1). Raw reads were concatenated and then all possible 31mers were extracted to form a master list of k-mers using Jellyfish v.2.2.6 (Marçais & Kingsford, 2011) running on Compute Canada. The master list was then used to query the reads of each individual sample using Jellyfish to count all present 31mers. A sex-specific analysis was then conducted whereby females and males were compared within each population. Sex-specific k-mers were considered as those sequences for which all of the opposite sex in the population had two or fewer copies (to allow

306 for sequencing errors), and the target sex individuals all had at least one k-mer and the sum of all males
307 had more than seven instances of the k-mer. The resulting sex-specific k-mer sets for each population
308 were mapped back to the reference genome using bwa-aln v.0.7.13-r1226 (Li, 2013). Following the
309 alignment, the number of sex-specific k-mers per population were counted in 10 kb windows across
310 the genome using bedtools v2.26.0 (Quinlan & Hall, 2010) and plotted in R v.3.5.3 using ggplot2
311 (Wickham, 2016).

312 Second, a genome-wide association analysis was performed using sex as the phenotype. The
313 individuals with sex phenotype data were extracted from the VCF. This extraction also included a filter
314 to keep variants with a minor allele count of at least two (i.e., *--mac* 2), resulting in the analysis of 17
315 females and 21 males. The analysis was performed using plink v.1.9b_5.2-x86_64 (Purcell et al., 2007)
316 using the *fisher-midp* option to run Fisher's exact tests and output p-values. The -log(p-value) of each
317 SNP was visualized in a Manhattan plot using qqman v.0.1.4 (Turner, 2014). Significant associations
318 were considered when Bonferonni corrected $p \leq 0.05$.

319 Third, a DAPC was conducted using sex as the differentiating variable using all resequenced
320 individuals with sex phenotypes within populations and defined groups from PCA and DAPC (*see*
321 *above*) in adegenet. Population and group-specific SNPs were extracted from the VCF file. Each group
322 was filtered independently using a custom script in R to remove variants that were not expected to be
323 related to the sex determination system (i.e., homozygous alternate variants). These variants were
324 removed because the reference genome is female and the sex determination system expected to be XY,
325 and therefore the homozygous alternate variants were not expected to be part of this system. The
326 remaining sites were those with homozygous reference and heterozygous genotypes. Subsequently, the
327 DAPC analysis was performed on each subset. Output loadings tables were inspected for genotypes
328 where all males or all females were heterozygous. From this information, lists of genomic locations
329 with putatively sex-specific signatures were generated for each population. Histograms of sex-specific
330 SNP occurrences were then generated along LGs using ggplot2.
331

332 *Sex markers and histology*

333 PCR-based inspection for genetic sex markers was conducted for individuals used in resequencing (*see*
334 *above*), as well as additional samples obtained from Castlegar, BC (n = 2 females and 5 males) and
335 from the Minto Flat region of Alaska, USA (n = 6 females, 14 males, 1 undetermined). PCR used
336 primers designed to amplify regions of *amhby* (i.e., SeqAMH1-4 and ConserveAMH1-1; Pan et al.,
337 2021), as well as newly developed primers from this study (i.e., set_24.5; Table S1). The set_24.5

338 primers amplify a 500 bp amplicon in both sexes, and an additional 250-300 bp amplicon specifically
339 in genetic males using a nested primer pair based on male-specific SNPs. PCR was conducted using 2
340 mM $MgCl_2$, 0.2 mM dNTPs, 1X reaction buffer, 0.25 units of Taq polymerase (Promega), 0.5 µM of
341 each primer, 50-100 ng of DNA, a remaining volume of molecular grade water up to 10 µl total
342 reaction volume. The nested primer pair used 0.2 µM forward primer instead of 0.5 µM and
343 included 0.3 µM of the probe. All PCR used the following thermal regime: 95˚C for 5 min., 35
344 cycles of 95˚C for 0.5 min., 0.5 min. anneal, 72˚C extension for 1 min. (or 0.5 min. for set_24.5),
345 and a final 10 min. extension at 72˚C. PCR products were visualized on 1% agarose gels and scored
346 as female, male, or undetermined.

347 Histological determination of phenotypic sex was conducted using gonadal tissue from the
348 Minto Flats and the St. Lawrence River populations. In brief, Minto Flats samples were saturated with
349 a 1:1 solution of 100% ethanol:LR White resin (hard grade) for 24 hours, then with pure LR White for
350 24 hours (resin replaced fresh after 6 hours). Each sample was then placed in separate gelatin capsules
351 with fresh catalyzed LR white and polymerized at 60˚C for 24 hours. Samples were cut to 1-micron
352 sections, stained with Richardson's LM stain, and examined microscopically. St. Lawrence River
353 samples were preserved in Davidson's solution for 48 hours, transferred to 70% ethanol prior to
354 histological processing as described above. Following processing, tissues were embedded in paraffin
355 wax and 5-micron sections were cut and stained with hematoxylin and eosin (H&E).

356

357 # Results

358 ## Genome assembly and annotation

359 A northern pike female was sampled from the Canadian portion of the Columbia River at Castlegar,
360 British Columbia (BC) and DNA was extracted from liver tissue to preserve high molecular weight
361 genomic DNA. The DNA from this individual was sequenced to approximately 80X depth using 8
362 SMRT cells on a Pacific Biosciences Sequel instrument. The initial contig-level assembly generated
363 by Canu (see *Methods*) yielded a total length of 939.0 Mbp in 1,258 contigs (contig N50: 3.9 Mbp).
364 Polishing and scaffolding occurred in several stages as described in brief here (see detailed description
365 in *Methods*). First, Hi-C data was applied to the assembly resulting in 941 scaffolds (scaffold N50:
366 18.8 Mbp). Second, 10X Chromium data was applied to the assembly, which split problematic contigs
367 and allowed for additional scaffolding. The error correction step increased the number of contigs to
368 1,395 and reduced contig N50 (contig N50: 3.4 Mbp), but the additional scaffolding increased the
369 scaffold N50 (scaffold N50: 23.3 Mbp). Third, the Hi-C data was further applied and manually

370 reviewed to yield a total of 811 scaffolds (final scaffold N50: 37.6 Mbp; Figure S1). This final northern
371 pike assembly (v.4.0) is 941 Mbp in length, with a maximum scaffold length of 52.6 Mbp, and includes
372 25 scaffolds of the expected chromosome lengths. In addition, the assembly contains 785 unplaced
373 scaffolds with a total length of 23 Mbp.
374 The 25 chromosome-length scaffolds were then assigned to the 25 linkage groups of northern
375 pike (Rondeau et al., 2014). Linkage groups were oriented by density of inter-chromosomal contacts,
376 where the greatest density of repeat elements was oriented to the 3' end of the chromosomes (Figure
377 S2). This fits with the original orientation by Rondeau and co-workers (2014), but should be noted that
378 the centromeres for the linkage map and assembly are therefore positioned at the 3' ends of all LGs or
379 chromosomes. A full inventory of observed repeat elements and their percentages is shown in Table
380 S2. The mitochondrial genome was identified by BLAST alignment of the previous assembly
381 (NC_025992.1; Rondeau et al., 2014) against the present assembly. Once identified, the mitochondrial
382 genome was manually circularized. The v.4.0 assembly has been submitted to NCBI under assembly
383 accession GCA_004634155.1 (see *Data Availability*).
384 Annotation of the assembly was performed by the NCBI eukaryotic genome annotation
385 pipeline using RNA-sequencing data (Pan et al., 2019; Rondeau et al., 2014) and EST data (Leong et
386 al., 2010). This identified 24,843 protein-coding genes. A BUSCO analysis found that 95.4% and
387 97.1% of the Actinopterygii (odb10) and Vertebrate (odb10) genes were complete, respectively.
388 Relative to the former database with 3640 genes, 94.1% (i.e., 3426) were identified as single copies,
389 and only 1.3%, 1.3%, and 3.3% were duplicated, fragmented, or missing, respectively.
390

391 **Genetic variation and population genomics**

392 Northern pike were obtained for whole-genome resequencing from across North America (Figure 1;
393 Table 1). Samples included those from Chatanika River (CHT) in Interior Alaska, USA (n = 10), Yukon
394 River at Hootalinqua (HOO) in the Whitehorse Region of the Yukon Territories, Canada (n = 5),
395 Palmer Lake (PAL) in northwestern British Columbia (BC; n = 4), Charlie Lake (CHA) in northeastern
396 BC (n = 1), Columbia River at Castlegar (CAS) in southeastern BC (n = 1), Whiteshell Hatchery (WHI)
397 in Manitoba (n = 6), the St. Lawrence River (SLA) off New York State (n = 11), and at Hackettstown
398 Hatchery (HCK) of New Jersey, USA (n = 9). In total, this included 17 phenotypic females, 21 males,
399 and nine individuals with undetermined sex (n = 47 total; Table 1). The individual sampled at Charlie
400 Lake and the individual sampled at Castlegar were used for whole-genome sequencing, and the
401 remainder of the samples were used for whole-genome resequencing (see *Data Availability*). The

402 Charlie Lake assembly was not used in this study, but the sample's data was used in the rest of the
403 analyses as a representative for this collection site.

404 In total, 1,363,731 raw SNP variants were called in the whole-genome resequencing data, and
405 after all quality filtering, 1,127,943 SNPs were retained (i.e., 82.7% retention; see Table 2). On
406 average, the mean (± s.d.) depth per individual per variant site was 20 ± 3.3x (range = 15x to 30x).
407 There was a large difference in genetic variation observed across sampling sites (Figure 1B; Table 3).
408 Interior Alaska samples (CHT) had the highest rate of heterozygous genotypes (mean per individual =
409 180,943 SNPs), followed by the Yukon River (HOO; 126,720 SNPs). All other populations had fewer
410 heterozygous SNPs, with 58,129 heterozygous SNPs in northwestern BC (PAL), 49,401 in Manitoba
411 (WHI), 64,394 in New York (SLA), and 67,944 in New Jersey (HCK). Additionally, Alaska, Yukon,
412 and northwestern BC had similarly high numbers of homozygous alternate genotypes (mean per
413 individual: 116,276, 84,775, and 101,465, respectively), and the eastern BC and eastern Canada/US
414 samples had similarly lower numbers of homozygous alternate genotypes ranging from 18,525 to
415 31,385 on average per individual (Additional File S1). Considering the ungapped genome length of
416 940.8 Mbp, and including both heterozygous and homozygous alternate variants, the CHT and SLA
417 populations have an average per individual difference from the reference genome of 0.0316% and
418 0.00891%, respectively. In terms of the average SNP per kbp per individual, CHT and SLA have a
419 SNP for every 3,165 bp and 11,222 bp, respectively. The average observed heterozygosity ($H_{OBS}$) per
420 population followed a similar trend as heterozygous SNP counts, with $H_{OBS}$ of CHT and HOO
421 estimated at 0.161 and 0.113, respectively, whereas PAL, WHI, SLA, and HCK ranged from 0.044-
422 0.060. More pronounced negative Tajima's D values were observed in WHI and SLA, with values of
423 -1.07 and -1.30, respectively, than HCK (-0.33) or the higher diversity populations CHT and HOO (-
424 0.22 and -0.331, respectively; Table 3). Results for the populations with only a single individual should
425 be taken with caution given the low sample size, different technology applied to sequencing, and for
426 the Castlegar sample (CAS), being the reference genome used in the study.

427 Genetic similarity between individuals and populations based on clustering in a genetic
428 dendrogram followed a similar trend to that observed in overall variation (i.e., considering the sum of
429 both heterozygous and homozygous alternate genotypes), where the Alaska (CHT), Yukon (HOO),
430 and northwestern BC (PAL) populations clustered separately from all other samples (Figure 2). Most
431 of the populations had very strong support (100%) for grouping all individuals in the population
432 together, and only the Yukon (HOO) population did not show a highly supported single cluster. Within

433 each separate section of the dendrogram for high or low diversity samples, the closest populations
434 geographically were also the most proximal in the dendrogram.

435 An analysis of private alleles and fixed differences further indicates the genetic separation and
436 isolation of the different populations. Of the 1.1 M variants characterized, 0.75 M were only present in
437 a single population (i.e., private alleles), and although many are of low frequency, each population
438 with at least four individuals had between 21,290-214,742 private alleles present. As an example, CHT
439 has 81,263 private alleles with at least four instances of the allele being observed, whereas HOO has
440 5,395, HCK has 26,173 and SLA has 5,989 private alleles with at least four instances of the allele.
441 Furthermore, when considering fixed homozygous alternate genotypes, a large proportion of these
442 were unique to each population. For example, there were 34,561 biallelic variants that were observed
443 as only homozygous alternate genotypes in Alaska (CHT), and of these, 34,002 were fixed reference
444 in St. Lawrence River (SLA) and 33,285 were fixed reference in the Yukon (HOO). As another
445 example, there were 13,777 biallelic variants observed as only homozygous alternate genotypes in
446 SLA, and of these, 10,975 and 9,993 were observed as fixed homozygous reference in the
447 Hackettstown (HCK) or Alaska (CHT) populations, respectively. Therefore, many variants identified
448 here were specific to one of the populations, and many variants were fixed in specific populations,
449 even in geographically proximal populations.

450 The distribution of variants was not equal across the northern pike genome (Figure 3A). When
451 considering the sum of variants within 1 Mbp windows across chromosomes (n = 930 windows total),
452 59 windows were considered outliers in terms of elevated polymorphism (see *Methods*) with more than
453 2,341 variants per window in comparison with the genome-wide average of 906 variants per window
454 (median: 1,202; Additional File S2). Several of the chromosomes have large sections of elevated
455 variation where multiple consecutive windows are outliers, specifically chromosomes 5 (3' end, 24-29
456 Mbp), 9 (5' end, 4-10 Mbp), 11 (3' end, 35-44 Mbp, excluding 41-42 Mbp), and 24 (throughout, 0-5
457 Mbp excluding 3-4 Mbp, and 18-23 Mbp; Figure 3A). Furthermore, when considering a per-
458 chromosome average number of variants per kbp, chromosome 24 has a clear elevation overall in terms
459 of polymorphism level relative to all other chromosomes (Figure 3B). Notably, in populations that still
460 use *amhby* as the sex determining gene, this gene is in chromosome 24. Some preliminary exploration
461 of elevated polymorphic regions suggests presence of genes related to immunity or having multiple
462 copy numbers, for example chromosome 09 contains zinc-finger genes, immune-associated nucleotide-
463 binding protein genes (GIMAPs), as well as multiple genes of the major histocompatibility complex
464 (MHC) class II type, but it is beyond the scope of this paper to describe all of the regions of the northern
465 pike genome with elevated SNP density.

466

467 **Sex determination**

468 The dataset was restricted to only include individuals with phenotypic sex recorded, and following
469 filters, there were 672,565 SNPs retained across 17 females and 21 males. All individuals were initially
470 analyzed together (i.e., on a continental scale) using three different approaches to search for a sex-
471 specific locus in the data. Both a genome-wide association study (GWAS) and a DAPC with sex as the
472 separating variable were performed but found no significant or suggestive associations (*data not*
473 *shown*). Additionally, a k-mer analysis of raw sequence reads and therefore operating independently
474 of a reference genome also failed to identify sex-specific polymorphism at this scale. The reference-
475 independent approach was valuable given that the reference genome used to score SNPs was developed
476 from a female individual, and therefore the male-specific sex determining locus would be expected to
477 be missing (Pan et al., 2021).

478 After the continental-scale analyses failed to find sex-associated loci, population-specific
479 analyses were conducted. With this focused analysis, a sex-specific signal was detected in the Alaskan
480 sample (CHT; n = 5 individuals per sex), where male-specific heterozygosity was observed on LG24
481 through both the DAPC and k-mer analyses (Figure 4), as previously observed (Pan et al., 2019). In
482 the CHT samples, a 500 kb region between 650-1150 kb of LG24 (total chromosome length: 29.5 Mbp)
483 was detected, containing 3,552 male-specific SNPs. In addition, 1,137 male-specific and 39 female-
484 specific SNPs from other sections of the genome were also identified by DAPC analysis (Additional
485 File S3). Although none of the Yukon River (HOO) samples had associated phenotypic sex
486 information, one of the five individuals also held the male-specific SNPs on LG24 (and elsewhere)
487 identified from the Alaska population on LG24. However, all other populations assessed in North
488 America did not hold these sex-specific variants (Figure 4).

489 To further characterize the genetic sex of the sampled individuals across North America, since
490 the reference genome is generated from a female and therefore is not expected to have the male-specific
491 sex-determining locus *amhby*, the presence of *amhby* was investigated using *amhby*-specific primers
492 in a PCR analysis (Pan et al., 2021). This confirmed that *amhby* was present in the phenotypic males
493 from Alaska (CHA) and the putative male from the Yukon (HOO) as observed in the sequence data.
494 None of the other samples in any collection used for WG-resequencing in North America amplified
495 *amhby*. The PCR results therefore agree with the lack of a sex-specific signal in the other populations
496 by GWAS, DAPC, or k-mer analyses (*see above*).

Additional sampling was undertaken to further characterize the geographic regions in which the transition from *amhby*(+)-males to *amhby*(-)-males occurs. Two additional females and three males were obtained from the Columbia River at Castlegar, BC (CAS), and six females and 14 males were obtained from Minto Flats, Alaska. These samples were tested for *amhby* by PCR markers (Pan et al., 2021), as above. Interestingly, all three of the Castlegar males were positive for *amhby* (females were negative); this contradicts the expectation of the absence of *amhby* in males in this region. Also unexpectedly, the Minto Flats males, expected to all be *amhby*(+), did not show a uniform positive detection with only six of the 14 phenotypic males with a positive *amhby* detection (the females were all negative). All three primer pairs, including *amhby_conserve* (partial exon 2) and *seqAMH_1* (partial exon 7) from Pan et al. (2021), and primer pair set_24.5 (LG24: 996,878-997,339 bp) found concordant results for these samples. Sex phenotypes were further evaluated through gonadal histology, which confired testes development in both *amhby*(+) and *amhby*(-) males (Figure 5). Therefore, there are both *amhby*(+) and *amhby*(-) males in Alaska, and *amhby*(+) males further southeast than expected in North America.

# Discussion

## Genetic diversity across North America

In all the populations assessed across North America from Alaska through New Jersey, the northern pike from the west in Interior Alaska and the Yukon River have the highest genomic diversity. A striking decrease in overall diversity coincides with the location of the North American Continental Divide (NACD; i.e., the separating point between eastward or westward-flowing watersheds). West of the NACD, a stepwise decrease in diversity occurs from Interior Alaska to the Yukon, then to northwestern British Columbia, as measured by observed heterozygosity and the total number of heterozygous variants. East of the NACD, the populations have uniformly low diversity. The eastern region samples at the Manitoba hatchery (Whiteshell) and the St. Lawrence River had more pronounced negative Tajima's D values, suggesting that these populations were reduced in number and then more recently expanded, resulting in an excess of low frequency polymorphism (Hedrick, 2005; Tajima, 1989). This was not universally observed in the eastern samples (e.g., Hackettstown had a similar Tajima's D value as western sites). The NACD may not be causal to the diversity decrease, although it is an important land formation that influences contemporary water catchments and flow, but it does serve as a marker beyond which all populations surveyed to date have been found to have universally low diversity. The west to east diversity decrease is in agreement with theory that the

western populations are the oldest in North America, and that while these populations expanded eastward from the Beringia refugium, for example while recolonizing post-glacial recession, they were impacted by population bottlenecks, founder effects, and possibly other diversity-reducing impacts (Crossman & Harington, 1970; Skog et al., 2014).

Even in the most diverse population in the dataset (i.e., Interior Alaska; CHT), genome-wide genetic diversity is very low; per individual, a SNP is observed on average every 3.2 kb. In the eastern populations (e.g., St. Lawrence River; SLA), this number reduces to a SNP on average every 11.2 kb. This is remarkably low when compared to marine fishes including Atlantic herring *Clupea harengus* (1 variant per 0.3 kb; Martinez Barrio et al., 2016) and Atlantic cod *Gadus morhua* (1 variant per 0.5 kb; Star et al., 2011), as well as freshwater fishes such as rainbow trout *Oncorhynchus mykiss* (1 variant per 0.75 kb; Gao et al., 2018). Furthermore, northern pike were estimated to be on average 0.0089% - 0.032% different from the reference genome. For comparison, humans *Homo sapiens*, known to contain low levels of genetic variation, are four to 13-times more genetically diverse than the northern pike as assessed here (i.e., 0.11% - 0.13% difference from the reference; Auton et al., 2015). However, this low level of genetic diversity does not preclude evolutionary successful strategies; northern pike are renowned for their ability to thrive in a variety of environments and to expand into and colonize new areas (Luan et al., 2021).

In the region that separates high and low diversity North American northern pike, local effects may be further impacting diversity levels. The BC population Palmer Lake (PAL) is from a lake off the large Atlin Lake, which eventually flows into the Yukon River. The Palmer Lake population has lower genetic variation than the Yukon River population (HOO), even though these two sites are relatively proximal and may have at one time been directly connected. However, the Palmer Lake population may have lost variation due to inbreeding and drift. The apparent very low diversity of the Columbia River sample (CAS), and the lower diversity of Charlie Lake sample (CHA) relative to the Palmer Lake population may be due to technical artefacts, since both samples only had a single sample characterized from the population and underwent different sequencing technology to characterize their variation. Furthermore, the CAS individual was the source of the reference genome assembly, which would further reduce the characterized variation given that there would be no homozygous alternate genotypes observed. However, the Columbia River population (CAS) is in fact invasive from east of the NACD (Carim et al., 2019), and therefore it may have low diversity relative to BC populations west of the NACD. Additional samples would be needed from these two populations to accurately place them in the context of other populations genotyped here. Furthermore, introductions of northern pike have occurred in BC waters (Harvey, 2009), which could impact spatial trends in unexpected

562 ways depending on the source population and size of the introduction. In conclusion, local impacts
563 such as isolated populations or translocation history may explain differences from the expected level
564 of diversity based on the broader regional location and proposed expansion history. The boundary
565 region between elevated diversity and low diversity east of the NACD provides an interesting area of
566 contrast that requires further characterization.

567 Although northern pike from east of the NACD had low genetic variation, in the present
568 analysis each population was still consistently clustering with members of its own population with
569 strong bootstrap support. Therefore, although previous studies have been unable to distinguish different
570 populations of northern pike in eastern North America (Miller & Senanan, 2003; Senanan &
571 Kapuscinski, 2000; Skog et al., 2014), in agreement with Ouellet-Cauchon et al. (2014) our results
572 suggest separability of populations in the low diversity collections. The high degree of private alleles
573 per population observed here is notable and may be explained by northern pike experiencing high
574 genetic drift and expansion from a few individuals. These high levels however should be further
575 investigated once more samples are included per population. Similar to distinct groupings observed
576 here, although low nucleotide diversity was observed in eastern Russia, high genetic differentiation
577 was observed in populations sharing a common origin, likely due to founder effects (Bachevskaja et
578 al., 2019). Less differentiation may occur in central Europe. For example, in Germany, erosion of
579 population substructure occurs alongside habitat degradation, which may be due to eutrophication and
580 habitat loss leading to natural dispersal (Eschbach et al., 2021). The loss of differentiation also occurs
581 by secondary contact of separate populations through human-mediated stocking, and diversity erosion
582 was observed to be most severe in rivers or in modified water bodies (Eschbach et al., 2021). Given
583 the well-supported clustering observed here across different populations in North America, population
584 management may be feasible as reported for the congener *Esox masquinongy* (Rougemont et al., 2019).
585 As an example, genetic markers have been used to determine the likely source of an introduced invasive
586 population of northern pike in eastern Washington State, finding that it was likely a human-mediated
587 transfer from disconnected populations in Idaho (Carim et al., 2022). To determine the ability to
588 perform stock identification in the populations identified here would require further analysis with
589 larger sample sizes to determine the expected power of genetic assignment tests (e.g., Moran &
590 Anderson, 2019).
591

592 **Genetic diversity across the genome**

593 Genetic diversity is not universally low across the northern pike genome, as certain regions are
594 enriched for polymorphism (e.g., regions of chromosomes 5, 9, 11, and 24). These chromosomes have
595 large regions of elevated diversity, which may indicate the presence of specific genes or high variability
596 regions. Although a full inspection of the regions has not been conducted, some regions hold immune-
597 related genes (e.g., major histocompatibility complex (MHC) region in chromosome 9). The MHC is
598 known to be highly variable in vertebrates, and its variability is valuable for presentation of a diverse
599 range of antigens for pathogen recognition (Piertney & Oliver, 2006; Unanue et al., 2016). Regions of
600 elevated diversity in the genomes of otherwise low diversity species have been identified previously.
601 For example, channel island foxes (*Urocyon littoralis*) have highest SNP density in olfactory receptor
602 genes (Robinson et al., 2018). Humans, also known for low genetic diversity, have an enrichment of
603 polymorphism around genes related to sensory perception and neurophysiology (Redon et al., 2006).
604 Enrichment of polymorphism in certain areas is also observed in species with generally higher overall
605 diversity; stickleback are known to have an enrichment of polymorphism around genes controlling
606 bony plate number and morphology (Nelson et al., 2019). Collectively, these observations suggest that
607 regions of high variation in the genome are associated with phenotypic variation that may be
608 instrumental to the survival of the species. The maintenance of genetic variation may therefore be most
609 crucial in genetic regions that affect the ability of a species to detect or respond to stimuli that are both
610 fundamental to survival and that fluctuate in their habitat. The enriched polymorphic regions identified
611 here in northern pike merit further investigation.

612 The observation of chromosome 24 having multiple separate regions of elevated diversity is in
613 contrast with single regions as observed in chromosomes 5, 9, and 11. Importantly, in the populations
614 that retain *amhby* in males, chromosome 24 contains the master sex determining gene (Pan et al., 2019).
615 Differentiation of sex chromosomes in northern pike has previously been determined to be generally
616 low (Pan et al., 2019). Lower levels of recombination between the X and Y chromosomes can occur,
617 which may result in degeneration of Y due to the accumulation of mutations (Charlesworth, 1991).
618 This can provide a benefit by sequestering sexually antagonistic alleles on sex chromosomes with
619 reduced recombination (Charlesworth & Charlesworth, 2005; Mackay, 2001), but this is not expected
620 in northern pike (Pan et al., 2019). In salmonids, for example, heterochiasmy occurs whereby a lack of
621 recombination in males occurs throughout the genome, with crossovers primarily occurring at
622 telomeric regions opposite to the centromere (Moen et al., 2004; Sakamoto et al., 2000; Sutherland et
623 al., 2017). Heterochiasmy also occurs in humans (Broman et al., 1998). However, any heterochiasmy

observed in northern pike appears to be low, as observed by Rondeau et al. (2014) that only a slight skew in recombination rates occurs between the sexes. The presence of multiple regions and generally elevated overall polymorphism enrichment in chromosome 24 merits further investigation, especially within populations that have lost the sex determining gene from this chromosome. With more populations and individuals characterized in the future, the precision of the location of these genomic regions or interpopulation variation that may exist will be valuable to be explored further.

**Sex determination and a patchwork of *amhby* presence in western North America**

Range expansion, if impacted by founder effects or population bottlenecks, may not only result in a loss of neutral genetic variation but also in a loss of functional genetic variation. The master sex determining gene *amhby* is considered to have been lost during expansion into North America (Pan et al., 2021), which coincides with a loss of overall genetic diversity (Skog et al., 2014), and our results are concordant with this finding. Northern pike can undergo sex reversal, possibly influenced by environmental signals (Pan et al., 2019) as has been observed in other teleosts (Rajendiran et al., 2021), as well as through disruptions of other genes in the pathway including *amh receptor II* (Pan et al., 2023). Therefore, as described by Pan et al. (2021), the founding of a population comprised of *amhby*(-) individuals (i.e., no genetic males) followed by sex reversal could presumably result in the establishment of populations without the master sex determining gene.

In the present study, *amhby*(-) populations are largely what was observed in populations southeast to the Yukon Territory. However, this observation was not universal, as there were also detections of Alaskan *amhby*(-) males and Columbia River *amhby*(+) males. North American northern pike provides a valuable model to study sex determination evolution; as Pan et al. (2021) describe that the old genetic sex determining system is lost before a new system has taken its place, and environmental sex determination may play a transition state for these populations. The observations here point to a mosaic of *amhby* presence and presumed activity in males in the boundary region from Alaska through British Columbia. East of British Columbia, however, there were no observations of *amhby*(+) males. It is unclear how the Castlegar males retained *amhby*, as they may have retained the gene from the source population in Alaska, or they may be descendants of another lineage that still uses the *amhby* system. Further characterization of populations in this area may further elucidate the interaction between an environmentally mediated sex determination system and the potential for emergence of a new genetic system for sex determination (Pan et al., 2021).

The lack of the *amhby* gene in eastern populations does not coincide with a lack of males in eastern Canada. However, it is notable that upper St. Lawrence River populations have experienced sex ratio shifts from male to female dominance in recent decades (Farrell & Barry, 2012). An earlier attempt by Rondeau and co-workers (*unpublished*) to map the sex determining locus using five biparental crosses of St. Lawrence region northern pike where phenotypic sex was determined by histology resulted in three highly male-biased families (from 76-92% male) and two moderately male-biased families (from 56-65% male), and the inability to detect sex-linked markers was thought to be due to the strong sex bias combined with low polymorphism (Rondeau, *unpublished*). However, more recent analyses, as discussed above, suggest that this was due to the lack of a genetic sex determination system at play currently in these eastern populations. Additional analysis of these samples using the markers developed by Pan et al. (2021) also failed to detect *amhby*, and so the resequencing approach here was taken as reported in the present study.

**Genome assembly**

A new individual was used for the latest version of the assembly due to decreasing quantity and quality of any remaining material used for earlier versions. Regardless, with the significant advances that occurred in sequencing and assembly, the latest assembly was superior in metrics and utility relative to earlier assemblies. In human medical, agricultural, or even now conservation and population genomic fields, there are an increasing number of assemblies for the same species, with researchers generating references specific to their population; this can have benefits in terms of alignment success and variant calling (Thorburn et al., 2023), but also needs to be considered in terms of detecting potential reference bias, in particular when populations are highly divergent (Bohling, 2020). In northern pike, there now exists both the North American female assembly from the Castlegar population of the Columbia River (presented here), as well as a high quality European male assembly (Pan et al., 2019) that was optimal for characterizing the sex determination region within the genome. Future additions of other assemblies from Asian or other lineages may eventually produce a pan-genome for the species that represents the diversity of this Holarctic species.

## Conclusions

In the present work we generate a long-read assembly with chromosome-length scaffolds, and use this resource to characterize population genetic trends at a genome-wide scale, as well as genomic diversity trends within and between chromosomes. By combining sex-specific analyses of whole-genome resequencing data with PCR-based assays for sex markers, we were able to identify a patchwork of populations with or without the male-based sex determination system *amhby*. This patchwork, so far including *amhby*(+) and *amhby*(-) males in Alaska and *amhby*(+) males in the Columbia River, points to a more complex landscape of the loss of this sex determining system that occurred from west to east in North America. The large number of private alleles and clear clustering of individuals within each population points to substantial drift in individual populations. Population distinctiveness would be expected to enable genetic stock identification of different populations, although this remains largely untested. The clear trends of decreased diversity (and loss of functional genetic sex determination gene) with distance from Alaska is in agreement with the theory that northern pike expanded outward from Alaska after the deglaciation of the last ice age, and that significant founder effects impacted populations in the recolonized range. Regional effects in population diversity were observed as well, and from evidence so far, the North American Continental Divide provides a marking point after which diversity is uniformly low. Additional samples from more southern regions including the Columbia River would be beneficial to understand these post-glacial trends.

## Acknowledgements

We are grateful to the following people and organizations for providing samples for this study: Charles O. Hayford State Fish Hatchery (Hackettstown, NJ, USA); the Alaska Department of Fish and Game; the Manitoba Fisheries Branch; Whiteshell fish Hatchery (West Hawk Lake, Manitoba); Jeremy Baxter of Mountain Water Research; and Marco Marrello of Terraquatic Resource Management. We thank Dr. Miya Qiaowei Pan and Dr. Yann Guiguen for collaboration and insights supporting the genetic sex analyses in this study. Thanks to Dr. Sergio Cortez Ghio of Ex Machina Biostats for discussions on identifying regions of polymorphism enrichment. 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.

## Competing Interests

Ben Sutherland is affiliated with Sutherland Bioinformatics. The author has no competing financial interests to declare. The other authors declare that no competing interests exist.

## Data Availability

The reference genome v.4.0 is available through NCBI WGS accession SAXP01000000.1 and assembly accession GCA_004634155.1. Whole genome resequencing data is available through BioProject accession PRJNA512507 and SRA accessions SAMN10685075 – SAMN10685119.

The filtered VCF file with dataset SNPs from alignment to reference genome v4.0 is available on FigShare: https://doi.org/10.6084/m9.figshare.25230146

This work was supported by the following GitHub repositories:
General data analysis: https://github.com/bensutherland/ms_pike_popgen
Genome data analysis scripts: https://github.com/enormandeau/Scripts

## Author Contributions

Designed Research: Ben F Koop, Eric B Rondeau
Performed Research: Eric B Rondeau, Hollie A Johnson, Ben J G Sutherland, 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, Ben J G Sutherland, David R Minkley, Jong S Leong, Joanne Whitehead
Wrote the Paper: Hollie A Johnson, Eric B Rondeau, Ben J G Sutherland

# Figures

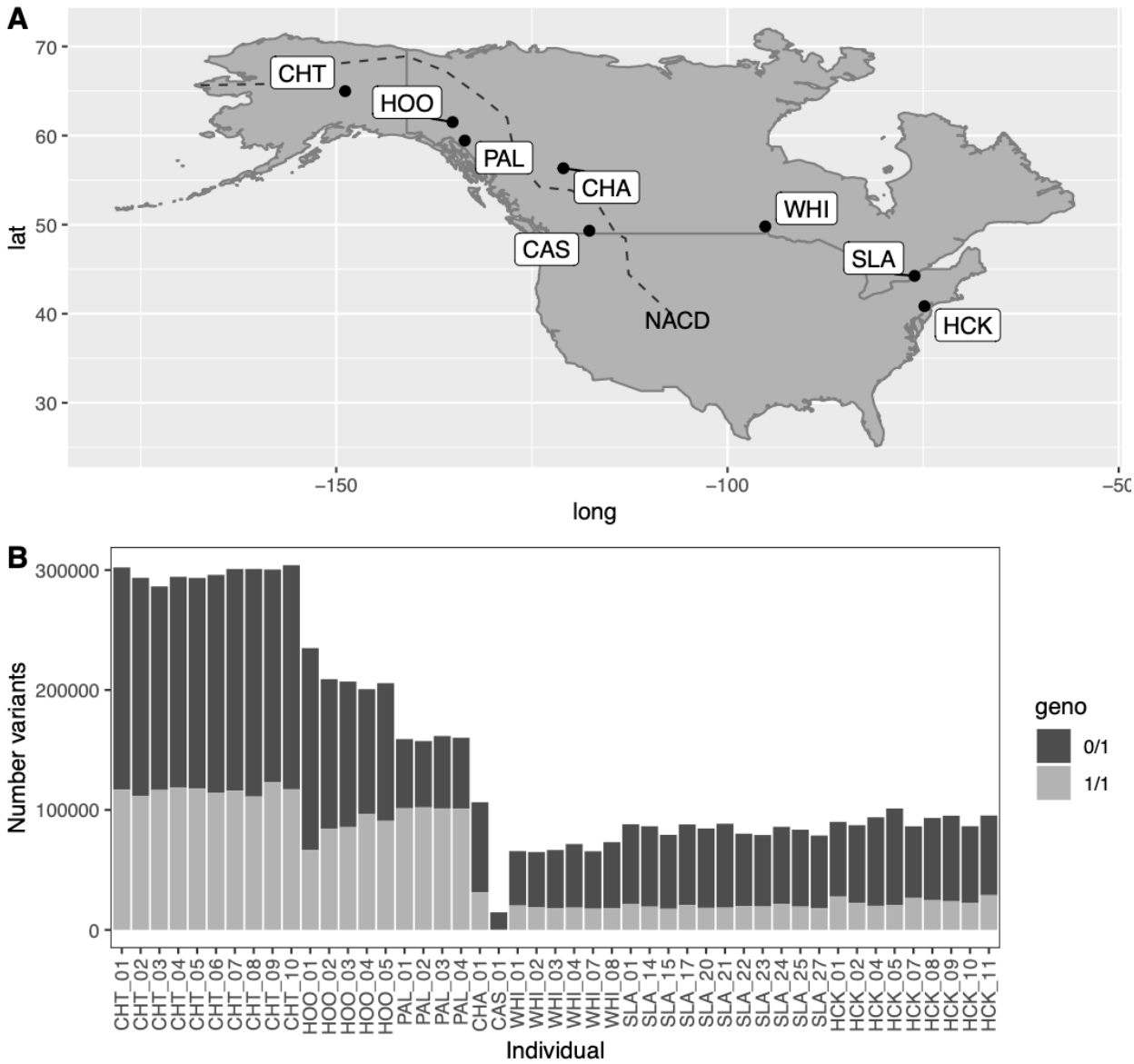


**Figure 1.** (A) Sampling locations for northern pike across North America, including Chatanika River, Alaska (CHT), Yukon River at Hootalinqua, Yukon Territory (HOO), Palmer Lake, British Columbia (B.C.; PAL), Charlie Lake, B.C. (CHA), Columbia River at Castlegar, B.C. (CAS), Whiteshell Hatchery, Manitoba (WHI), St. Lawrence River at New York (SLA), Hackettstown Hatchery, New Jersey (HCK). (B) Per individual total counts of homozygous alternate (light shading) or heterozygous (dark shading) variants.

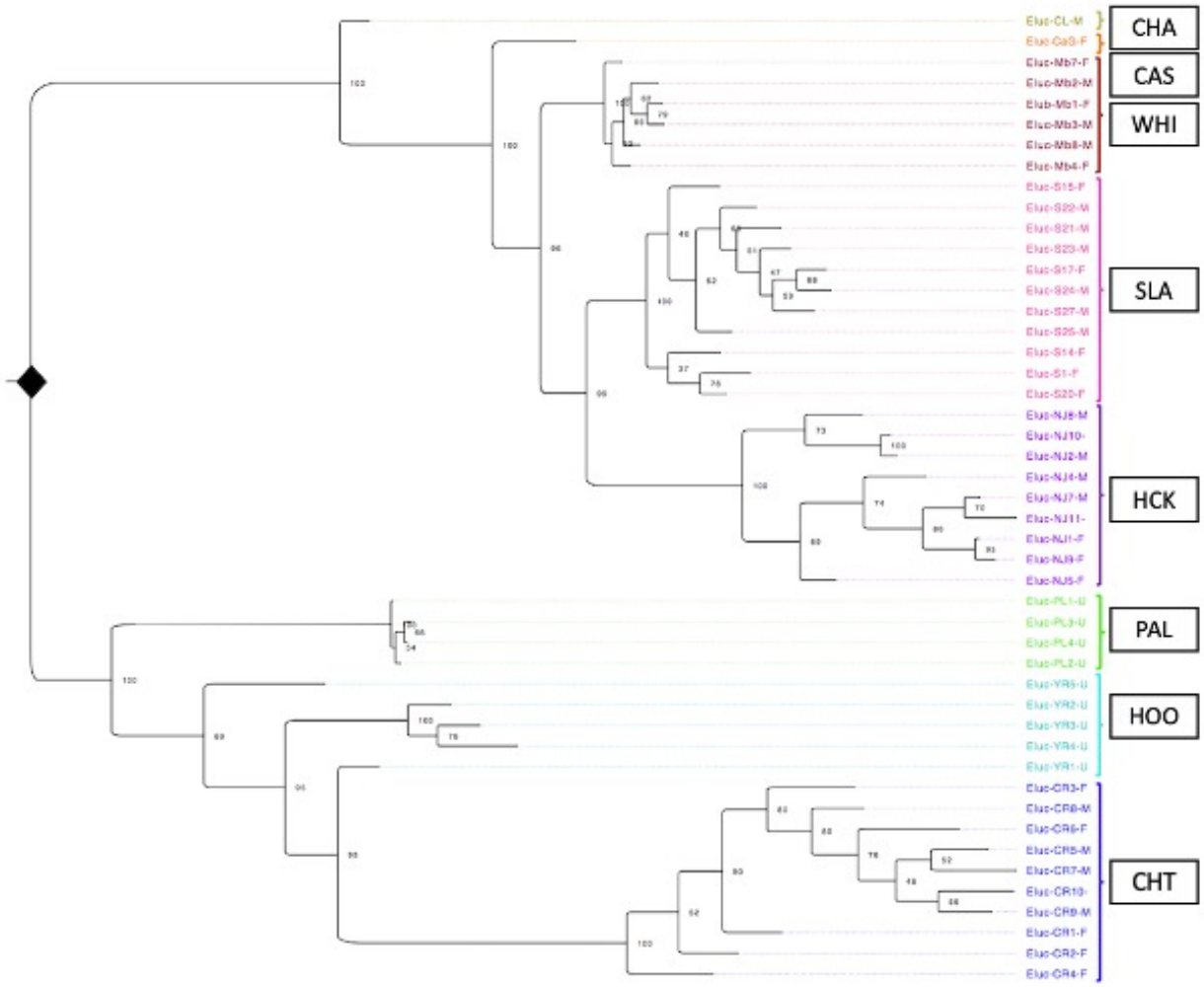


**Figure 2.** Genetic dendrogram clustering of individual northern pike using all filtered SNP variants. Bootstrap support is indicated at branch nodes as evaluated by 1,000 bootstraps. The largest separation in the data separates samples across the North American Continental Divide (NACD), with populations from Alaska (CHT), Yukon (HOO), and northwestern British Columbia (PAL) separating from all other populations. See Table 1 for all acronyms.

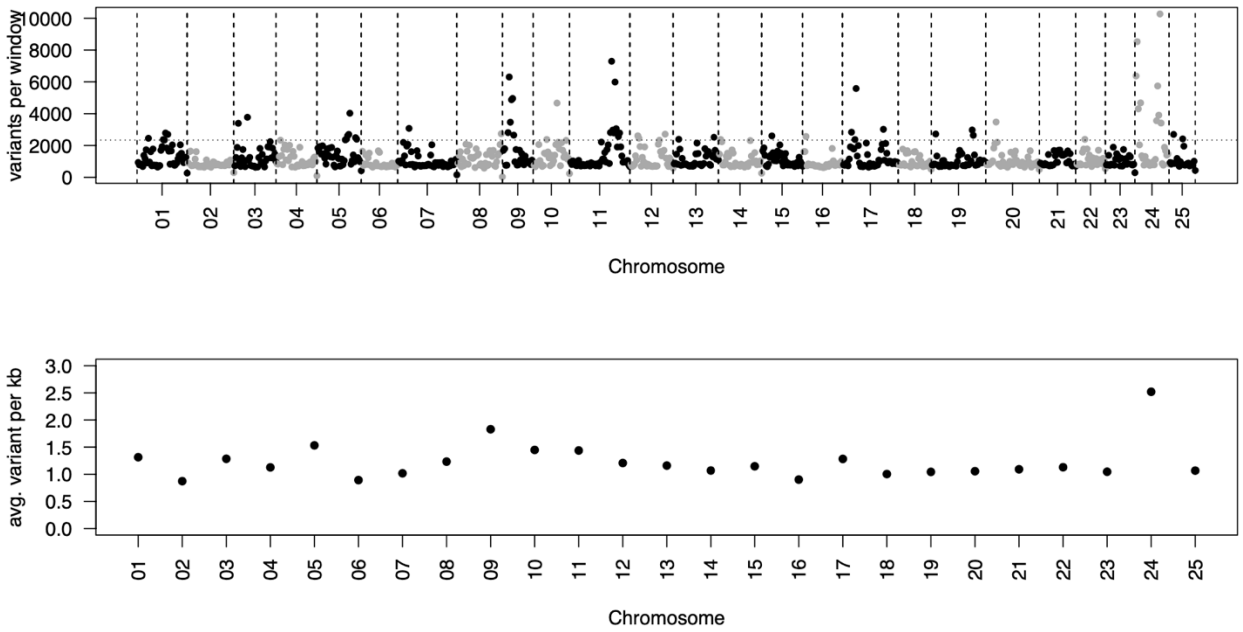


**Figure 3.** (A) Genome-wide SNP density by chromosome showing all SNP variants retained in the dataset, with variants summed across 1 Mbp windows (n = 930 windows total). Windows above the outlier threshold are indicated by the horizontal hatched line (*see Methods*). (B) The per chromosome average number of variants per kbp are shown, indicating chromosome 24 as a clear outlier overall.

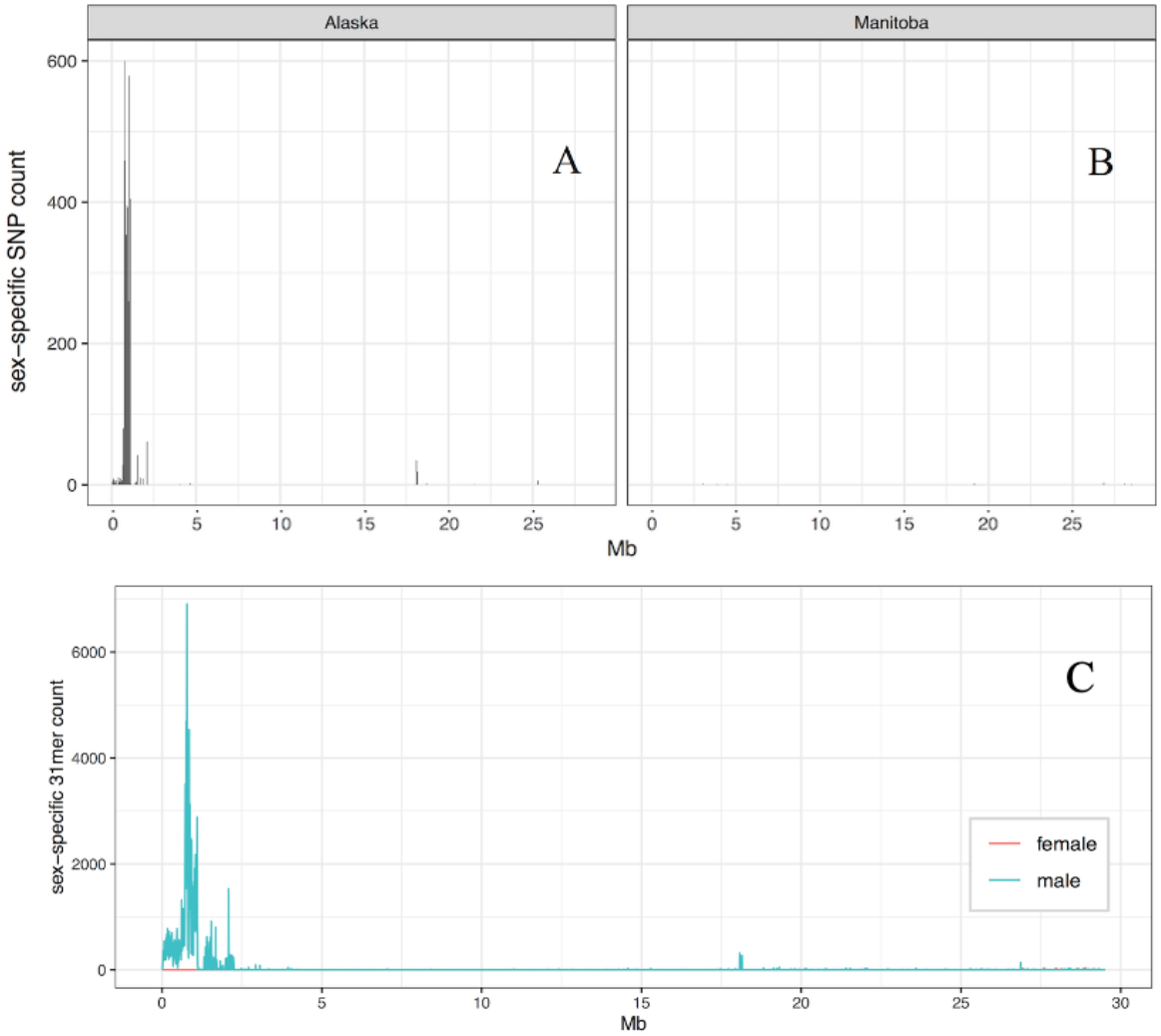


**Figure 4.** Sex-specific SNP counts for LG24 shows a strong signal in the Alaska population (panel A; CHT), and a lack of signal for populations east of the NACD (panel B; WHI used as example). The sex-specific signal is also observable in the Alaskan LG24 when viewing k-mer content rather than variant content. The k-mers were identified independent to the reference genome, but were aligned to the genome after identification.

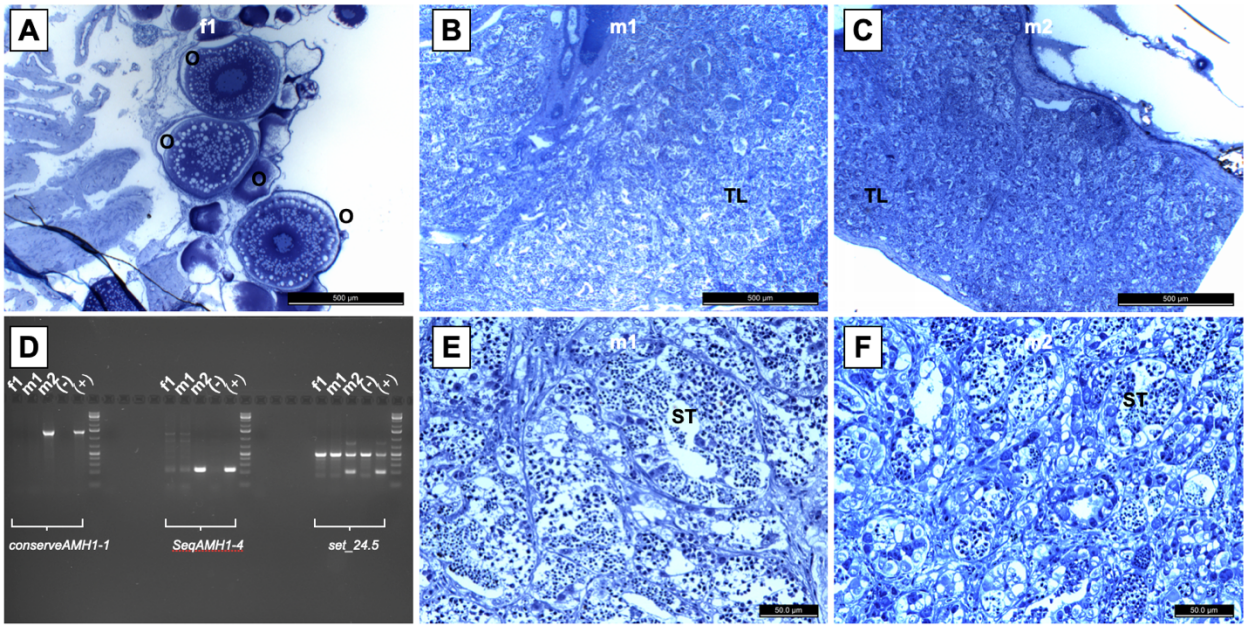


**Figure 5.** Gonadal histology and *amhby* PCR testing for sex genotypes and phenotypes in additional samples from Minto Flats, Alaska. The phenotypic female (f1) is shown in panel (A) with ooctyes (O) indicated at various stages of vitellogenesis. The two phenotypic males m1 and m2 are shown in panels (B) and (E), and panels (C) and (F), respectively, with testis lobules (TL) and seminiferous tubules (ST) indicates with gametes at various stages of spermatogenesis. In panel (D), PCR results of amplicons developed with the three primer pairs applied (see *Methods*) for the female (f1), an amhby(-) male (m1), and an amhby(+) male (m2), as well as a negative and positive control sample for each primer pair, selected from samples from Alaska (CHT) in the resequencing data. Primer sets *conserveAMH1-1* and *SeqAMH1-4* by Pan et al. (2021) indicate *amhby* by a single band, and *set_24.5* indicates presence of *amhby* by a second band.

# Tables

**Table 1.** Sampling location details and sample sizes classified using phenotypic sex information. Sample sites with unrecorded phenotypic sex are indicated as unknown (*unkn*.). The most recent genome assembly (v.4.0) was constructed using the Castlegar female. Samples were obtained throughout North America including from Alaska (AK), the Yukon Territories (YT), British Columbia (BC), Manitoba (MB), New York (NY), and New Jersey (NJ). Populations obtained from hatcheries are indicated with an asterisk. The short names for each sample site are provided.

| Country | Prov/ State | Population | Short name | Latitude | Longitude | Mal. (n) | Fem. (n) | Tot. (n) |
|---|---|---|---|---|---|---|---|---|
| USA | AK | Chatanika R. | CHT | 64.98396 | -148.86032 | 5 | 5 | 10 |
| Canada | YT | Hootalinqua | HOO | 61.51156 | -135.13209 | *unkn.* | *unkn.* | 5 |
| Canada | BC | Palmer Lk. | PAL | 59.43708 | -133.57592 | *unkn.* | *unkn.* | 4 |
| Canada | BC | Charlie Lk. | CHA | 56.32853 | -120.97835 | 1 | - | 1 |
| Canada | BC | Castlegar | CAS | 49.31538 | -117.65344 | - | 1 | 1 |
| Canada | MB | Whiteshell | WHI* | 49.80051 | -95.17243 | 3 | 3 | 6 |
| USA | NY | St. Lawrence | SLA | 44.24787 | -76.09785 | 6 | 5 | 11 |
| USA | NJ | Hackettstown | HCK* | 40.84155 | -74.83359 | 6 | 3 | 9 |
| | | **Total** | | | | 21 | 17 | 47 |

**Table 2.** Variant filtering steps and parameters applied after alignments of reads against the northern pike reference genome (v.4.0).

| Filter Step | Parameter value | Program | No. variants retained |
|---|---|---|---|
| Raw genotypes (includes indels) | - | - | 1,910,789 |
| SNP variants only | - | GATK | 1,363,731 |
| GATK hard filter | - | GATK | 1,189,068 |
| Min. quality | 20 | VCFtools | 1,186,793 |
| Min. mean depth | 10 | VCFtools | 1,153,668 |
| Max. mean depth | 60 | VCFtools | 1,152,122 |
| Max. missing count | 10 | VCFtools | 1,129,884 |
| Minor allele count | 1 | VCFtools | 1,129,701 |
| Per locus heterozygous genotypes | < 100% | R/ VCFtools | **1,127,943** |

**Table 3.** Per population summary statistics following genotyping, including the average number of heterozygous or alternate homozygous variants per individual, the per population observed heterozygosity, Tajima's D, and the number of private alleles. Sample size was rounded to nearest whole number. Acronyms: CHT = Chatanika R.; HOO = Hootalinqua; PAL = Palmer Lake; CHA = Charlie Lake; CAS = Castlegar; WHI = Whiteshell; SLA = St. Lawrence River; HCK = Hackettstown.

| Pop | (n) | Average no. variants per indiv. (ref/alt) | Average no. variants per indiv. (alt/alt) | Obs. heterozyg. ($H_{OBS}$) | Tajima's D | Private alleles |
|---|---|---|---|---|---|---|
| CHT | 10 | 180,943.2 | 116,275.7 | 0.1609 | -0.221 | 204,410 |
| HOO | 5 | 126,719.8 | 84,774.6 | 0.1127 | -0.331 | 67,803 |
| PAL | 4 | 58,128.5 | 101,464.5 | 0.0516 | -0.072 | 21,290 |
| CHA | 1 | 74,996.0 | 31,385.0 | 0.0668 | - | 14,736 |
| CAS | 1 | 14,552.0 | 96.0 | 0.0129 | - | 17,974 |
| WHI | 6 | 49,401.8 | 18,524.8 | 0.0438 | -1.068 | 101,149 |
| SLA | 11 | 64,393.5 | 19,442.5 | 0.0571 | -1.300 | 214,742 |
| HCK | 9 | 67,943.5 | 24,188.0 | 0.0604 | -0.328 | 107,085 |

## Additional Files

**Figure S1.** Hi-C contact map as visualized in Juicebox v 1.8.8 for assembly v4.0 GCF_004634155.1.

**Figure S2.** Distribution of repeats along the assembly v4.0 GCF_004634155.1, displayed in bins of 100,000 bp along the chromosome, with counts of masked bases per bin.

**Table S1.** Primer names, sequences, and annealing temperatures for sex markers. SeqAMH1 and ConserveAMH1 are reported in Pan et al. (2021).

**Table S2.** RepeatMasker summary table for assembly version 4.0 (GCF_004634155.1). Description of custom repeat library generation and masking methods described in Rondeau et al, (2014).

**Additional File S1.** Counts and locus identifiers of heterozygous or homozygous alternate variants per individual for all individuals in the study.

**Additional File S2.** Sum of variants per window across the genome in 1 Mbp windows.

**Additional File S3.** Sex-specific SNPs identified from the regional analysis of the whole-genome sequence data, shown per genomic coordinate as a 0 (ref/ref) or 1 (ref/alt) for each sexed northern pike from Chatanika River.

# REFERENCES


Arima Genomics. (2023). Arima Genomics. Retrieved from https://arimagenomics.com

Aury, J-M, Engelen, S, Istace, B, Monat, C, Lasserre-Zuber, P, Belser, C, et al. (2022). Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *GigaScience, 11*. doi:10.1093/gigascience/giac034

Auton, A, Abecasis, GR, Altshuler, DM, Durbin, RM, Abecasis, GR, Bentley, DR, et al. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Bachevskaja, LT, Pereverzeva, VV, Agapova, GA, & Grunin, SI. (2019). Genetic diversity of the population of northern pike *Esox lucius* L. from the rivers of the northeastern part of Russia. *Biology Bulletin, 46*(2), 154-160. doi:10.1134/S1062359019020031

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

Bernatchez, L, & Wilson, CC. (1998). Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular Ecology, 7*(4), 431-452. doi:https://doi.org/10.1046/j.1365-294x.1998.00319.x

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution, 10*(14), 7585-7601. doi:https://doi.org/10.1002/ece3.6483

Bongartz, P. (2019). Resolving repeat families with long reads. *BMC Bioinformatics, 20*(1), 232. doi:10.1186/s12859-019-2807-4

Bosworth, A, & Farrell, JM. (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. doi:10.1577/M05-060.1

Broad Institute. (2023). Picard toolkit. In. GitHub Respository.

Broman, KW, Murray, JC, Sheffield, VC, White, RL, & Weber, JL. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *The American Journal of Human Genetics, 63*(3), 861-869. doi:10.1086/302011

Carbine, WF. (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. doi:https://doi.org/10.1577/1548-8659(1941)71[149:OOTLHO]2.0.CO;2

Carim, KJ, Caleb Dysthe, J, McLellan, H, Young, MK, McKelvey, KS, & Schwartz, MK. (2019). Using environmental DNA sampling to monitor the invasion of nonnative *Esox lucius* (northern pike) in the Columbia River basin, USA. *Environmental DNA, 1*(3), 215-226. doi:https://doi.org/10.1002/edn3.22

Carim, KJ, Eby, LA, Miller, LM, McLellan, H, Dupuis, V, & Schwartz, MK. (2022). Mechanism of northern pike invasion in the Columbia River Basin. *Management of Biological Invasions, 13*(1).

Casselman, JM. (1975). Sex ratios of northern pike, *Esox lucius* Linnaeus. *Transactions of the American Fisheries Society, 104*(1), 60-63. doi:10.1577/1548-8659(1975)104<60:SRONPE>2.0.CO;2

CCAC. (2005). Canadian Council on Animal Care guidelines on: the care and use of fish in research, teaching and testing. In (pp. 94). Ottawa, Ontario: Canadian Council on Animal Care.

Chaisson, MJP, Wilson, RK, & Eichler, EE. (2015). Genetic variation and the *de novo* assembly of human genomes. *Nature Reviews Genetics, 16*(11), 627-640. doi:10.1038/nrg3933

Charlesworth, B. (1991). The evolution of sex chromosomes. *Science, 251*(4997), 1030-1033.

Charlesworth, D, & Charlesworth, B. (2005). Sex chromosomes: evolution of the weird and wonderful. *Current Biology, 15*(4), R129-R131. doi:https://doi.org/10.1016/j.cub.2005.02.011

Chin, C-S, Alexander, DH, Marks, P, Klammer, AA, Drake, J, Heiner, C, et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods, 10*(6), 563-569. doi:10.1038/nmeth.2474

Clark, CF. (1950). Observations on the spawning habits of the northern pike, *Esox lucius*, in Northwestern Ohio. *Copeia, 1950*(4), 285-288. doi:10.2307/1437909

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

Crossman, EJ, & Harington, CR. (1970). Pleistocene pike, *Esox lucius*, and *Esox* sp., from the Yukon Territory and Ontario. *Canadian Journal of Earth Sciences, 7*(4), 1130-1138. doi:10.1139/e70-107

Danecek, P, Auton, A, Abecasis, G, Albers, CA, Banks, E, DePristo, MA, et al. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Danecek, P, Bonfield, JK, Liddle, J, Marshall, J, Ohan, V, Pollard, MO, et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). doi:10.1093/gigascience/giab008

Dekker, J, Rippe, K, Dekker, M, & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295*(5558), 1306-1311. doi:doi:10.1126/science.1067799

DePristo, MA, Banks, E, Poplin, R, Garimella, KV, Maguire, JR, Hartl, C, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-498. doi:10.1038/ng.806

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

DFO. (2012). *2010 Survey of recreational fishing in Canada*. Economic Analysis and Statistics Strategic Policy Retrieved from http://www.dfo-mpo.gc.ca/stats/rec/can/2010/index-eng.htm.

Dudchenko, O, Batra, SS, Omer, AD, Nyquist, SK, Hoeger, M, Durand, NC, et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science, 356*(6333), 92-95. doi:doi:10.1126/science.aal3327

Durand, NC, Robinson, JT, Shamim, MS, Machol, I, Mesirov, JP, Lander, ES, et al. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems, 3*(1), 99-101. doi:https://doi.org/10.1016/j.cels.2015.07.012

Durand, NC, Shamim, MS, Machol, I, Rao, SSP, Huntley, MH, Lander, ES, et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems, 3*(1), 95-98. doi:10.1016/j.cels.2016.07.002

Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323*(5910), 133-138. doi:10.1126/science.1162986

Eschbach, E, Nolte, AW, Kohlmann, K, Alós, J, Schöning, S, & Arlinghaus, R. (2021). Genetic population structure of a top predatory fish (northern pike, Esox lucius) covaries with anthropogenic alteration of freshwater ecosystems. *Freshwater Biology, 66*(5), 884-901. doi:https://doi.org/10.1111/fwb.13684

Farrell, JM, & Barry, CC. (2012). *Northern pike monitoring and management in the upper St. Lawrence River.* . Retrieved from

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. doi:https://doi.org/10.1111/jfb.12712

Gao, G, Nome, T, Pearse, DE, Moen, T, Naish, KA, Thorgaard, GH, et al. (2018). A new single nucleotide polymorphism database for rainbow trout generated through whole genome resequencing. *Frontiers in Genetics, 9*. doi:10.3389/fgene.2018.00147

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. doi:10.1186/s12864-017-3879-z

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

Goodwin, S, Gurtowski, J, Ethe-Sayers, S, Deshpande, P, Schatz, MC, & McCombie, WR. (2015). Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res, 25*(11), 1750-1756. doi: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. doi:https://doi.org/10.1016/j.aquaculture.2005.10.009

Grande, T, Laten, H, & López, JA. (2004). Phylogenetic relationships of extant Esocid species (Teleostei: Salmoniformes) based on morphological and molecular characters. *Copeia, 2004*(4), 743-757, 715.

Gruber, B, Unmack, PJ, Berry, OF, & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources, 18*(3), 691-699. doi:https://doi.org/10.1111/1755-0998.12745

Harris, RS. (2007). *Improved pairwise alignment of genomic DNA.* (Ph.D.), The Pennsylvania State University, Pennsylvania. Retrieved from https://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf

Harvey, B. (2009). *A biological synopsis of northern pike (Esox lucius)*. Canadian Manuscript Report of Fisheries and Aquatic Sciences.

Hedrick, PW. (2005). Tests of the neutral theory and evidence of selection: Tajima test. In *Genetics of populations* (3 ed., pp. 433-436). Sudbury, Massachusetts: Jones and Bartlett Publishers.

Höglund, J. (2009). *Evolutionary Conservation Genetics*. Oxford, United Kingdom: Oxford University Press.

Huffman, K, Farrell, JM, & Whipps, CM. (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*. Paper presented at the American Fisheries Society (AFS), Quebec City, QC. https://afs.confex.com/afs/2014/webprogram/Paper16111.html

Jackman, SD, Coombe, L, Chu, J, Warren, RL, Vandervalk, BP, Yeo, S, et al. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics, 19*(1), 393. doi:10.1186/s12859-018-2425-6

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405. doi: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. doi:10.1093/bioinformatics/btr521

Kamvar, ZN, Tabima, JF, & Grünwald, NJ. (2014). poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ, 2*, e281. doi:10.7717/peerj.281

Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649. doi:10.1093/bioinformatics/bts199

Knaus, BJ, & Grünwald, NJ. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources, 17*(1), 44-53. doi:https://doi.org/10.1111/1755-0998.12549

Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, JT, Ganapathy, G, et al. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology, 30*(7), 693-700. doi:10.1038/nbt.2280

Koren, S, Walenz, BP, Berlin, K, Miller, JR, Bergman, NH, & Phillippy, AM. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res, 27*(5), 722-736. doi:10.1101/gr.215087.116

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

Leong, JS, Jantzen, SG, von Schalburg, KR, Cooper, GA, Messmer, AM, Liao, NY, et al. (2010). *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *BMC Genomics, 11*(1), 279. doi:10.1186/1471-2164-11-279

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv:1303.3997. Retrieved from https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L

Li, H. (2017). Bioawk. In: GitHub.

Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Luan, P, Huo, T, Ma, B, Song, D, Zhang, X, & Hu, G. (2021). Genomic inbreeding and population structure of northern pike (*Esox lucius*) in Xinjiang, China. *Ecology and Evolution, 11*(10), 5657-5668. doi:https://doi.org/10.1002/ece3.7469

Mackay, TFC. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics, 35*(1), 303-339. doi:10.1146/annurev.genet.35.102401.090633

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

Martinez Barrio, A, Lamichhaney, S, Fan, G, Rafati, N, Pettersson, M, Zhang, H, et al. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife, 5*, e12081. doi:10.7554/eLife.12081

McKenna, A, Hanna, M, Banks, E, Sivachenko, A, Cibulskis, K, Kernytsky, A, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Miller, LM, & Kapuscinski, AR. (1996). Notes: Microsatellite DNA markers reveal new levels of genetic variation in northern pike. *Transactions of the American Fisheries Society, 125*(6), 971-977. doi:10.1577/1548-8659(1996)125<0971:NMDMRN>2.3.CO;2

Miller, LM, & Senanan, W. (2003). A review of northern pike population genetics research and its implications for management. *North American Journal of Fisheries Management, 23*(1), 297-306. doi:https://doi.org/10.1577/1548-8675(2003)023<0297:ARONPP>2.0.CO;2

Moen, T, Hoyheim, B, Munck, H, & Gomez-Raya, L. (2004). A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. *Animal Genetics, 35*(2), 81-92. doi:https://doi.org/10.1111/j.1365-2052.2004.01097.x

Moran, BM, & Anderson, EC. (2019). Bayesian inference from the conditional genetic stock identification model. *Canadian Journal of Fisheries and Aquatic Sciences, 76*(4), 551-560. doi:10.1139/cjfas-2018-0016

Mostovoy, Y, Levy-Sakin, M, Lam, J, Lam, ET, Hastie, AR, Marks, P, et al. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature Methods, 13*(7), 587-590. doi:10.1038/nmeth.3865

Nelson, TC, Crandall, JG, Ituarte, CM, Catchen, JM, & Cresko, WA. (2019). Selection, Linkage, and Population Structure Interact To Shape Genetic Variation Among Threespine Stickleback Genomes. *Genetics, 212*(4), 1367-1382. doi:10.1534/genetics.119.302261

Nicod, J-C, Wang, YZ, Excoffier, L, & Largiadèr, CR. (2004). Low levels of mitochondrial DNA variation among central and southern European *Esox lucius* populations. *Journal of Fish Biology, 64*(5), 1442-1449. doi:https://doi.org/10.1111/j.0022-1112.2004.00397.x

Ouellet-Cauchon, G, Mingelbier, M, Lecomte, F, & Bernatchez, L. (2014). Landscape variability explains spatial pattern of population structure of northern pike (*Esox lucius*) in a large fluvial system. *Ecology and Evolution, 4*(19), 3723-3735. doi:https://doi.org/10.1002/ece3.1121

Pan, Q, Feron, R, Jouanno, E, Darras, H, Herpin, A, Koop, B, et al. (2021). The rise and fall of the ancient northern pike master sex-determining gene. *eLife, 10*, e62858. doi:10.7554/eLife.62858

Pan, Q, Feron, R, Yano, A, Guyomard, R, Jouanno, E, Vigouroux, E, et al. (2019). Identification of the master sex determining gene in northern pike (*Esox lucius*) reveals restricted sex chromosome differentiation. *PLOS Genetics, 15*(8), e1008013. doi:10.1371/journal.pgen.1008013

Pan, Q, Herpin, A, & Guiguen, Y. (2023). Inactivation of the anti-Müllerian hormone receptor type 2 (*amhrII*) gene in northern pike (*Esox lucius*) results in male-to-female sex reversal. *Sexual Development, 16*(4), 289-294. doi:10.1159/000521003

Piertney, SB, & Oliver, MK. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity, 96*(1), 7-21. doi:10.1038/sj.hdy.6800724

Poplin, R, Ruano-Rubio, V, DePristo, MA, Fennell, TJ, Carneiro, MO, Auwera, GAVd, et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178. doi:10.1101/201178

Priegel, GR, & Krohn, DC. (1975). *Characteristics of a northern pike spawning population*. Madison, Wisconsin: Madison, Wisconsin: Wisconsin Department of Natural Resources Retrieved from https://search.library.wisc.edu/digital/AMHTIK6R4JXFCA8F.

Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, MA, Bender, D, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Putnam, NH, O'Connell, BL, Stites, JC, Rice, BJ, Blanchette, M, Calef, R, et al. (2016). Chromosome-scale shotgun assembly using an in *vitro* method for long-range linkage. *Genome Research, 26*(3), 342-350. doi:10.1101/gr.193474.115

Quinlan, AR, & Hall, IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

R Core Team. (2023). R: A Language and Environment for Statistical Computing. In. Vienna, Austria.

Rajendiran, P, Jaafar, F, Kar, S, Sudhakumari, C, Senthilkumaran, B, & Parhar, IS. (2021). Sex determination and differentiation in Teleost: roles of genetics, environment, and brain. *Biology (Basel), 10*(10). doi:10.3390/biology10100973

Rambaut, A. (2016). Figtree (v.1.4.3). Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Redon, R, Ishikawa, S, Fitch, KR, Feuk, L, Perry, GH, Andrews, TD, et al. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454. doi:10.1038/nature05329

Robinson, JT, Turner, D, Durand, NC, Thorvaldsdóttir, H, Mesirov, JP, & Aiden, EL. (2018). Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst, 6*(2), 256-258.e251. doi:10.1016/j.cels.2018.01.001

Rondeau, EB, Minkley, DR, Leong, JS, Messmer, AM, Jantzen, JR, von Schalburg, KR, et al. (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. doi:10.1371/journal.pone.0102089

Rougemont, Q, Carrier, A, Le Luyer, J, Ferchaud, A-L, Farrell, JM, Hatin, D, et al. (2019). Combining population genomics and forward simulations to investigate stocking impacts: A case study of Muskellunge (*Esox masquinongy*) from the St. Lawrence River basin. *Evolutionary Applications, 12*(5), 902-922. doi:https://doi.org/10.1111/eva.12765

Sakamoto, T, Danzmann, RG, Gharbi, K, Howard, P, Ozaki, A, Khoo, SK, et al. (2000). A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. *Genetics, 155*(3), 1331-1345. doi:10.1093/genetics/155.3.1331

Seeb, JE, Seeb, LW, Oates, DW, & Utter, FM. (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. doi:10.1139/f87-068

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

Seppey, M, Manni, M, & Zdobnov, EM. (2019). BUSCO: assessing genome assembly and annotation completeness. In M. Kollman (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227-245): Springer.

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

Skov, C, & Nilsson, PA. (2018). *Biology and Ecology of Pike* (1 ed.). Boca Raton: CRC Press.

Star, B, Nederbragt, AJ, Jentoft, S, Grimholt, U, Malmstrøm, M, Gregers, TF, et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature, 477*(7363), 207-210. doi:10.1038/nature10342

Stoddart, D, Heron, AJ, 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, 106*(19), 7702-7707. doi:doi:10.1073/pnas.0901054106

Sun, H, Jiao, W-B, Krause, K, Campoy, JA, Goel, M, Folz-Donahue, K, et al. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics, 54*(3), 342-348. doi:10.1038/s41588-022-01015-0

Sunde, J, Yıldırım, Y, Tibblin, P, Bekkevold, D, Skov, C, Nordahl, O, et al. (2022). Drivers of neutral and adaptive differentiation in pike (*Esox lucius*) populations from contrasting environments. *Molecular Ecology, 31*(4), 1093-1110. doi:https://doi.org/10.1111/mec.16315

Sutherland, BJG, Rico, C, Audet, C, & Bernatchez, L. (2017). Sex chromosome evolution, heterochiasmy, and physiological QTL in the salmonid brook charr *Salvelinus fontinalis*. *G3 Genes|Genomes|Genetics, 7*(8), 2749-2762. doi:10.1534/g3.117.040915

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3), 585-595. doi:10.1093/genetics/123.3.585

Thorburn, D-MJ, Sagonas, K, Binzer-Panchal, M, Chain, FJJ, Feulner, PGD, Bornberg-Bauer, E, et al. (2023). Origin matters: Using a local reference genome improves measures in population genomics. *Molecular Ecology Resources, n/a*(n/a). doi:https://doi.org/10.1111/1755-0998.13838

Turner, SD. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*, 005165. doi:10.1101/005165

Unanue, ER, Turk, V, & Neefjes, J. (2016). Variations in MHC class II antigen processing and presentation in health and disease. *Annu Rev Immunol, 34*, 265-297. doi:10.1146/annurev-immunol-041015-055420

Van der Auwera, GA, Carneiro, MO, Hartl, C, Poplin, R, Del Angel, G, Levy-Moonshine, A, et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 43*(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Walker, BJ, Abeel, T, Shea, T, Priest, M, Abouelliel, A, Sakthikumar, S, et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE, 9*(11), e112963. doi:10.1371/journal.pone.0112963

Warren, RL, Yang, C, Vandervalk, BP, Behsaz, B, Lagman, A, Jones, SJM, et al. (2015). LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience, 4*(1), 35. doi:10.1186/s13742-015-0076-3

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*: Springer-Verlag New York.

Wickham, H, François, R, Henry, L, Müller, K, & Vaughan, D. (2023). dplyr: a grammar of data manipulation. Retrieved from https://github.com/tidyverse/dplyr

Wilson, MVH. (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. doi:10.1139/e80-030

Wilson, MVH, Brinkman, DB, & Neuman, AG. (1992). Cretaceous Esocoidei (Teleostei): early radiation of the pikes in North American fresh waters. *Journal of Paleontology, 66*(5), 839-846. doi:10.1017/S0022336000020849

Wooller, MJ, Gaglioti, B, Fulton, TL, Lopez, A, & Shapiro, B. (2015). Post-glacial dispersal patterns of Northern pike inferred from an 8800 year old pike (*Esox* cf. *lucius*) skull from interior Alaska. *Quaternary Science Reviews, 120*, 118-125. doi:https://doi.org/10.1016/j.quascirev.2015.04.027

Yeo, S, Coombe, L, Warren, RL, Chu, J, & Birol, I. (2017). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics, 34*(5), 725-731. doi:10.1093/bioinformatics/btx675

Yuan, Y, Scheben, A, Edwards, D, & Chan, TF. (2021). Toward haplotype studies in polyploid plants to assist breeding. *Mol Plant, 14*(12), 1969-1972. doi:10.1016/j.molp.2021.11.004