# Northern Pike of North America: Population Genomics and Sex Determination

by

Hollie Johnson
B.Sc., University of Victoria, 2010

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Biology

Northern Pike of North America:
Population Genomics and Sex Determination

by

Hollie Johnson
B.Sc., University of Victoria, 2010

Supervisory Committee

Dr. Ben Koop, Supervisor
Department of Biology

Dr. John Taylor, Co-Supervisor
Department of Biology

Dr. John Nelson, Departmental Member
Department of Biology

# Abstract

Northern Pike (*Esox lucius*) is an economically and ecologically valuable species with a circumpolar distribution across the Northern Hemisphere. Northern Pike have been shown to have low levels of genetic variation despite their great capacity to colonize new environments. Here, high-resolution resequencing data from 47 Northern Pike from across North America was used for SNP discovery and population analysis. Our analysis reveals an extraordinary lack of genetic variation among Northern Pike with observed heterozygosity ($H_O$) of just 0.0835. Our analyses suggest that two major groups of Northern Pike exist in North America that are separated by the North American Continental Divide. Genetic variation associated with the stratification of these two groups resides across the genome particularly in gene regions with multiple copy number variants and functions related to immunity, tissue permeability, and development. Northern Pike from Alaska and the Yukon River harbour about two times more heterozygosity than Northern Pike east of the Continental Divide with an average of one heterozygous SNP every 6,250 bases. Populations east of the Continental Divide possess a remarkable level of genetic homogenization with an average of just one heterozygous SNP every 16,500 bases. For comparison, an average of one heterozygous SNP per 309 bases was reported in herring (Martinez Barrio et al., 2016), one per 500 in Atlantic cod (Star et al., 2011), and one per 750 bases in Coho and chinook salmon (Koop, 2018). This is at least 5 – 10 fold less variation than is seen in humans (the 1000 Genomes Project Consortium, 2015).

We observed a recently described master sex-determining gene, *amhby*, in three western North American populations but not in populations east of the Continental Divide. We could not resolve any signals indicating a genetic sex determination system was present in populations

from southern Manitoba or the St. Lawrence River. This may indicate that environmental sex determination is at play in these populations. We found evidence of a possible female-heterozygous, male homozygous ZW-ZZ genetic sex-determination system in New Jersey Northern Pike.

With the highest average of 181,268 heterozygous SNPs genome wide and the greatest $H_o$ (0.3228) of all populations, as well as the presence of the sex-determining gene *amhby* indicate that Northern Pike from our Alaskan population are the oldest in North America. Fewer numbers of heterozygous SNPs (61,073), low $H_o$ (0.0922), and the absence of *amhby* in Northern Pike east of the Continental Divide suggests that these are relatively young populations and are descended from a small founding population. These results imply that Northern Pike first came to North America through Beringia and colonized its North American range from there, possibly via pro-glacial lake formation and drainage. However, from the data herein it was not possible to trace how re-colonization occurred after the final retreat of glaciers at the end of the last ice age.

This thesis provides a genetically high-resolution snapshot of Northern Pike population structure in North America. It demonstrates that organisms with largely homogenous genomes can be incredibly successful and resilient. Finally, it adds to the complex subject of sex determination in fish and provides insight into a sex determination system in transition.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms

*amh* – Anti-Müllerian Hormone
*amhby* – Anti-Müllerian Hormone B on the Y-chromosome
BAM – Binary Alignment/Map File
DAPC – Discriminant Analysis of Principle Components
GATK – Genome Analysis Toolkit
GWAS – Genome Wide Association Study
HWE – Hardy-Weinberg Equilibrium
IGV – Integrated Genomics Viewer
InDels – Insertions and Deletions
KYA – Thousand Years Ago
MHC – Major Histocompatibility Complex
MSD – Master Sex Determination
MYA – Million Years Ago
PCA – Principle Component Analysis
PCR – Polymerase Chain Reaction
SNP – Single Nucleotide Polymorphism
VCF – Variant Call Format
WGD – Whole Genome Duplication

# Acknowledgements

I would like to send my deepest and most sincere thanks to Dr. Ben Koop for sparking my interest in molecular evolution a decade ago, for being inspiring and fun to work with, and for broadening my focus from details to big picture implications. Thank you, Ben.

A big thank you to Dr. John Taylor for careful and thorough consideration of my writing and for asking me questions that shed insight onto my own hidden assumptions. I love the opportunity to contemplate my own beliefs.

Thank you, Dr. John Nelson and Dr. Chris Darimont, for participating in this process and for your support.

Last but not least, thanks to my children, Julian and Isabel Prieto, and my partner Geoff de Ruiter for support, patience, inspiration, comical relief, love and acceptance. Thank you for helping me realize the importance of balance.

# Dedication

Dedicated to my Mum, Dad, and brother, and everyone else who has ever been bitten by a pike.

# Chapter 1

# Introduction

## Thesis Overview

Whole genome data from forty-seven Northern Pike (*Esox lucius*) from North America were aligned to the Northern Pike reference genome. Single nucleotide polymorphisms (SNPs) were identified and analyzed for patterns of genetic variation, population structure, and sex-specific variation. Numerous reduced representation sequencing studies have reported very low levels of polymorphism in Northern Pike. The paradox of how a species with such low amounts of genetic variation can be so widespread and prosperous was an inspiration for this thesis. Another inspiration was observations of skewed sex ratios and a lack of sex-specific genetic variation in North American Northern Pike. Herein, I quantified and analyzed the distribution of variation across genomes of North American Northern Pike.

In this first chapter of the thesis, relevant background information was reviewed and pertinent concepts introduced. The ecology and phylogenetic position of Northern Pike was discussed, highlighting their widespread distribution as an apex predator and their importance as an outgroup to the heavily studied salmonids. The role of genetic variation in evolution was summarized, and studies that have reported on genetic variation in Northern Pike and other teleosts were discussed. The fossil history of Esociformes and Northern Pike was discussed, as well as hypotheses surrounding the colonization and re-colonization of North America. Mechanisms of sex determination in teleosts were reviewed, as well as what is known about sex determination in Northern Pike.

In the second chapter of this thesis population genomics was investigated. Sample origin and composition was discussed, as well as sequencing technology and variant discovery techniques. Patterns of genetic variation were then analyzed. First, the number of variants in each individual was quantified and compared between populations. A phylogeny was constructed, and SNPs were used to identify clustering patterns through principle component analysis (PCA). SNPs that stratified clusters were identified with a discriminant analysis of principle component analysis (DAPC). The discriminant axes of the DAPC were examined to pinpoint genomic locations and clustering patterns that delineated the strongest trends in the data. Genes present at these genomic locations were identified. Tests for Hardy Weinberg Equilibrium (HWE) were performed, as well as Tajima's D neutrality test. Wright's fixation index ($F_{ST}$) was used to compare the two most differentiated groups of Northern Pike and used to visualize regions of peak differentiation throughout the genome. Genes located in highly differentiated regions were identified and reviewed.

The third chapter of my thesis investigated genetic variation related to sex. This work used the same set of SNPs discussed in chapter two, but associated genetic variation with sexual phenotype. DAPC was performed with the constraint of differentiating males and females, among and within populations. The distribution sex-specific SNPs was visualized across the genome to identify potential sex determining regions. Genes located in region of dense sex-specific variation were identified and discussed. A genome wide association study was performed, as well as comparisons of $F_{ST}$ between males and females.

In the fourth and final chapter of my thesis summarized conclusions from chapters one and two. The implications of my results on the role of genetic variation and biogeographical history of Northern Pike were discussed. The loss of the sex determination system was

emphasized and the proposed alternate region of sex determination in New Jersey pike was highlighted.

This thesis illustrates that the Northern Pike genome is remarkably homogenous, yet still allows for immense success and ability to invade and recolonize vast geographical areas. It suggests that maintenance of genetic variation in immune-related genes is pivotal for Northern Pike. This work also helps to clarify the colonization of North America by Northern Pike, and exemplifies the plasticity of sex determination mechanisms in teleost fish

## Northern Pike ecology

Northern Pike are an important species economically and ecologically. Economically, they are valued in sport fishing and recreation. Anglers contribute approximately 8 million dollars annually to local economies in Canada and Northern Pike comprise 10% of sport fishing catch (Government of Canada, 2016). Ecologically, Northern Pike are a top tier ambush predator found in lakes, rivers, and mildly saline habitats in the circumpolar region of the northern hemisphere. They exert a significant role in freshwater ecosystems in the northern hemisphere (Craig, 2008).

In early spring, when water temperatures are between 8-12°C, Northern Pike spawn in shallow waters over vegetation (Casselman and Lewis, 1996). Approximately 6-8 weeks after spawning, eggs hatch and fry disperse into deeper water with macrophyte cover of 30 – 70% (Casselman and Lewis, 1996; Craig, 2008). Macrophytes serve as nurseries for juveniles and also provide coverage for adults (Craig, 2008). Macrophyte cover is especially important for smaller individuals to protect against predation. Northern Pike are cannibalistic, and because of this, exert a direct effect on their own population structure (Craig, 2008). Natal-site and spawning-site fidelity has been observed in Pike, and tagging experiments have shown that females move

greater distances than males during the spawning season (Craig, 2008; Koed et al., 2006). Pike reach sexual maturation by age 3 – 4 (Casselman, 1974), reaching a length of 50 – 130 cm, and can live for longer than 10 years (Forsman et al., 2015; Senay et al., 2017).

## Genetic variation

Genetic variation can be described as the differences in nucleotide sequence within and among individuals, populations, and species. Within an individual, variation appears as heterozygosity. In diploid and polyploid systems, a locus is said to be heterozygous when the nucleotide at that location is different between chromosomes. Homozygosity is the term used to describe the situation where nucleotides are the same at the loci in question. When heterozygosity or differential homozygosity at the same locus across individuals is observed, the locus is classified as a single nucleotide polymorphism (SNP). Collections of SNPs can be analyzed to inform genotype-phenotype relationships, population structure, divergence times, and give clues about stochastic events that could affect the gene pool (Graur and Li, 2000).

Factors influencing the amount of variation include mutation, random genetic drift, population size or age, bottleneck events, selection pressure, recombination, gene flow, as well as genomic processes such as movement of transposable elements, gene duplication or whole genome duplication, and gene conversion (Wayne and Miyamota, 2006). Mutation, recombination, gene flow, and genomic processes act to increase the amount of variation through the incorporation of new nucleotide sequences. Selection pressure generally reduces diversity in the genome through the conservation of nucleotide sequences that produce individuals with the greatest fitness. An exception to this is overdominance, or heterozygote advantage, where the phenotype of the heterozygote is more beneficial than either of the two homozygous states (Sellis et al., 2011). Random genetic drift and bottleneck events drive populations toward

homozygosity, and are largely influenced by effective population size. A smaller effective population size essentially limits the gene pool in a population, and alleles tend to be driven toward fixation in fewer generations when the effective population size is small (Gillespie, 2004).

In the face of climate change and the onset of the Anthropocene (Davies, 2016), species are challenged with rapidly changing environments. Genetic variation is considered pivotal for the success of populations (Lande, 1988; Reed and Frankham, 2003). Nucleotide sequence within gene coding regions determines the amino acid sequence of proteins and can produce a spectrum of gene variants through post-transcriptional modification. Outside of gene coding regions, nucleotide sequence affects binding affinity for transcription factors, can cue epigenetic effects such as methylation, or affect the shape and availability of DNA for transcription. Thus, when one organism holds two chromosomes with different sequences, it has the capability of producing gene products with different characteristics, and reacting in more than one way to the same external signals. A greater number of alleles in both individuals and populations provides resistance to selective pressures by increasing the potential for adaptation (Barrett and Schluter, 2008; Höglund, 2009). It follows, then, that knowledge of genetic variation can help us understand how species may respond to environmental change and disturbance.

## Variation in Northern Pike

Low levels of polymorphism have been observed in many Northern Pike populations (Hansen et al., 1999; Healy and Mulcahy, 1980; Miller and Kapuscinski, 1996, 1997; Miller and Senanan, 2003; Rondeau et al., 2014; Seeb et al., 1987; Skog et al., 2014). Observed heterozygosity ($H_o$) is a parameter used to describe variability and is defined as the observed proportion of individuals that are heterozygous at a specific locus (Gillespie, 2004). Early

investigations into polymorphism assessed allozyme heterozygosity (Healy and Mulcahy, 1980; Seeb et al., 1987) (Table 2) Allozymes are enzymes that share function but differ in amino acid sequence (and therefore, structure), and can be resolved electrophoretically. Because of the degenerate nature of the genetic code (i.e. the ability of multiple codons to specify the same amino acid), allozymes represent a subset of the actual genetic variation at a given locus. Healy and Mulcahy (1980) examined 26 loci in 15 allozymes from seven Northern Pike populations across Europe and North America. Of the 26 loci analyzed, only three were found to be polymorphic, and $H_o$ was low (0.019 among all populations; 0 – 0.075 within populations). Seeb et al. (1987) found only 2 of 65 loci to be polymorphic among eight North American populations also with low $H_o$ (0 – 0.002). Seeb et al. (1987) also saw a fixation of a different allozyme at one North American location that was absent in all the others. An allozyme analysis of yellow perch, a species with a similar habitat and distribution to Northern Pike, detected 29 polymorphic loci of 54, and reported $H_o$ between 0.017 and 0.054 (Marsden et al., 1995). Despite the low levels of polymorphism and $H_o$ in many Northern Pike populations, Healy and Mulcahy (1980) detected greater $H_o$ in Swedish and Nordic populations than others in Europe and North America. Although allozyme studies were restricted to a relatively small number of loci, they were still useful for providing information about two measures of variation: (1) allelic polymorphism - the fraction of loci with greater than one allele out of the total loci examined), and (2) heterozygosity – the fraction of the population that is heterozygous at an allele). Allozyme studies and gave an important initial insight into large scale population structure of Northern Pike.

Microsatellite DNA (repeating elements of short nucleotide sequence) has also been used to assess population structure. Indeed, the observed heterozygosity values reported for microsatellite data are ten to one hundred times greater than those reported for allozyme data

(see Table 2). This is expected, in part, because of the ability to resolve nucleic acid fragments at a finer scale and with greater precision than polypeptides. An analogy is measuring in millimetres (nucleic acids) vs centimetres (polypeptides). Although microsatellites do not occur often in gene coding regions, the redundancy of the genetic code (i.e. the ability of more than one codon to specify the same amino acid), also allowed for deeper insight into heterozygosity. Microsatellites also are known to have more polymorphic loci than gene coding regions (i.e. the focus of allozyme studies) because of elevated mutation rates in repetitive DNA. Elevated mutation rates in microsatellites are attributable to an increase in DNA replication errors in these regions, and a relaxation in selection pressure. This contributes to greater variation between individuals and populations than can be seen with allozymes. Despite this greater capability to detect variation, Northern Pike microsatellites were also observed to be less variable than those in other species (Table 2) (Hansen et al., 1999; Miller and Kapuscinski, 1996, 1997; Miller and Senanan, 2003; Ouellet-Cauchon et al., 2014; Wang et al., 2011).

Authors have been able to distinguish smaller scale population structure in Northern Pike using microsatellite data. For example, Hansen et al. (1999) correctly assigned individuals to one of two Danish populations based on microsatellite patterns (Hansen et al., 1999). Another study found microsatellite markers associated with water drainage systems in Danish populations (Bekkevold et al., 2015). However, in central North America, defining population structure has been more difficult with microsatellites because of fewer numbers of polymorphic alleles and less heterozygosity within polymorphic alleles.

Further demonstrating low levels of variation in Northern Pike, Rondeau et al. (2014), reported only one SNP per 6 -10 kilobases while constructing the Northern Pike genome. In other organisms, an average of one polymorphic SNP every 1000 bases is observed in humans

(the 1000 Genomes Project Consortium, 2015), one every 750 bases in salmonids (Koop, 2018), one every 500 bases in Atlantic Cod (Star et al., 2011), and one every 300 bases in herring (Martinez Barrio et al., 2016).

**Table 1. Heterozygosity in Northern Pike populations.**

| Population | Observed Heterozygosity | Polymorphic loci | Author |
|---|---|---|---|
| **Allozyme** | | | |
| Ireland | 0 | 3 of 26 | Healy & Mulcahy, 1980 |
| England | 0 | | |
| Sweden | 0.02 | | |
| Swedish Baltic | 0.075 | | |
| Netherlands | 0.033 | | |
| US | 0.003 | | |
| Canada | 0.001 | | |
| Nebraska | 0.002 | 2 of 65 | Seeb et al., 1987 |
| Dakota | 0 | | |
| Wisconsin | 0 | | |
| Manitoba 1 | 0.002 | | |
| Manitoba 2 | 0 | | |
| Ontario | 0 | | |
| Saskatchewan | 0.002 | | |
| Alberta | 0 | | |
| Yellow Perch North America & Europe | 0.017 - 0.054 | 29 of 54 | Marsden et al., 1995 |
| **Microsatellite** | | | |
| Minnesota and Wisconsin | 0.07 – 0.31 | 4 of 9 | Miller and Kapuscinski, 1996 |
| Danish Populations | 0.11 - 0.79 | 9 of 13 | Hansen et al., 1999 |
| North America (including Alaska) | 0.10 – 0.27, | 8 of 13 | Senanan and Kapuscinski, 2000 |
| Finland | 0.48 – 0.50 | 8 of 13 | Senanan and Kapuscinski, 2000 |
| Ulungur, China | 0.15 – 0.85 | 18 of 48 | Wang et al., 2011 |
| Balaton, Hungary | 0.04 - 0.92 | 18 of 48 | Wang et al., 2011 |
| St. Lawrence | 0.03 – 0.97 | n/a | Ouellet-Cauchon et al., 2014 |
| Muskie - Great Lakes Region | 0.12 – 0.71; Average 0.50 | 13 of 13 | Turnquist et al., 2017 |
| Yellow Perch, Pennsylvania | 0.07 – 0.81 | 12 of 13 | Zhan et al., 2008 |
| Freshwater Fish | Average 0.46 | 7.5 of 75 | DeWoody and Avise, 2000 |
| Marine Fish | Average 0.79 | 20.6 of 66 | DeWoody and Avise, 2000 |
| Anadromous Fish | Average 0.68 | 11.3 of 43 | DeWoody and Avise, 2000 |
| Other Animals | Average 0.58 | 7.1 of 340 | DeWoody and Avise, 2000 |

In Northern Pike, low levels of variation are thought to be a product of small founding populations and small effective population sizes (the number of reproducing individuals). Both of these effects decrease variability and drive alleles to fixation. As top tier predators, Northern Pike are known to have low effective population sizes, and typically the largest are the most successful spawners, and the most likely to survive (Skov and Nilsson, 2018). Because of their position as top tier generalist predators, habitat expansion by only a few numbers of individuals can be successful. As such, expansion by a series of founding events by few individuals can lead to an extensive lack of variation (Skov and Nilsson, 2018).

The effect of introductions and stocking programs on Northern Pike population structure have also been considered, although not well resolved (Miller and Senanan, 2003). In Canada, Northern Pike introductions are usually extensions of the local native range, although some long distance transfers have been documented (Harvey, 2009). Northern Pike from Alberta and the Yukon were transferred to British Columbia in 1986 (Crossman, 1991), reportedly in small numbers and to locations where Northern Pike already inhabited nearby waters. However, the locations, exact numbers transferred, sex, and age are unknown. Concern that stocking of Northern Pike can lead to genetic homogenization has been expressed (Bradford et al.; DFO, 2011; Harvey, 2009). Yet the ability for stocked Northern Pike to contribute genetically to local population is unclear, as the age of the stock has been shown to influence their survivability and genetic contribution (Guillerault et al., 2018; Larsen et al., 2005; Skov et al., 2011). Skov et al. (2012) found that stocked Northern Pike fry were recovered in smaller number and smaller sizes at the end of the season than wild fry, despite being larger and more abundant at the beginning of the season. Larsen et al. (2005) reported a low rate of genetic introgression from stocked fry to a wild population, suggesting that stocked fry contribute little to long term genetic structure.

Guillerault et al. (2018) found that large, adult stock could persist in the environment for at least two years and be active spawners. Therefore, without knowledge of the age and origin of the Northern Pike stock, the effects of stocking on genetic variation is unclear, especially across great distances.

## Importance of the Northern Pike Reference Genome

In 2014, Rondeau et al. published the first construction of the Northern Pike genome, version 3. Since this publication, long-read and linked-read sequence data such as PacBio Sequel, 10x Chromium, and Hi-C have been produced and incorporated with the original paired end sequence data to build a greatly improved chromosome level assembly of the Northern Pike genome (Rondeau, personal communication). This new genome (version 4) is currently available on NCBI under GenBank assembly accession GCA_004634155.1. The addition of long-read and linked-read data improved the contiguity of the reference genome such that 97.6% of the sequence reads are arranged into chromosomes, and only the most difficult to resolve repeat regions are contained in unanchored scaffolds. This thesis utilizes this improved reference genome v. 4 as a basis for SNP discovery. A high-quality long-read reference genome such as this one is invaluable to dependable SNP discovery because duplicated genes, repeats, and complex regions are assembled unambiguously, resisting collapse of similar but distinct genomic regions. This allows for correct alignment of resequencing data and therefore reduces the occurrence of false positive SNP calls.

## Phylogenetic position of Northern Pike

Northern Pike are grouped in the genus *Esox*, in the family Esocidae, order Esociformes. The genus includes the sub genera *Esox* and *Kenoza*, representing pikes and pickerels, respectively. In addition to *E. lucius,* the sub-genus *Esox* contains *E. masquinongy* (Muskie or

Muskellenge) and *E. reicherti* (Amur Pike), to which Northern Pike is most closely related (Grande et al., 2004). Two additional species of pike have been recently described. In Italy, *E. cisalpinus*, was distinguished from *E. lucius* (Bianco and Delmastro, 2011; Lucentini et al., 2011), and in France *E. aquitanicus* (Denys et al., 2014). The pickerels (sub-genus *Kenoza)* has two species: *E. niger* and *E. americanus*, and one sub-species *E. americanus vermiculatus* (Figure 1). Northern Pike are the considered the type species for the genus *Esox*. In recent years, an increasing number of studies have been performed on Northern Pike, and as such this species is an emerging model organism in ecology and evolutionary biology (Forsman et al., 2015).

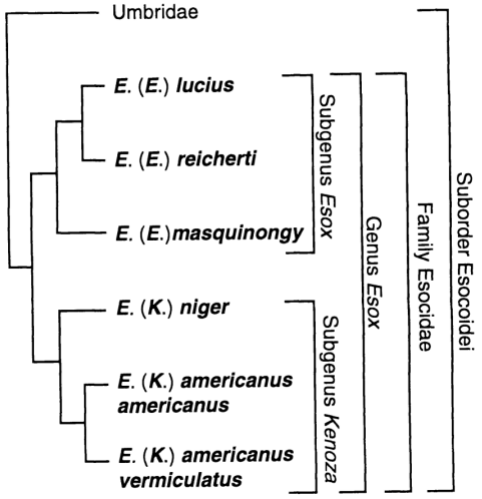



**Figure 1. Phylogeny of Esociformes**. From Grande, 1999. The recently described species *E. cisalpinus* and *E. aquitanicus* are not seen on this phylogeny as they have not been genetically integrated.

Analysis of microsatellite markers and mitochondrial DNA (mtDNA) sequences suggests that Northern Pike is a monophyletic group with three major lineages across Europe and North America. (Jacobsen et al., 2005; Skog et al., 2014). Skog et al. (2014) conclude that one lineage has a circumpolar distribution, while two lineages are restricted to Europe. Both studies detected a distinct group in southern Europe, which is thought to be the newly described *E. cisalpinus*.

## Phylogeography

The earliest esocid fossils were found in what is now North America about 100 million years ago (MYA) during the Cretaceous period, and esocids are thought to have radiated around this time point (Wilson et al., 1992). At this time, the Atlantic Ocean had not yet formed and North America was linked with Greenland and Eurasia forming Laurasia. Fossils of extinct *Esox* species have been unearthed in Alberta and Wyoming and dated to 65 and 56 MYA, respectively (Grande, 1999; Wilson, 1980). The earliest fossils of Northern Pike (*E. lucius*) were found in Germany and date to the Pliocene epoch (Grande, 1999), which occurred between 2.5 – 5.3 MYA. By this time, Eurasia and North America were well separated by the Atlantic Ocean, and continents were in their current positions. In North America, the earliest known fossils of Northern Pike date to the last Pleistocene, 126-12 thousand years ago (KYA). Wilson (1980) suggested that the modern *Esox* species found in North America today may be relics from ancient fauna. Indeed, the absence of million-year-old *E. lucius* fossils in North America does not exclude the possibility that they speciated in North America 65 – 56 MYA, and travelled to Europe while the continents were still connected. However, the current fossil record does not support this scenario.

The earliest known fossils of Northern Pike were identified in Germany and dated to 2.6 – 5.3 MYA. Many glacial expansions and retreats occurred after this date, as the Quaternary Glaciation began about 2.6 MYA (Ehlers and Gibbard, 2011). These glaciations re-shaped landscapes and drainage systems on a global scale (Jones et al., 2018; Sealy et al., 2016) and undoubtedly influenced the current day demography of Northern Pike. The most recent glacial period of the Quaternary Glaciation occurred between 20 - 11 (KYA). During this period, most of Canada and the northern United States were buried in ice, but parts of Alaska and much of Eurasia remained ice free. The most extensive ice coverage in Europe and Asia occurred before

this time (Ehlers and Gibbard, 2011). European populations of Northern Pike appear to show greater levels of variation than North American populations (Jacobsen et al., 2005; Senanan and Kapuscinski, 2000; Skog et al., 2014). This observation is supported by the geological history of the Northern Hemisphere, as populations in Europe and Asia would have re-colonized at an earlier time point than North American populations, and because of the complex mountainous geography, likely from a greater number of refugia (Senanan and Kapuscinski, 2000; Skog et al., 2014). Both recolonization from multiple refugia and additional time for the accumulation of variation could account for the greater levels of genetic variation in European Northern Pike. The relative lack of genetic variation in North American populations have led authors to suggest that North America was recolonized from one refugium (Senanan and Kapuscinski, 2000). However, microsatellite and mtDNA hint at the possibility of a second founding population in North America. Two studies found Alaskan populations to have unique genetic signatures that distinguished them from other North American populations (Senanan and Kapuscinski, 2000; Skog et al., 2014), hinting that Alaskan Northern Pike may have descended from a refugium in Beringia. Thus, questions still remain: was North America colonized by more than one founding population after the last ice age, and did these founding populations originate from Beringia, an ancient population that persisted south of the North American ice sheets, or both?

## Esociformes and relationship to Salmoniformes

Phylogenetic analyses based on mitochondrial sequence data and several nuclear loci place Esociformes and Salmoniformes as sister groups (Campbell et al., 2017; Ishiguro et al., 2003). This was a surprising revelation at the time, and prompted research into the evolutionary forces behind salmonid life histories (Ramsden et al., 2003). Based on expressed sequence tags (ESTs), Northern Pike were shown to be 89.6% similar to Atlantic Salmon (Koop et al., 2008).

Soon after divergence, about 80 MYA, salmonids experienced a whole genome duplication (WGD) (Macqueen and Johnston, 2014), that esocids clearly lacked (Leong et al., 2010; Rondeau et al., 2014). Thus, Northern Pike can serve as a valuable outgroup for the study of Salmonids, providing a basis for understanding evolutionary processes that occur after a genome duplication, as well as evolutionary forces affecting life history strategies.

Researchers have used the evolutionary relationship between esocids and salmonids to investigate genome re-diploidization after duplication. It is important to have a pre-WGD with which to compare evolutionary processes so that divergence rates between duplicated paralogs can be assessed (Leong et al., 2010). With Northern Pike as an outgroup, Leong et al. (2010) showed selection pressures on Atlantic Salmon paralogs (homologous genes resulting from a genome duplication). Rondeau et al. (2014) showed that synteny (the grouping of genetic loci on chromosomes) is conserved between each Northern Pike linkage group and two associated chromosomes on Atlantic salmon. These observations highlight the similarity between the esocid and salmonid genomes and the usefulness of the Northern Pike genome in studying the evolution of salmonids. Genetic resources generated through this work are available for public use and can further these investigations. Although no direct analytical comparisons are made here, similarities and differences between the sex determination systems in salmonids and Northern Pike are discussed, as well as the possible function of immune related genes in sex determination.

## Sex determination

The sex of individuals is a factor of major consideration in biological studies. Not only can traits and behaviours be influenced by an individual's sex, but the success of a population is

largely influenced by the relative abundance of males and females (Ancona et al., 2017). What, then, determines if an individual will develop as a male or female?

Sex is determined in early development when gonadal cells detect, or fail to detect, the presence of a signal that triggers a sexual differentiation pathway toward either male or female physiology. The signal can originate from genetic or environmental sources, and sometimes both. Most familiar is the model of genetic sex determination (GSD) in which the genotype set at conception determines male or female phenotype (Valenzuela and Lance, 2004). In this model, expression of genetic instructions stimulates the sex differentiation cascade, that, through complex pathways, initiates or suppresses the development of testes or ovaries in the developing embryo (Suzuki et al., 2002). This is the case in humans and almost all mammals, where the male-specific Y chromosome holds the gene responsible for initiating male differentiation: *sry* (Swain and Lovell-Badge, 1999). A gene that initiates the cascade of events toward male or female differentiation is called a master sex determiner (MSD)

Environmental sex determination (ESD) occurs when the primary sex of an individual is dependent on the state of its environment during a sensitive period in embryogenesis (Valenzuela and Lance, 2004). In this case environmental factors such as pH, social cues, stress, and most commonly temperature, instigate the sex differentiation cascade. Temperature is thought to strongly influence the activity of enzymes and receptors involved in the differentiation cascade (Devlin and Nagahama, 2002). Environmental sex determination has been observed in reptiles, amphibians, and fish. In the salt water crocodile, for example, incubation temperatures determine sex (Mohanty-Hejmadi et al., 1999). In goldfish, incubation temperatures around 20°C yield 95% female offspring, while temperatures around 30°C yield just 5% females (Goto-Kazeto et al., 2006).

Factors determining sex are diverse in fishes as both genetic and environmental factors have been observed to influence sex (Devlin and Nagahama, 2002). Genotypically, species can utilize a female homozygous, male heterozygous system (XX-XY), or a female heterozygous, male homozygous system (ZW-ZZ) (Heule et al., 2014; Mank et al., 2006), where the 'switch' for male or female differentiation are coded on the heterozygous sex. Genetic determiners of sex can be much more complex than this classical 'presence or absence' model of genetic sex determination. Systems can involve more than one gene one on more chromosome, or rely on a combination of environmental cues and genetic expression thresholds that initiate the differentiation cascade (Heule et al., 2014; Mank et al., 2006; Wu and Chang, 2013).

## Locating sex determining loci

Attempts to elucidate sex-determining genes or genomic regions requires extensive sequencing and knowledge of genetic markers. The process often begins with identification of either an XY or ZW heterogametic system and the identification of the Y or W chromosome. In the first investigations into sex determination, heteromorphic X and Y chromosomes could be identified through karyotyping. However, in most fish species, sex chromosomes are homomorphic (i.e. not visually distinctive) and cannot be identified through karyotyping. In these cases, linkage studies allowed for sex determining loci to be broadly identified and subsequently for specific regions to be identified. This was accomplished by associating sexual phenotype with the inheritance patterns of sex-linked traits or markers (Devlin and Nagahama, 2002; Matsuda et al., 2002). The theory behind this approach is that genetic or phenotypic markers located close to one another will likely be inherited together. Thus, if genes controlling phenotypic sex and genes of an associated phenotype (of known sequence) are tightly linked, they will be in linkage disequilibrium (non-randomly associated). Those low frequency cases

where tightly linked markers are not inherited together are highly informative as they indicate that recombination has occurred between the markers. The sequence of the known genetic marker can then be used to probe the region for sex determining genes in recombinant individuals (Matsuda et al., 2002). This can be accomplished through the use of bacterial artificial chromosome library creation – a lengthy and expensive procedure that involves rounds of cloning and sequencing.

More recently, the increasing availability and affordability of next generation sequencing has allowed for genome construction and analysis, which greatly facilitates investigations of sex determination mechanisms. In genome construction, recombination analysis is performed on sequence data from many related individuals of both sexes. This type of analysis tells us how closely genetic markers are linked and allows for genome mapping (i.e. allows sequences to be assembled in a manner that represents their position along the physical chromosome) (Gharbi et al., 2006; Rondeau et al., 2014). Because X and Y regions most commonly do not recombine, chromosomes that show low levels of recombination are candidates for sex chromosomes. Next generation sequencing makes it possible for whole genome data from many individuals to be obtained at a relatively low cost, and with a whole genome resource, these data can be quickly assembled and used to identify pools of sex-specific SNPs. Single nucleotide polymorphisms can be compared and used to pinpoint regions of the genome associated with sex, and point to a region where sex determination is encoded (Star et al., 2016).

In any case, the ability to detect a robust GSD system depends on the identification of sex-specific heterozygosity. Sex-specific heterozygosity is a condition where heterozygous SNPs are restricted to either male or female phenotypes; i.e. exclusively males or exclusively female heterozygous SNPs at particular loci. Additionally, only *aa* and *Aa* genotypes are present at these

loci. The presence of *aa*, *Aa*, and *AA* genotypes would negate the possibility of the locus having a sex determining effect, as it signifies that individuals of the same genotype are mating. A greater quantity and density of sex-specific heterozygous SNPs allows the detection of a GSD through sequencing and bioinformatics approaches.

Because older, more established sex determination systems have less recombination between sexes, more sex-specific mutations accumulate in the heterozygous sex (Abbott et al., 2017). Thus, older GSD systems have greater SNP densities and are easier to detect than newly acquired GSD systems. When a genetic sex determination system is overturned, environmental factors can influence sex determination and a combination of ESD and GSD can become the determiner, or, alternatively a new mutation in the same or another gene can begin to bear the function of sex determination. If a new GSD system evolves, it can begin with just one mutation. Over time, heterozygosity will accumulate at this region as recombination frequency diminishes (Abbott et al., 2017; Matsuda, 2018). As such, younger, more recently developed GSDs have fewer and possibly less dense sex-specific heterozygous SNPs, making detection more difficult and requiring a large number of individuals for confident detection.

## Sex determining genes in fish

In fish, eight different master sex-determining (MSD) genes have been identified (Table 2). These genes have been shown to be necessary for the development of testes based on knockout and transgenic experiments (Hattori et al., 2012; Kamiya et al.; Li et al., 2015; Matsuda et al., 2002; Myosho et al., 2012; Pan, 2017; Takehana et al., 2014; Yano et al., 2012). In the Medaka fish *Oryzias latipes*, it is the *dmy* gene that signals male differentiation (Matsuda et al., 2002) In another species of Medaka, *Oryzias luzonensis*, the gene *gsdfy* is the master sex determining gene (Myosho et al., 2012). In yet another Medaka species, *Oryzias dancena*, male

differentiation is initiated by *sox3y* (Takehana et al., 2014). In rainbow trout, the signal to induce male sex features is the encoded by the *sdy* gene (Yano et al., 2012). The Patagonian pejerry (*Odontesthes hatcheri*) uses a male-specific duplicate of the anti-Müllerian hormone (*amh*) gene, denoted *amhy*, to initiate male differentiation (Hattori et al., 2012). In the pufferfish *Takifugu rubripes*, a missense SNP in the *amh* receptor, *amhr2*, is only present in males and is the sole determinant of sex determination (Kamiya et al., 2012). In Nile tilapia, both temperature and the presence of a duplicated and modified *amh* gene in males determines sex (Baroiller et al., 2009; Li et al., 2015). While many of the species mentioned above use different MSD genes, most of these genes, with the exception of *sdy*, are known players in sex determination pathways (Navara, 2018). A growing number of MSD genes are variants of the *amh* gene. In mammals, this gene is known to induce regression of the Müllerian ducts in males during embryogenesis, and play a role in sexual development in both sexes. Fish do not have Müllerian ducts, but studies on the role of *amh* in fish have shown it to inhibit gonadal gem cell proliferation and steroidogenesis (Pfennig et al., 2015). The last exon of *amh* codes for transforming growth factor beta (TGF-β). This growth factor is well studied and can stimulate or inhibit cell proliferation (Morikawa et al., 2016). Intriguingly, other sex determining genes have links to the TGF-β pathway, highlighting the importance of this signalling cascade and *amh* in sex determination in fish.

**Table 2. List of known master sex determining genes in fish**.

| Species | Gene | Description | Reference |
|---|---|---|---|
| Medaka (*Oryzias latipes*) | *dmy* | *dmy* is only found in males. Mutation and reduced expression of the gene result in female offspring. | Matsuda et al., 2002(Matsuda et al., 2002) |
| Medaka (*Oryzias luzonensis*) | *gsdfy* (gonadal some derived growth factor on the Y chromosome) | *gsdfy* differs from *gsdfx* in 12 nucleotide positions. All are synonymous substitutions and the amino acid sequence of both are the same. Expression analysis suggest that these substitutions lead to higher expression of GSDF in males during sex differentiation. | Myosho et al., 2012 |
| Rainbow trout (*Oncorynchus mykiss)* | *sdy* (sexually dimorphic on the Y chromosome) on linkage group 1 | Found on the sex chromosome (linkage group1), *sdy* is only present in males. Future paper by Yano et al show this gene to be conserved on the Y chromosome for the majority of salmonids(Yano et al., 2013). | Yano et al., 2012 |
| Patagonian pejerry (*Odontesthes hatcheri*) | *amhy* (Y-chromosome specific *amh*) | Males have a duplicated copy of *amhy* not present in females. | Hattori et al., 2012 |
| Pufferfish (*Takifugu rubripes*) | *amhr2* (anti-Müllerian hormone receptor type II) | C/G SNP in *amhr2*. Males are exclusively heterozygous C/G. Females are homozygous C. SNP is a missense mutation changing Histidine to Asparagine (Asp) | Kamiya et al., 2012 |
| Indian ricefish (*Oryzias dancena*) | *sox3y* | Tissue specific regulatory element downstream of *sox3* initiates early expression in the gonads of males. | Takehana et al., 2014 |
| Nile Tilapia (*Oreochromis niloticus)* | *amhy* | There is a duplicated copy of *amh* on the Y chromosome. Both *amh* copies on the Y differ from the one copy on the X. A missense mutation on the downstream *amhy* (C/T) changes serine to leucine. | Li et al., 2015 |
| European Northern Pike (*Esox lucius*) | *amhby* | A duplicated copy of *amh* is located on linkage group 24 in males only, and is highly differentiated from the autosomal copy. | Pan et al., 2019 |

These genes have been shown to be absolutely necessary for sex determination either through knockout or transgenic experiments.

When this thesis was initiated in September 2017, the sex determination mechanism for Northern Pike had not been identified, and preliminary experiments suggested that European and North American Northern Pike may have different sex determination pathways. During my

studies, a Ph.D. thesis and subsequent paper were published identifying a duplicated copy of *amh* (termed *amhby*) in male Northern Pike, demonstrated that it is responsible for sex determination, and reported that this mechanism was lacking in North American Northern Pike (Pan, 2017; Pan et al., 2019). This result was obtained using a RAD-Sequencing (a reduced representation method), different from the whole-genome resequencing approach undertaken in this thesis.

In many other fish studies, the genomic region associated with a putative sex determining locus was obtained through the characterization of male or female specific SNPs, insertions, or deletions (Purcell et al., 2018; Rondeau et al., 2013; Star et al., 2016; Yano et al., 2013). Known sex determining genes are often observed in these areas where sex-specific variation is identified. For example, Rondeau et al., (2013) examined differences between male and female sablefish (*Anoplopoma fimbria*) and characterized sex-specific insertions upstream of the known sex determiner *gsdf*. The upstream location of sex-specific SNPs suggests differential gene expression and regulation patterns may play a role in sex determination (Rondeau et al., 2013). Purcell et al., (2018) identified a female-specific 61-base deletion in the California Yellowtail (*Seriola dorsalis*), upstream of the estradiol 17-beta-dehydrogenase 1 gene (*hsd17b1*). In males, this region contains binding sites for *Sry*, *Sox9*, and *Sox3* – all known to be involved in the sex determination pathway. This, along with the observation of greater heterozygosity in females in the sex determining region, lead the authors to nominate *hsd17b1* as the sex-determining gene in a ZZ-ZW system for *S. dorsalis* (Purcell et al., 2018). However, a known sex-determining gene is not always found in regions of sex-specific variation. In Atlantic Cod, for example, genotypic differences associated with sex were located outside of gene coding regions and were not in proximity to any known sex determining gene, preventing the authors from identifying a gene candidate in this area (Star et al., 2016). The authors suggested that an unknown sex

determination mechanism may be at play in Atlantic Cod. A recent study suggested sex-specific splice variants of an estrogen related gene located in the region identified by Star et al., (2016) is responsible for male sex determination in this species (Bao et al., 2019). Altogether, these studies highlight the diversity in sex determination mechanisms utilized in fish, and underscore the importance of species-specific investigations when researching sex determination mechanisms.

## Sex determination in Northern Pike

External sex determination of Northern Pike is unreliable. Although some physical characteristics appear to differ between the sexes, such as females having a deeper body and more slender caudal peduncles, these observations seem to be variable and more reliable in mature than juvenile fish (Casselman, 1974; Senay et al., 2017). Casselman (1974) describes a method in which the urogenital opening can be examined to determine sex externally. Females have an area of convoluted tissue between the anus and urogenital pore, and the entire area has a pinkish colour. This is not seen in males. Males have a transverse, slit-like depression just posterior to the urogenital opening, and the pigmentation is the same as the surrounding scales and tissue. Using these features for determining sex was accurate in the winter months, but was unreliable in the summer months and spawning season as the urogenital areas of males develop a pinkish colour. As Northern Pike mature in the first three years of life, the morphology of the urogenital region changes. Because close examination of an individual in the field is not always possible, and because of the variable morphology of the urogenital region at different life stages, a more accurate and less invasive was to determine sex would be desirable.

A study by Luczynski et al., (1997) suggested that European Northern Pike utilize an XX-XY male heterozygous sex determination system (Luczynski et al., 1997). Observations of North American Northern Pike populations have revealed both male and female biased sex ratios

(Huffman et al., 2014; Priegel and Krohn, 1975). Attempts by our own lab in partnership with Fisheries and Oceans Canada (Winnipeg, Manitoba) to rear Northern Pike resulted in a highly female biased sex ratio. In one mating, all 30 hatchlings were female (determined by dissection). In a second attempt, 14 of 18 progeny were female. These observations strongly suggest that the sex of Northern Pike is influenced by environmental factors, as a purely genetic sex determination mechanism would result in a 1:1 sex ratio. In 2014, Rondeau et al. published the Northern Pike genome and microsatellite-based linkage map, and reported an almost equal recombination rate between males and females (1.07:1) (Rondeau et al., 2014). As previously stated, Pan et al., (2017) identified *amhby* to be responsible for male sex differentiation in European Northern Pike. They concluded that the duplicated *amhby* was highly differentiated from the autosomal gene copy and was an ancient sex determination gene in Esocids. However, this gene could not be located in North American Northern Pike, consistent with previous observations. This suggests that the sex determination mechanism utilized by North American Northern Pike differs from that of their European counterpart. Northern Pike are considered one species across the Northern Hemisphere, so it is surprising that North American and European lineages utilize different sex determining mechanisms. As North American and European lineages have been separated only since the last glaciation (Skog et al., 2014), a divergence in sex determining mechanisms would be a recent evolutionary event. With our data, we hope to clarify some of these confounding observations and determine if evidence exists for a genetic mode of sex determination in North American Northern Pike.

## Benefits of identifying sex determination systems

Understanding sex determination systems has a myriad of benefits. Identification of genetic loci that are specific to males or females allows for development of simple diagnostic

laboratory tests, such as PCR screening, to be developed and applied. Such a test requires a small amount of tissue or fin clipping and thus provides a non-lethal and affordable method of sex identification. Such test have been developed and applied in aquaculture to maximize productivity (Li and Wang, 2017). These tests can also be applied in recreational fisheries and conservation management.

Knowledge of an individual's sex is a necessary piece of information in population genetics studies. Because genetic signatures can differ between males and females and failure to account for this difference can lead to incorrect conclusions about population structure (Benestan et al., 2017; Fowler and Buonaccorsi, 2016). This is demonstrated by Benestan et al. (2017), who falsely reported population structure between onshore and offshore populations of the American Lobster. By adjusting the number of males and females from each population included for analysis, they falsely amplified $F_{ST}$, such that comparing all females from one location with all males from the other resulted in a significant $F_{ST}$ value that supported isolated populations. When the sex ratio was balanced (an equal number of males and females were compared from each population) the authors obtained the correct result that the populations were panmictic. In situations where sex cannot be easily distinguished externally, the ability to identify sex via genetic methods is vital for conducting non-bias investigations into population structure.

The identification and comparison of sex determination systems provides insight into the pathways that control this pivotal event and the perishable nature of master sex determining genes. After a whole genome duplication, two copies of the sex determining loci are present. To avoid confounding signals, one of the loci must either become non-functional, or another master switch must take control of the system (Davidson et al., 2009). In the majority of salmonids, *sdy,* a gene with no previous history in the sex determination system, has become the master switch

that controls sex determination (Yano et al., 2013). This gene could not be detected in male or female Northern Pike (Yano et al., 2013). An understanding of the dynamics of sex determination in Pike may provide insight into the sex determination system in salmonids.

An additional benefit of sex determination studies is the potential to identify an environmentally influenced sex determination system or the lack of a genetically based sex determination system. In light of climate change, such a realization is pivotal for population management. Female-biased sex ratios have been reportedly increasing in sea turtles in response to warming temperatures, and data suggests that populations that nest along the beaches of the Great Barrier Reef may completely feminize (Jensen et al., 2018). The observation of skewed sex ratios in natural environments and laboratory-bred families suggests that an environmental factor may play a role in the sex determination pathway of Northern Pike.

## Summary of research questions

Here, we hope to address the following questions in regards to Northern Pike: What is the nature of the variation within the Northern Pike genome? How is it distributed and what genes does it affect? How does it group our samples phylogenetically? From where did Northern Pike colonize North America, and is there evidence for multiple refugia during the glacial maximum? Is there evidence for a genetic sex determination mechanism in Northern Pike across North America?

We will compare whole genome nuclear SNPs from 47 individual Northern Pike representing six populations from Alaska to New Jersey in attempt to answer these questions.

# Chapter 2

# Genetic variation and population genomics

## Summary

Northern Pike is an economically and ecologically valuable species with a circumpolar distribution across the Northern Hemisphere. They have been noted to have low levels of genetic variation despite their great capacity to colonize new environments successfully. Because of this low level of genetic variation, past population genetics studies have been unable to detect population structure. A high-resolution study of genetic variation is valuable to help understand the genome wide patterns of variation within this species in North America.

Resequencing data from 47 Northern Pike from across North America was used for SNP discovery and population analysis. These data have been uploaded to NCBI and will be available for public access under accession numbers SAMN10685075 – SAMN10685119. Sequences were aligned to the Northern Pike reference genome version 4 (Rondeau et al., in prep) and SNP discovery was performed. SNP and genotype counts reveal an extraordinary lack of genetic variation among Northern Pike, with a range of about 1 heterozygous SNP every 6,300 bases in Yukon River Drainage Basin Populations to 1 heterozygous SNP every 16,500 bases in all other North American populations observed here. Phylogenetic and principle component analyses show that individuals stratify by population into four main groups: 1 - Chatankia River (Alaska); 2- Hootalinqua (Yukon River); 3 - Palmer Lake (Northwestern B.C.); and 4 - Eastern North America (New Jersey, Upper St. Lawrence River, southern Manitoba, Charlie Lake in Northeastern B.C., and Castlegar, B.C.). Discriminant analysis of principal components (DAPC)

confirmed this structure and that the majority of variance is accounted for when samples are grouped as being from east or west of the great divide. This is supported by genome wide SNPs that are out of Hardy Weinberg Equilibrium (HWE) across North America, but within HWE when groups are analyzed independently. Genome wide Tajima's D, which is a measure of mutation-drift equilibrium and constant population size, is calculated to be very low, around -1.8 for Eastern North American Northern Pike, indicating this group has recently undergone a rapid population expansion. Comparison of $F_{ST}$ values between Eastern North American and Chatanika River pike show that populations are fixed at alternate alleles at a number of gene regions with functions related to immunity, cell structure and development.

Our results confirm that Northern Pike possess an extremely low level of genetic variation genome wide, and that Northern Pike from Alaska and the Yukon River harbour almost two times more heterozygosity than Northern Pike east of the Continental Divide. These results suggest that Alaskan Northern Pike are the oldest population in North America. Incredible genetic similarity, low levels of Tajima's D and low minor allele frequencies in Northern Pike across eastern North America suggest this is a younger population than that in Alaska. Following this result, we believe that extant populations of Northern Pike in North America originally colonized from Beringia and that a small founding population from the Yukon River drainage were the original colonizers of North America.

## Introduction

Nucleotide sequences have been studied for decades with intentions of elucidating genetic relationships among individuals, populations and species, and deciphering links between genotypes and phenotypes. Ongoing advances in sequencing technology and bioinformatics tools have allowed these studies to advance from small fragments to genome-wide analyses. A major benefit of genome-wide analysis is the incorporation of all nucleotide bases and polymorphisms present in an individual (assuming the complete genome was able to be sequenced and aligned). This overcomes a potential bias that can be introduced when inferring relationships from small nucleotide fragments or parts of the genome that may be under different selection pressures. Another benefit is the ability to perform investigations such as genome-wide association studies that link an area of the genome with a phenotype without prior knowledge of the associated gene or genomic region influencing the phenotype of interest. The resolution of whole genome data allows for the detection of small genomic regions that may contribute significantly to phenotype or population structure, but could potentially be overlooked by other methods that sample the genome. It also allows us to comprehensively quantify polymorphism on a genome-wide scale, and visualize how and where it is distributed throughout the genome.

In every genetic study conducted on northern Pike, authors have noted the low level of genetic variation present in this species. However, principles of evolution dictate that genetic variation is central for the ability of species to adapt and succeed in the face of environmental change and instability. Northern Pike challenge this notion as they are at least 2.5 million years old and have persisted through periods of intense climate change while attaining a distribution across the entire sub-arctic. This begs the question of how Northern Pike are able to be so successful despite the reported lack of variation. Currently, a high resolution, genome-wide analysis of Northern Pike genetics has not been performed.

The origin of Northern Pike in North America has been debated in the past. Some authors suggest that Northern Pike speciated in North America before the continents drifted apart some 100 MYA, while others suggest that Northern Pike speciated in Europe and colonized North America through Beringia. Fossil evidence suggests that Esocids and *Esox* spp. (representing pickerels) were present in North America about 60 MYA, but the earliest fossils of specific *Esox lucius* as a species are reported to be from Austria and Germany, and date to 2.6 – 5.3 MYA (Grande, 1999). Several genetic analyses performed to date have been unable to resolve possible origins due to low variation and reduced representation genetic methods.

Here, we use over 1,000,000 polymorphic SNPs obtained from whole-genome data to analyze the level and distribution of genetic variation in Northern Pike from across North America. We determine the relationships among populations of North American Northern Pike through variant quantification and distribution, analysis of phylogenetic relationships, PCA, DAPC, tests for neutrality (Tajima's D), genome wide mapping of genotype frequencies, nucleotide diversity ($\pi$), and Wright's fixation index ($F_{ST}$), and gain insight into the origin of North American Northern Pike.

# Methods

## Samples

Tissue samples from 47 Northern Pike from across Canada and the northern United States were provided by collaborators and from hatcheries as per Table 3 and Figure 2. We included the reads from the specimens used to generate the previous and current version of the reference genome (Rondeau et al., 2014). These fish came from Charlie Lake, British Columbia, Canada (version 1) and Castlegar, British Columbia, Canada (Rondeau et al., in prep.).

## DNA extraction and sequencing

DNA was extracted from a variety of tissues using DNEasy Blood and Tissue Kit (QIAGEN) following the manufacturer's protocols. Extracted DNA was quantified via Nanodrop ND-1000 (Thermo) spectrophotomer and Qubit® 2.0 (Life Technologies) Fluorometer. Samples were sent for sequencing to McGill University and Genome Quebec Innovation Centre, where 35 of the 47 samples underwent PE-150 (paired-end 150bp) PCR-free whole genome shotgun sequencing. Ten samples (the Alaskan population) were sequenced via PE-150 PCR shotgun sequencing because the amount of DNA extracted was insufficient for PCR-free libraries. All libraries were created on an Illumina HiSeq X Ten platform. Samples were pooled such that 5 – 7 samples were sequenced per lane. Lanes were designated for male or female samples exclusively as much as possible in order to reduce the possibility of index switching. Reads from the two reference individuals were sequenced as described in Rondeau et al. (2014) and in Rondeau et al. (in prep.).

**Table 3. Origin of Northern Pike samples.**

| Population | Latitude | Longitude | Males | Females | Total |
|---|---|---|---|---|---|
| Chatanika River, Alaska, U.S.A. | 64.98396 | -148.86032 | 5 | 5 | 10 |
| Hootalinqua, Yukon Territory, CANADA | 61.511562 | -135.132089 | ? | ? | 5 |
| Palmer Lake, British Columbia, CANADA | 59.43708 | -133.57592 | ? | ? | 4 |
| Charlie Lake, British Columbia, CANADA* | 56.32853 | -120.97835 | 1 | - | 1 |
| Castlegar, British Columbia, CANADA* | 49.31538 | -117.65344 | - | 1 | 1 |
| Whiteshell Hatchery, Manitoba, CANADA | 49.80051 | -95.17243 | 3 | 3 | 6 |
| St. Lawrence Waterway, New York, U.S.A. | 44.247868 | -76.097851 | 6 | 5 | 11 |
| Hackettstown Hatchery, New Jersey, U.S.A. | 40.84155 | -74.83359 | 6 | 3 | 9 |
| Total | | | 21+? | 17+? | 47 |

(?) Indicates that sex was unknown. (-) Indicates that no samples of this sex were collected. (*) Indicates Northern Pike that were used to build reference genomes.

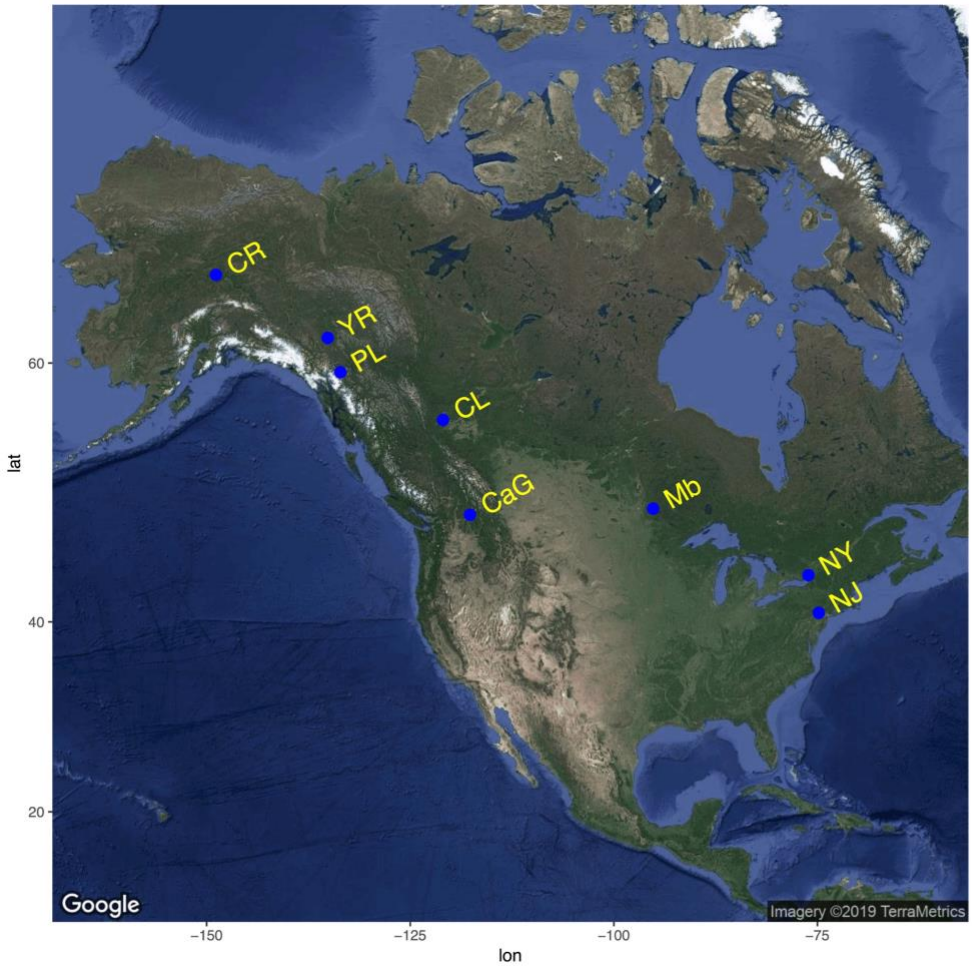



**Figure 2. Map of sampling locations across North America.** Y-axis is longitude and x-axis is latitude. CR = Chatanika River, Alaska. YR = Hootalinqua, Yukon Territory. PL = Palmer Lake, British Columbia. CL= Charlie Lake, British Columbia, CaG = Castlegar, British Columbia. Mb = Whitsehell Hatchery, Manitoba. NY = St. Lawrence Waterway, New York. NJ = Hackettstwon Hatchery, New Jersey.

## Read processing and variant calling

*Alignment*

Read processing and variant calling was based on GATK's best practices, and for all steps involving GATK we used version 3.8-0-ge9d806836 (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). Paired end reads were aligned to the Northern Pike genome version 4 (Rondeau et al., in prep) (WGS accession AZJR03000000.2; Assembly accession GCA_000721915.2) using the Burrows-Wheeler Aligner (BWA) version 0.7.13-r1126 (Li and Durbin, 2009) with the "-mem" algorithm, as recommended for reads of our length. Alignment files were piped to Samtools version 1.3 (Li et al., 2009b), converted to binary alignment/map (BAM) format, then sorted and indexed according to position. Information detailing the sequencing platform and multiplexing layout was incorporated and used to mark duplicates with Picard version 2.17.11 (Broad Institute, 2017). Because the reference individuals had read depths 5 – 7 times greater than our samples, we down-sampled their BAM files using the Samtools "view -s" command in order to obtain files that had similar depth to the rest of our samples. This was necessary in order to ensure downstream filters were working appropriately and to manage analysis time.

*Base quality score recalibration*

The purpose of base quality score recalibration is to correct known biases and systematic errors that occur in high throughput base quality scoring (McKenna et al., 2010; Van der Auwera et al., 2013). Recalibration broadens the range of quality scores, improves their accuracy, and closes the gap between empirical and reported scores. Base quality score recalibration requires known databases of SNPs and insertions and deletions (InDels) from which to base the recalibration algorithm. For Northern Pike, such sets were unavailable, so we produced

stringently filtered SNP and InDel files to fulfill this requirement following GATK's suggestions for non-model organisms. With GATK we called SNPs and InDels on each individual using HaplotypeCaller in GVCF mode and combined these calls as a cohort via GATK's GenotypeGVCF command. For each sample, HaplotypeCaller identifies regions of variation and reassembles the reads in this region. It then determines the likelihood of the reassembly given the reads, and assigns the sample a genotype based on greatest likelihood. Through GATK's GenotypeGVCF command, each sample's output from HaplotypeCaller is combined. Based on likelihoods, the merged file is re-genotyped and re-annotated to produce the working Variant Call Format (VCF) file, which contains all the variant sites identified in the genome, and each individuals' genotype, along with other information detailing the support for the genotype assignment. This VCF file is the starting point for downstream analysis. SNPs an InDels were extracted into separate files and each subject to a hard filter (GATK) that removed variants according to the following parameters (thresholds in brackets): quality by depth (2), fisher strand bias (60), root mean square mapping quality (30), mapping quality rank sum test (-12.5), and read position rank sum test (-8.0). We then applied the following quality control filters through VCFtools version 0.1.15 (thresholds in brackets): minor allele frequency (0.2), minimum mean depth (10), maximum mean depth (60), and minimum quality (30). Definitions for these filtering parameters can be found in the glossary. We used an in-house repeat library (Minkley et al., in prep) in BED format to exclude calls in repetitive regions using the VCFtools "--exclude-bed" function. As the current genome annotation was not available at this time, we created a file specifying transcribed regions using Standalone BLAT v. 34 (Kent, 2002). We then used blast2bed (nterhoeven, 2017) to convert the output blast file into a BED file, which is a text file that defines the co-ordinates of genomic features. We used this BED file to map the

transcriptome from the previous genome version to the current genome. We discarded calls outside of transcribed regions using the VCFtools "--bed" command. The resulting stringently filtered SNP and InDel files were used to recalibrate base quality scores in our alignment files with GATK's base quality score recalibration procedure.

### *SNP calling and filtration*

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

SNPs (1,363,731) were extracted and filtered according the parameters in Table 2.2. Through GATK we applied the same hard filter as described under Base Quality Score Recalibration. We applied further quality control filters through VCFtools version 0.1.15 (Table 2.2). Beyond standard quality control filters, we removed sites where more than 10 individuals were missing calls in order to ensure our analyses represented the majority of individuals (--max-missing-count 10). We applied a minor allele frequency filter of 1 (--mac 1). This filtered out sites that were a combination of only homozygous alternate alleles and missing calls. Finally, we applied a filter that required at least one of the calls to be homozygous (variant or reference). The VCF file produced after the last filtration step was the central file for our analyses. Any further filtration steps specific to particular analysis are discussed in their corresponding section.

## Population analysis

### *Heterozygosity and variation analysis*

The number of variant calls per individual was calculated with Real Time Genomics' RTG Tools version 3.9.1 (2018). An analysis of variance (ANOVA) and the Tukey post-hoc test

were performed on the number of heterozygous counts in individuals in different populations using the R statistical software "stats" (Chambers et al., 2017).

Genotype counts per site were obtained through GATK's "VariantsToTable" command for the following call categories: heterozygous, homozygous reference, homozygous variant, no call, total variants called, number of samples called. To obtain per site genotype frequencies, each category was divided by the number of samples called. Using VCFtools, we obtained nucleotide diversity values per site (Danecek et al., 2011). We assumed that at all positions where no variants were called (i.e. the remainder of positions in the genome not called in our VCF file), the variant call frequencies and nucleotide diversity values were zero and the homozygous reference frequency was 1. To highlight regions of heightened variation, we calculated a windowed mean of genotype frequencies and nucleotide diversity values across each linkage group using a 10 kb window and a 5 kb step including all sites using R v.3.5.2 (R Core Team, 2018) with the package zoo v. 1.8-1 (Zeileis and Grothendieck, 2005). All analyses done with R used the stated version. Results were plotted across the lengths of the linkage groups. Regions showing increased areas of variation were visually confirmed using the Broad Institute's Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

### *Phylogeny*

A maximum likelihood tree based on genome wide SNPs was generated using SNPhylo v. 20140701 (Lee et al., 2014). Default parameters were used, except to specify to perform 100 bootstraps. The resulting tree was visualized through Figtree version 1.4.3 (Rambaut, 2007), and rooted by midpoint.

### *Principal component analysis*

We used the R software package SNPRelate v.1.14.2 package (Zheng et al., 2012) to perform and plot a PCA using genome wide SNPs. For this analysis, we applied a minor allele count threshold of 2 to our VCF file. This excluded unique SNPs (SNPs only present in one individual) from the analysis, as these are unlikely to contribute to wider population structure.

*Discriminant analysis of principal components*

A DAPC was performed with bi-allelic genome wide SNPs using the R software package Adegenet v. 2.1.1 (Jombart, 2008; Jombart et al., 2010). Adegenet's "find.clusters" function grouped our samples into 4 clusters based on the lowest Bayesian Information Criterion value when all principal components were kept. We performed the DAPC on the groups identified by the "find.clusters" function, and retained 24 principal components and all three discriminant functions. We then used the "snpzip" command with the Ward clustering method to return lists of SNPs that had the greatest contribution to each of the three discriminant axis identified in the DAPC. We performed additional DAPCs on these high loading SNPs in order to compare their clustering patterns with that of the DAPC performed on all SNPs.

*Hardy-Weinberg equilibrium*

Given the absence of evolutionary influences such as genetic drift, selection, non-random mating, and mutation, the Hardy-Weinberg equilibrium (HWE) principle states that allele and genotype frequencies remain constant within a population. At each bi-allelic locus, this is defined by the following equation:

$$p^2 + 2pq + q^2 = 1$$

Where $p$ is the frequency of allele *A*, and $q$ is the frequency of allele *a*. When the frequencies of $p$ and $q$ are known, as they are with our SNP data, one can calculate the expected genotype frequencies and check for deviations from the HWE using a chi-square test. Deviations from the

expected values indicate that one or more of the evolutionary forces mentioned above is acting on a population. We used VCFTools to carry out this test at each bi-allelic SNP site on all of our samples together, and on the groups defined by PCA and DAPC analysis. We visualized the negative log of the *p*-values on a Manhattan plot generated with the R package qqman (Turner, 2017). The significance level was set using the Bonferroni correction.

*Tajima's D*

Tajima's D is a test statistic that can help characterise selection or demographic processes shaping genetic patterns in populations. It is based on the comparison of observed nucleotide diversity ($\pi$) to the expectation of what the nucleotide diversity should be ($\theta$), given the number of segregating sites, the number of sequences being compared, and the assumption that the individuals are evolving neutrally (mutation – drift equilibrium) at a constant population size (Tajima, 1989). When the population is evolving neutrally and not expanding nor contracting, genetic variation is maintained, $\pi$ and $\theta$ are equal, and Tajima's D is zero. When a population is losing genetic variation, or has less genetic variation than expected, $\pi$ is less than $\theta$ and Tajima's D is negative. Loss of genetic variation can come from population expansion after an extended bottleneck or from positive selection. When a population is gaining genetic variation, or has more variation than expected, $\pi$ is greater than $\theta$, and Tajima's D is negative. Excess genetic variation can be due to balancing selection or population stratification. The Tajima's D statistic does not fit a normal distribution, and so estimating the significance of the statistic can require one to build a distribution from which to assess significance. Because of the way the statistic is calculated, values for Tajima's D fall between -2 and 2 most (~95%) of the time. Values outside of this range are generally taken to be significant (Simonsen et al., 1995). When one has data for a large number of sites such as our data, another method to assess the significance of Tajima's D

is to plot the statistic along the length of the DNA sequence and assess trends and deviations. We took the later approach here. We used VCFTools to calculate Tajima's D in bin sizes of 10,000 for all of our Pike together and then for each group defined in PCA analysis. We then plotted this data along each chromosome.

### *Wright's fixation index ($F_{ST}$)*

Based on the clustering results from our phylogeny, PCA, and DAPC, we calculated Fixation index ($F_{ST}$) per site between each group using VCFtools (Danecek et al., 2011), which uses the Weir and Cockerham method (Weir and Cockerham, 1984):

$$F_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})}$$

where $\sigma_p^2$ is the variance in allele frequency between subpopulations and $\bar{p}$ is the mean allele frequency across all populations. Any negative values were assumed to be zero. We plotted the genome-wide $F_{ST}$ comparison between the two largest and most differentiated groups in our data set to identify regions of the genome where the groups were stratified. We visualized these areas in IGV and noted genes that were affected by differentiating SNPs in the two populations. We expect elevated regions of $F_{ST}$ to coincide with regions that are out of HWE.

# Results

## DNA sequence processing and variant discovery

Whole genome sequencing returned between 54 and 122 million reads per sample. Quality control checks on raw sequence data were performed by Genome Quebec and reviewed in-house to assure sequencing quality. After alignment to the reference genome, removal of duplicate reads, base quality score re-calibration, SNP calling and filtration, we obtained one VCF file that contained 1,910,789 SNPs and InDels across all 47 specimens. From this file, 1,363,731 SNPs were extracted. Filtering (Table 4) reduced the number of SNPs for use in analysis to 1,127,923 (Table 4, Figure 3).

## Raw SNP counts

We used RTG tools to count the number of heterozygous and homozygous alternate variants in each individual. We found a strikingly low number of heterozygous sites across the genome. On average, Northern Pike in North America possess one heterozygous SNP per 10,000 bases. The Castlegar individual had an incredibly low number of heterozygotes– just 14,565 (1 heterozygote every 50,000 bases). Chatanika River and Hootalinqua Northern Pike have significantly more heterozygous SNPs than Palmer Lake and all populations east of the Continental Divide (Figure 4, $p < 0.001$). Remarkably, they have 2 – 3 times as many heterozygous SNPs than any other population or individual (Figures 4 and 5).

Alaskan, Yukon, and Palmer Lake Northern Pike have the highest number of homozygous alternate SNPs compared to the reference genome (Figure 5), indicating that these populations are less closely related to the reference than the Charlie Lake, Manitoba, New York

**Table 4. SNP filtering parameters.**

| Filter Description | Program | Number of Variants Remaining |
|---|---|---|
| No Filter | - | 1,910,789 (SNPs and InDels) |
| Extract SNPs | GATK | 1,363,731 |
| GATK Hard Filter | GATK | 1,189,068 |
| Minimum Quality 20 | VCFtools | 1,186,793 |
| Minimum Mean Depth 10 | VCFtools | 1,153,668 |
| Maximum Mean Depth 60 | VCFtools | 1,152,122 |
| Max Missing Count 10 | VCFtools | 1,129,884 |
| Minor Allele Count 1 | VCFtools | 1,129,701 |
| All Heterozygote Filter | R/ VCFtools | 1,127,943 |

Extracting SNPs and applying filters removed 17% of variants and left 1,127,943 variants remaining for use in analysis.

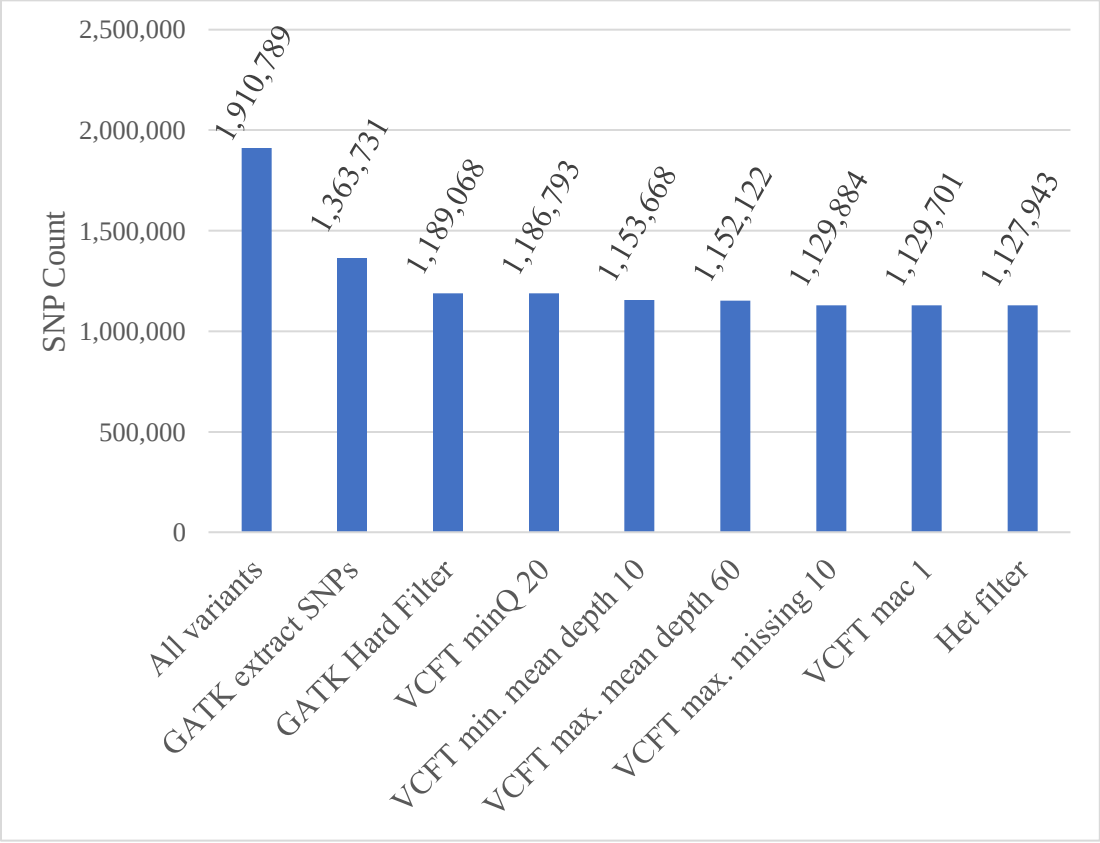



**Figure 3. SNP filtering barchart.** GATK = Genome Analysis Tool Kit. VCFT = VCFTools.

and New Jersey populations. As expected, we see almost no homozygous alternate SNPs in the reference individual.

Counting SNPs in adjacent bins of 10kb highlights the location of dense regions of variation and differentiation as shown in Figure 6. Most of the 10kb bins contain between 0 – 50 SNPs, however in some regions the density increases to a maximum of 400 SNPs per 10kb. Seven regions of high SNP density were identified and genes in these regions are described in Figure 6. The majority of genes in SNP dense regions have immune related functions.

A

| Comparison | *p* |
|---|---|
| CR-Mb | <0.001* |
| CR-NJ | <0.001* |
| CR-PL | <0.001* |
| CR-NY | <0.001* |
| CR-YR | <0.001* |
| Mb-NJ | 0.006* |
| Mb-PL | 0.676 |
| Mb-NY | 0.029* |
| Mb-YR | <0.001* |
| NJ-PL | 0.497 |
| NJ-NY | 0.954 |
| NJ-YR | <0.001* |
| NY-PL | 0.853 |
| YR-PL | <0.001* |
| YR-NY | <0.001* |

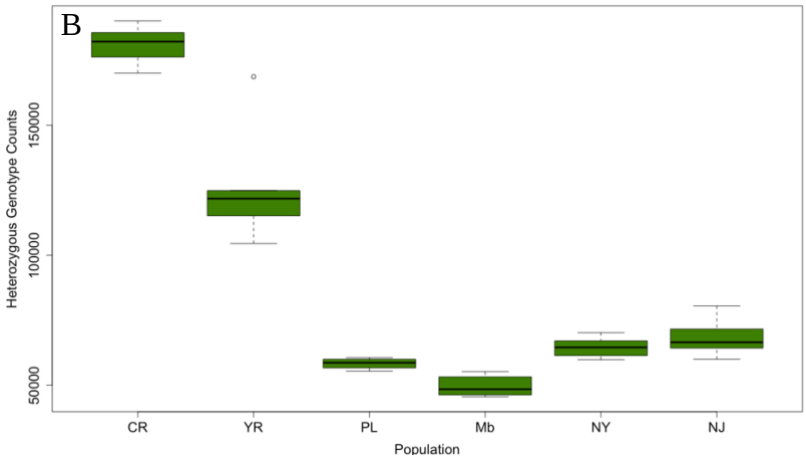



**Figure 4. Results of Tukey post-hoc test and box plots of heterozygous SNP counts per population.** Pike from Charlie Lake and Castelgar are not included in this analysis because n = 1 in these locations. (A) Tukey Post-hoc Test. Significant results are marked with an asterisk (*). (B) Boxplots.

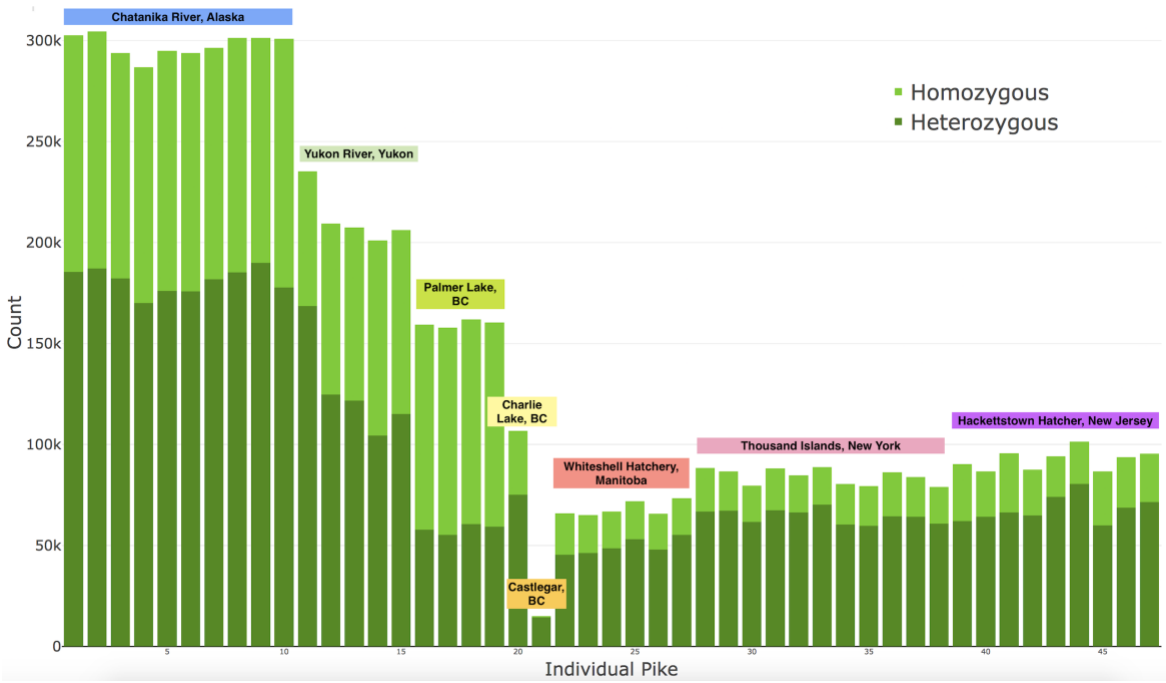
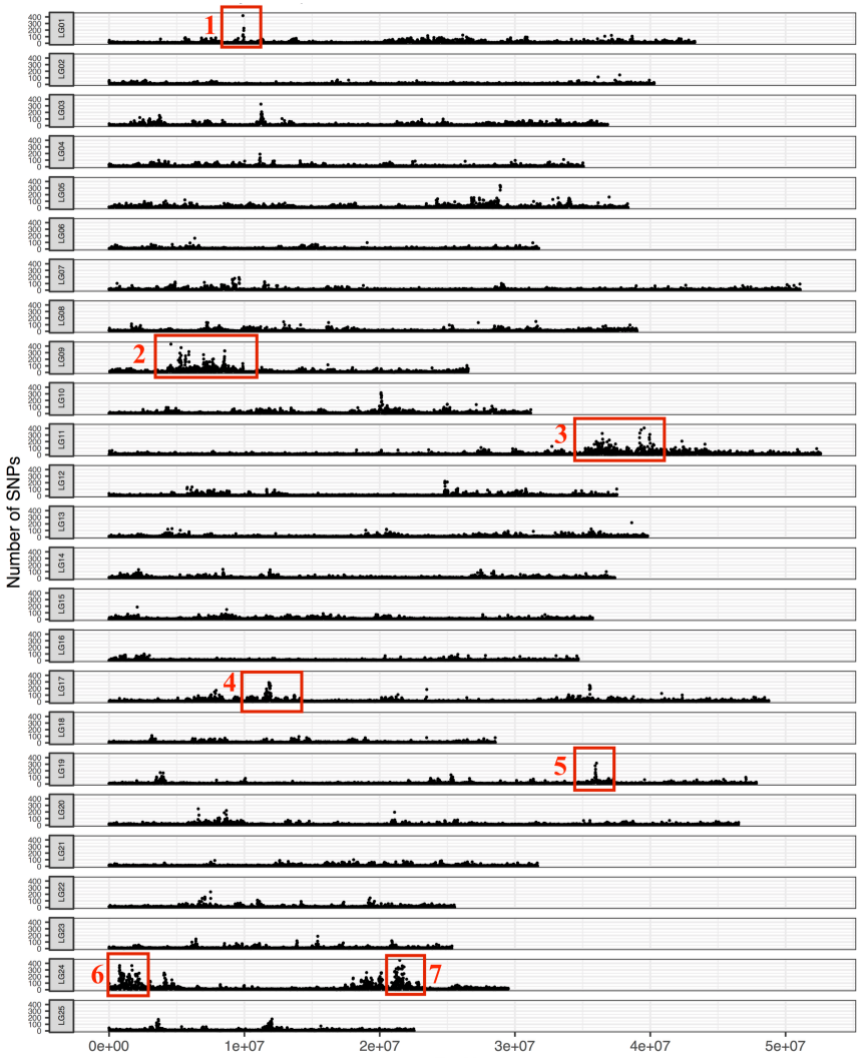



**Figure 5. Counts of heterozygous and homozygous SNPs in each individual**. The Castlegar, BC individual was used to create the reference genome version 4.




| # | LG : Region (MB) | Region Description |
|---|---|---|
| 1 | LG01 : 9.875 – 9.975 | Densely grouped heterozygous SNPs that overlap the collagen homolog, galaxin-like protein coding gene. |
| 2 | LG09 : 5 – 10 | Gene-rich region where SNPs are segregated by population. Both heterozygous sand homozygous alternate SNPs present. Many genes in this region are immune related. |
| 3 | LG11 : 35 – 40 | Dense, gene-rich region of SNPs that mainly occur in Chatanika and Hootalinqua. Multiple copy number variants of immune and inflammation related genes. |
| 4 | LG17 : 11.5 – 12.0 | Dense region of SNPs in all populations. Gene rich region. Most genes here are of unknown function. Two genes are voltage related and three are DNA binding. |
| 5 | LG19 : 35.8 -37.0 | SNPs are concentrated in Palmer Lake, Hootalinqua, and Chatanika River populations. Region contains multiple copy number variants of genes relation to development and proliferation |
| 6 | LG24: 0.6 – 2.4 | Male specific SNPs in Chatanika River. Immune related genes and others. |
| 7 | LG24 : 21 – 22 | SNPs occure in Chatankia River and Hootalinqua populations and affect immune related genes. |

**Figure 6. Distribution of SNPs across linkage groups.** The x-axis is linkage group position in base pair number. The y-axis is the number of SNPs per 10kb bin. Numbers and boxes denote positions of elevated SNP density. The accompanying table details the nature of the SNP dense region and notes associated functional themes.

## Phylogeny

Our tree clustered individuals according to their geographic locations (Figure 7). Individuals from Alaska, Hootalinqua (Yukon River), and Palmer Lake are more genetically similar to each other than they are to individuals from populations East of the Continental Divide (Charlie Lake, Castlegar, Manitoba, New York, and New Jersey). Almost all of the nodes identify separate local populations and are well supported with bootstrap values equal to or greater than 90/100. The exception to this is the topology of the Hootalinqua population. Here, Northern Pike are not clustered distinctly into one clade as the other populations are. Instead, Hootalinqua pike bridge the Chatankina River clade to an internal node which extends to the Palmer Lake clade on a terminal branch and to the eastern North American clades through the root of the phylogeny. The Chatanika River clade appears to be more closely related to just one of the Hootalinqua pike, and the placement of sample YR5 is supported by a bootstrap value of just 53. This suggests that other possible topographies for Hootalinqua pike are possible and indicates that Northern Pike from this region are less clearly related than Northern Pike sampled at other sites.

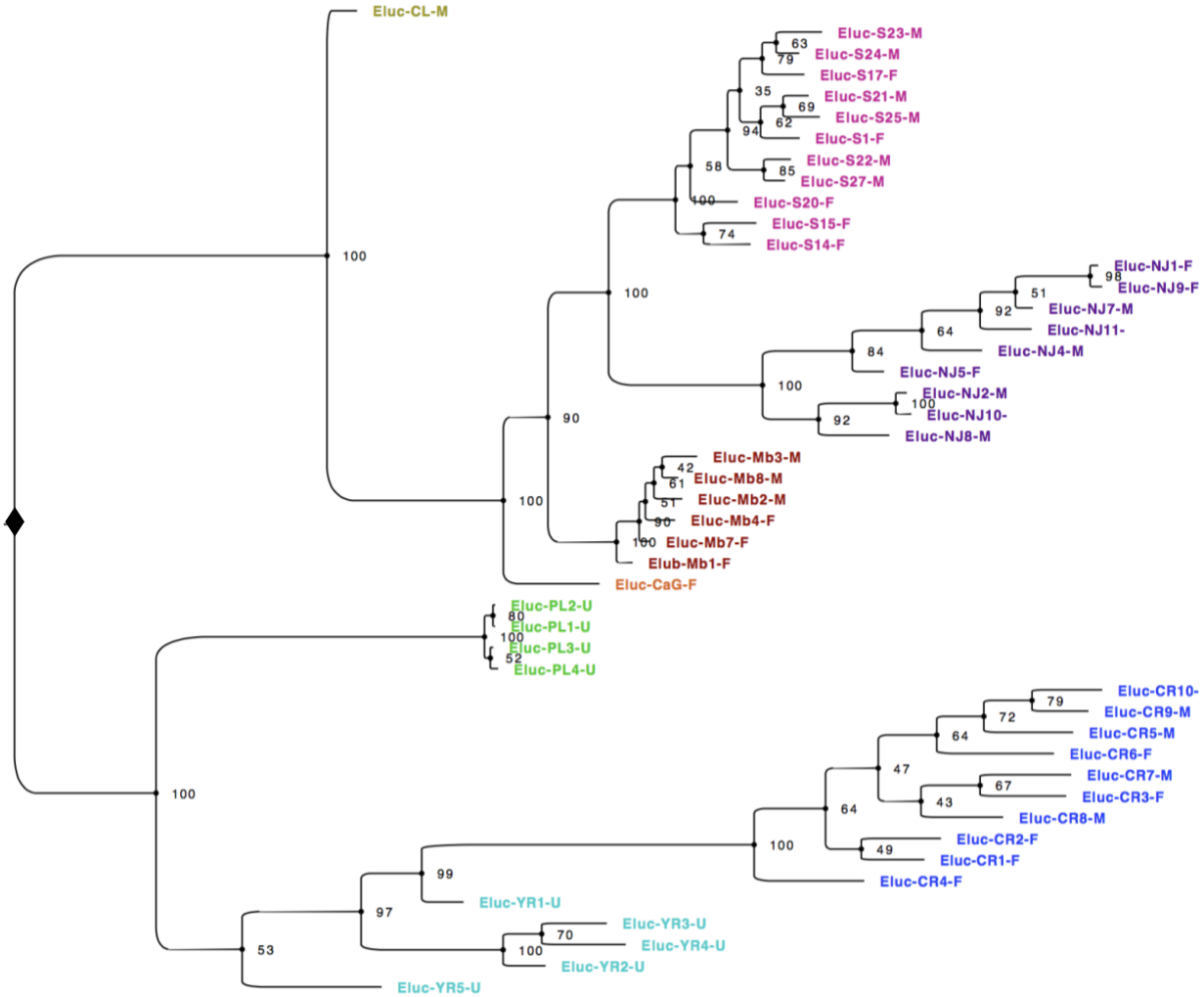



**Figure 7. Maximum likelihood phylogeny.** Tree was constructed using SNPhylo. Nodes are marked by circles. Number on branches indicate bootstrap value out of 100. Midpoint is marked by the diamond. Populations groups as follows: Alaska in blue, Yukon River in teal, Palmer Lake in green, Charlie Lake in dark yellow, Castlegar in orange, Manitoba in red, New Jersey in purple, and New York in pink.

## PCA

Principle component analysis (PCA) grouped individuals into four distinct clusters: Chatanika River (Alaska), Hootalinqua (Yukon River), Palmer Lake, and the remainder of North America (Figure 8). This analysis applied a minor allele count threshold of 2 to our data thereby reducing the number of SNPs for PCA analysis from 1,127,943 to 672,565. Despite the relative physical proximity of Charlie Lake to Palmer Lake and Alaska, Charlie Lake groups with Eastern North American populations.

The first two principal components (PCs) account for 40% percent of the variation in the data. The third PC accounts for 6.6%, and PCs thereafter account for less than 3% each. In contrast to the phylogenetic tree, the PCA clearly separates the Hootalinqua and Alaskan populations. In addition, populations from Eastern North America (Charlie Lake, Manitoba, New York, and New Jersey) are clustered tightly together in the principle component analysis, which is not apparent in the phylogenetic tree. PC1 represents a south/east to north/west distribution across North America, while to biological meaning of PC 2 is unclear, and could possibly represent consequences of random genetic drift in small populations.

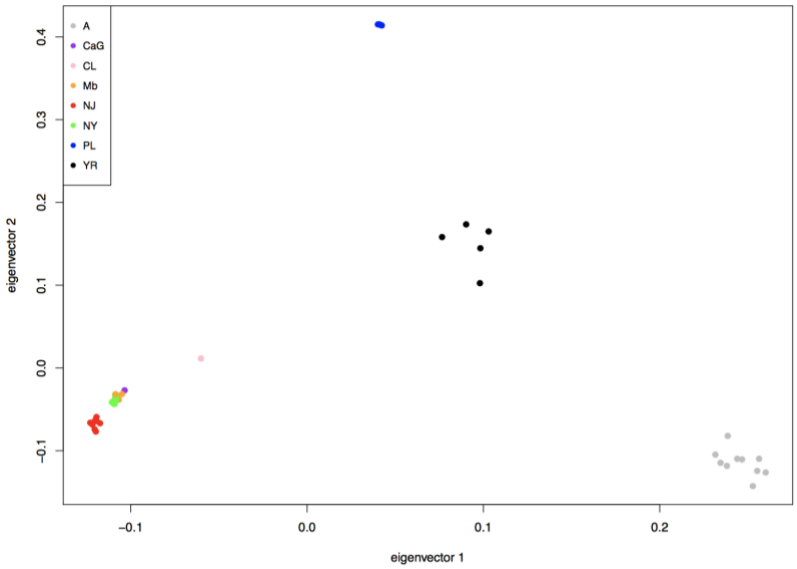



**Figure 8. PCA plot.** Four major clusters can be distinguished: grey – Chatanika River(1); black – Hootalinqua (2); blue – Palmer Lake (3); red – New Jersey, green – New York, orange - southern Manitoba, purple – Castlegar, and pink – Charlie Lake, all in group (4).

## DAPC and high-loading SNPS

Pike are separated into four distinct clusters by DAPC analysis. Chatanika River (Alaska), Hootalinqua (Yukon River), and Palmer Lake individuals each group into their own distinct cluster. The remaining individuals from Castlegar, Charlie Lake, Manitoba, New York and New Jersey all group together into the fourth cluster (Figure 9, plot A). This confirms the clustering seen in the PCA. Adegenet's command "snpzip" returned three loading plots (one for each discriminant axis retained) that rank each SNPs' contribution to separating clusters (Figure 9, plots B, D, and F). Performing DAPCs on the high loading SNPs from each linear discriminant allowed us to visualize grouping patterns associated with each axis (Figure 9, plots C, E, and G)

We used a discriminant analysis of principle components to not only to verify clustering but also to identify SNP sites that segregate populations. From the "snpzip" command, 309 SNPs were identified as the highest contributing SNPs to the first discriminant axis. These SNPs group

Chatanika River (Alaska) and Hootalinqua (Yukon River) Northern Pike together, and group Palmer Lake with the rest of North America (Figure 9, plots B and C). Of the 309 high loading SNPs, 249 are located on linkage group 21 between 17.2 – 17.6 Mb and fall on the transcribed region for a gene called partitioning defective 3 homolog (*pard3*). In humans, this gene plays a major role in epithelial cell tight junction formation (Chen et al., 2017), neuronal polarity (Khazaei and Püschel, 2009), and is involved in Schwann cell myelination (Beirowski et al., 2011). The next largest density of high loading SNPs is a group of 14 located on linkage group 15 between 7.972 Mb and 8.022 Mb. They fall on the transcribed region of the gene sodium bicarbonate transporter-like protein II (*slc4a11*).

A prominent feature of the first linear discriminant is the 2,115 SNPs distributed across the genome with a loading value of 1.375 e-05 (Figure 2.8, part B, blue arrow). Associated with this loading value are sites where Palmer Lake, Chatanika River, and Hootalinqua populations are all homozygous for the alternate allele, and the remainder of the individuals are homozygous for the reference allele.

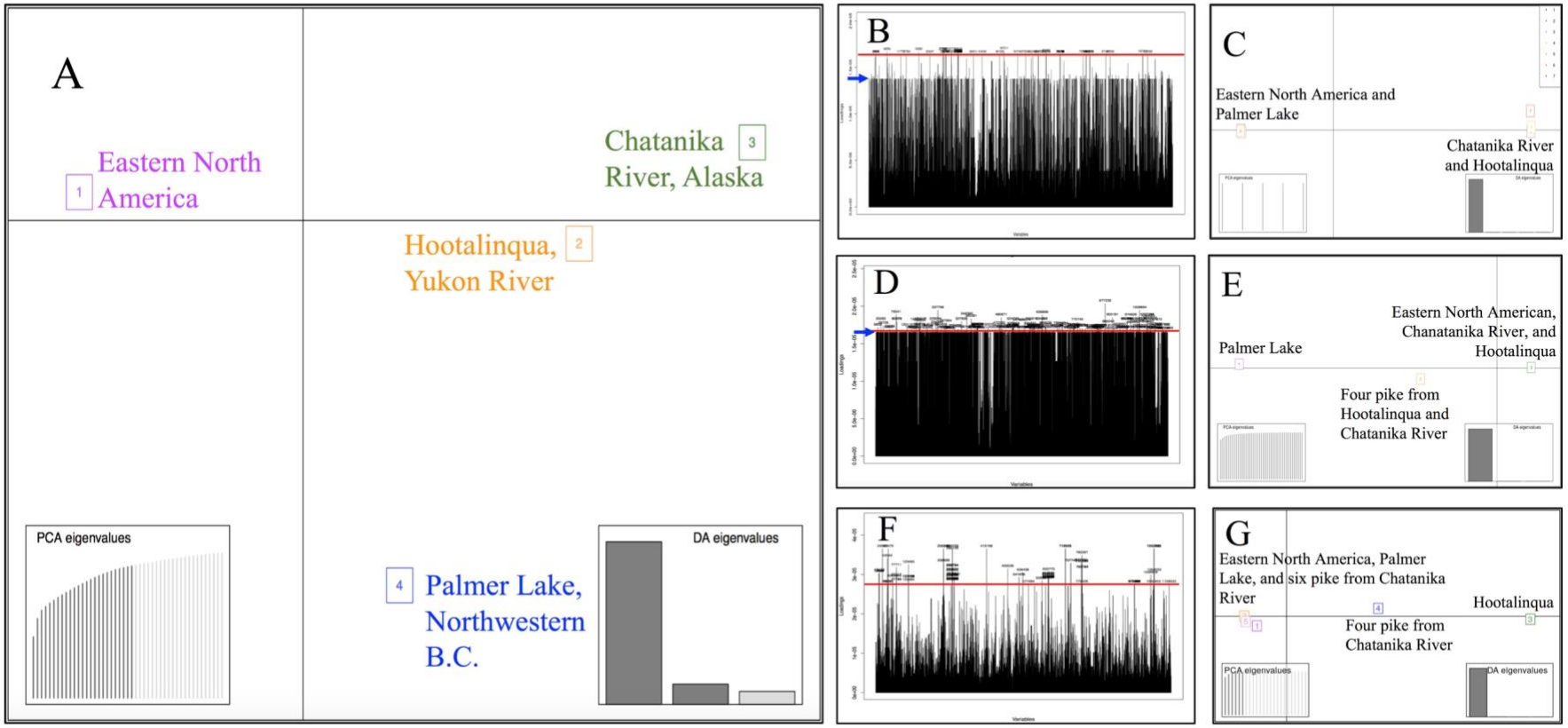



**Figure 9. Discriminant analysis of principle component scatter and loading plots.** (**A**) Is the main DAPC of where all SNPs are analyzed and. Cluster 1 contains all pike collected east of the Great Divide, cluster 2 contains only individuals from Hootalinqua, cluster 3 contains only individuals from Chatanika River, and cluster 4 contains only individuals from Palmer Lake. Individuals cannot be seen because they cluster so tightly they are hidden by group labels. **B**, **D**, and **F** are loading plots from each of the three retained discriminant axis (DAs) in **A** (DA1, DA2, and DA3, respectively). The blue arrow in plot **B** highlights genome wide SNPs that have gone to fixation at the alternate allele in Chatanika River, Hootalinqua, and Palmer Lake. The blue arrow in plot **D** indicates genome wide SNPs that have gone to fixation at the alternate allele in Palmer Lake only. The x-axis in these plots is the SNP ID number, and the y-axis is the loading value. Plot **C** is a DAPC scatter plot of high loading SNPs in **DA**1 above the red line in plot **B**. Plot E is a DAPC scatter plot of high loading SNPs in DA2 indicated above the red line in plot **D.** Plot **G** is a DAPC scatter plot of high loading SNPs in DA3 indicated above the red line in plot **F.**

The second linear discriminant is dominantly marked by genome-wide SNPs that load at 1.65e-05 (Figure 9, plot D, blue arrow). These SNPs are associated with sites where all four Palmer Lake individuals are homozygous alternate, and the rest of the individuals are all homozygous reference. There were 310 SNPs identified as having the highest contributions to the second discriminant axis (Figure 9, plot C, above the red line). Performing a secondary DAPC on these SNPs grouped Northern Pike into 3 groups: the first containing 39 of the 47 individuals, the second group containing one individual from Hootalinqua and three individuals from Chatanika River, and the third group containing the four Palmer Lake fish (Figure 9, plot E). The majority of these SNPs are located on linkage groups 9 (56 SNPs) and 11 (50 SNPs) and the remainder of the high loading SNPs are distributed fairly evenly across the genome. On linkage group 9, the high loading SNPs are located between 11.151 Mb and 11.216 Mb and fall in the transcribed region of the gene microtubule associated serine/threonine kinase 1 (*mast1*). On linkage group 11, the 50 high loading SNPs fall in the transcribed region of two genes: BTB/POZ domain containing protein 17 (*btbd17*) and dual specificity mitogen-activated protein kinase kinase 6 (*map2k6*).

The third linear discriminant in the DAPC analysis isolates Hootalinqua samples. High loading SNPs are associated with sites where Hootalinqua pike are all homozygous alternate or a combination of homozygous alternate and heterozygous. Some Chatanika River individuals are heterozygous at these locations, but Palmer Lake and Eastern North America are all homozygous reference. SNPZip identified 438 high loading SNPs for this axis. The majority (158) of these SNPs fall in a dense region on linkage group 6 between 6.334 Mb and 6.460 Mb where 3 genes are located: interferon induced very large GTPase I (*gvinp1*), BRCA2 & CDKNIA interacting

protein (*bccip*), and matrix-metalloproteinase 21 (*mmp21*). The remainder of densely arranged high loading SNPs fell outside of gene regions.

## Genotype frequencies and variant mapping

For each group identified by PCA analysis (and confirmed by DAPC), the genome wide mean variation in each category is listed in Table 5. In part A of Table 5, the means were calculated from variant sites only. In part B, all sites in the genome were included in the means. Since most of the sites in the genome are non-variant homozygous reference sites (i.e. variant frequencies of 0), there is a substantial difference in mean genotype frequencies depending on how they are calculated.

**Table 5. Genome-wide mean genotype frequencies.**

| Genotype | All Pike | Chatanika River | Hootalinqua | Palmer Lake | North America |
|---|---|---|---|---|---|
| **A** | | | | | |
| Heterozygous ($H_o$) | 0.08350 | 0.3228 | 0.3171 | 0.2849 | 0.0922 |
| Homozygous Alternate | 0.04875 | 0.2097 | 0.2163 | 0.5429 | 0.0319 |
| Variant | 0.1323 | 0.5325 | 0.5334 | 0.8277 | 0.1241 |
| Homozygous Reference | 0.08676 | 0.4675 | 0.4666 | 0.1723 | 0.8759 |
| Nucleotide Diversity | 0.1250 | 0.2906 | 0.3288 | 0.2135 | 0.0915 |
| **B** | | | | | |
| Heterozygous ($H_o$) | 1.001e-04 | 1.973e-04 | 1.385e-04 | 5.818e-05 | 6.437e-05 |
| Homozygous Alternate | 5.8443e-05 | 1.280e-04 | 9.303e-05 | 1.097e-04 | 2.279e-05 |
| Variant | 1.5854e-04 | 3.254e-04 | 2.315e-04 | 1.679e-04 | 8.717e-05 |
| Homozygous Reference | 9.998e-01 | 9.997e-01 | 9.998e-01 | 9.998e-01 | 9.999e-01 |
| Nucleotide Diversity | 1.498e-04 | 1.772e-04 | 1.434e-04 | 4.310e-05 | 6.383e-05 |

In part (A), means are calculated with variant sites only. In part (B) means are calculated with all site in the genome, where non-variant sites are assumed to have variant frequencies and nucleotide diversity values of 0 and homozygous reference frequencies of 1. Heterozygous frequencies are representative of $H_o$.

From the genotype frequencies, we can see immediately how incorporating all of non-variant sites into the mean dilutes discriminating population identifying signals. This is

indicative of the relatively few variant sites in Northern Pike, as we have seen in the individual genotype counts in prior sections. For example, in Chatanika River, the mean heterozygous genotype frequency among variant sites is 0.3228. Incorporating all sites into the mean reduces this to 0.0001973.

These genotype frequencies mirror the relationships we have observed between the Northern Pike populations from other analyses such as the individual genotype counts, the phylogenetic tree, and the PCA. Palmer Lake has the highest frequency of variants at any variant site, and the majority of these are homozygous alternate genotypes. Chatanika River pike have a higher $H_o$ than any other group, and the greatest variant frequency genome wide (Table 5 part B). Hootalinqua has the greatest mean nucleotide diversity at variant sites, and this could reflect the opportunity for this group to breed with other Northern Pike that inhabit the Teslin River and other tributaries of the Yukon River near Hootalinqua. Pike that inhabit Eastern North America have the lowest mean $H_o$ (Table 5, part A). However, on a genome-wide level, this group has a slightly greater $H_o$ than Palmer Lake (Table 5, part B).

Mapping the windowed mean of genotype frequencies across the genome is useful and interesting because it allows us to see the distribution of variation across the genome. By plotting the populations in different colours on the same plot, we can compare patterns of variation between populations. In Northern Pike, these plots show low levels of variation across the genome, with peaks that are indicative of dense regions of heterozygous or homozygous SNPs, regions where the SNPs have high genotype frequencies, or both (dense regions of SNPs where multiple individuals in the population have the SNP). For this section, linkage group 3 will be used as an example, but the trends discussed here hold true on a genome wide level.

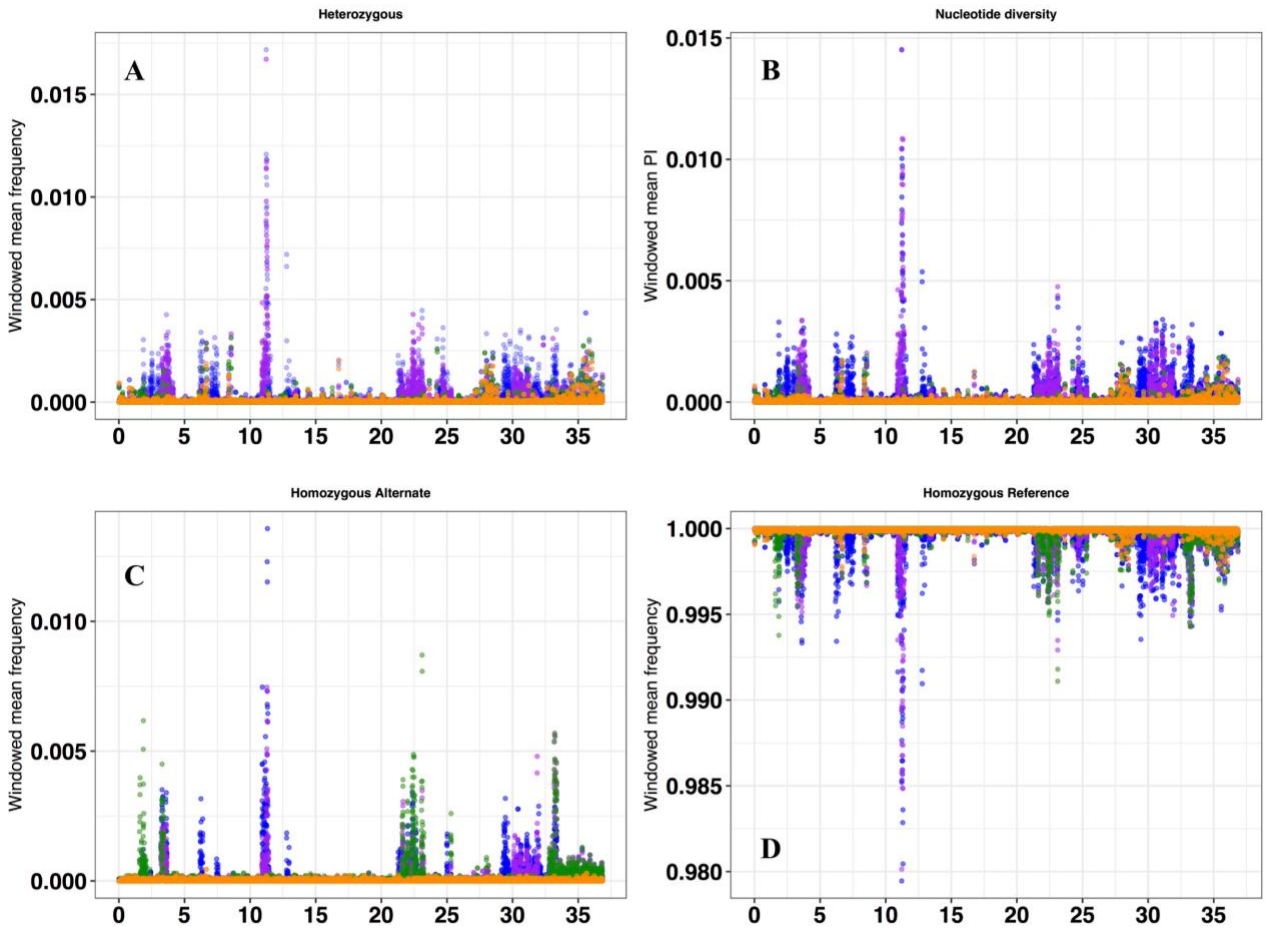



**Figure 10. Linkage group 3: windowed means of genotype frequencies.** Windowed means of (A) heterozygous genotype frequencies, (B) Nucleotide diversity, (C) Homozygous alternate genotype frequencies, and (D) Homozygous reference genotype frequencies, all calculated with a window size of 10Kbp and a step of 5Kbp. The x-axis on all plots is chromosomal position in MB. Populations are differentiated by colour: Blue –Chatanika River. Purple – Hootalinqua, Green – Palmer Lake, Orange – Eastern North America.

In linkage group 3, Alaskan and Hootalinqua (blue and purple in Figure 10, respectively) populations have more peaks, and peaks of higher amplitude, than Palmer Lake or North American populations in measures of heterozygosity and nucleotide diversity. Palmer lake has regions of homozygosity not seen in any other population (Fig 10 C, green). Eastern North America (Figure 10, orange) appears as a flatline across much of the linkage group for all parameters. These observations are reflective of the means observed in Table 5. On linkage group 3 between 20-25MB, we see that the Chatanika River and Hootalinqua have an increase in heterozygous genotype frequencies (Fig 10, part A). At the same position, Palmer Lake has an

increase in homozygous alternate frequencies (Fig 10, part C). This circumstance is observed genome wide in other linkage groups, as well as the trend discussed above. This indicates that Hootalinqua and Chatanika river populations represent a genetic middle ground between Palmer Lake and eastern North American populations.

## Hardy-Weinberg equilibrium

Of the 1,117,361 chromosomal variant sites tested across all populations, 4.3% (48,436) were significantly out of Hardy-Weinberg equilibrium ($p < 5 \times 10^{-8}$). These SNPs were located across the genome and also clustered in peaks associated with regions of interest, such as the sex-determination region on linkage group 24 and the massive and complex repeat region found on linkage group 11 at around 30 – 40 MB (Figure 11). Genome-wide significant SNPs are associated with sites that have gone to fixation at alternate alleles in different populations. Looking at HWE within separate populations, we see hardly any significant results. This could be indicative of the true nature of these population, but could also be due to the lower power associated with testing few numbers of individuals. As a continental population, North American Northern Pike defy HWE genome wide, indicating that at least one of the HWE assumptions has been violated.

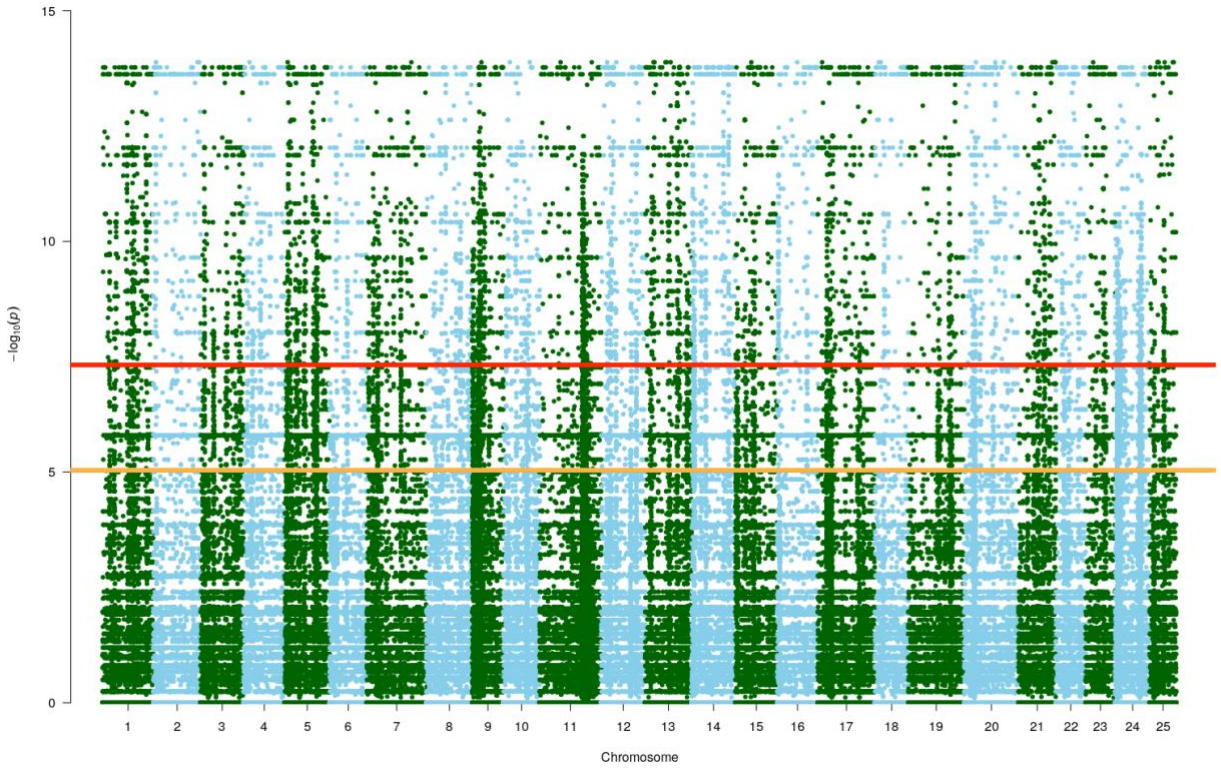



**Figure 11. Manhattan plot of HWE Chi-square test p-values.** All Northern Pike tested as a single population. Bi-allelic SNPs only. Chromosomal SNPs included only (no unplaced scaffolds or mitochondria). Red line signifies the Bonferroni corrected significance level ($p = 5x10^{-8}$). Orange line signifies the approximate chromosomal-significance level ($p=1x10^{-6}$).

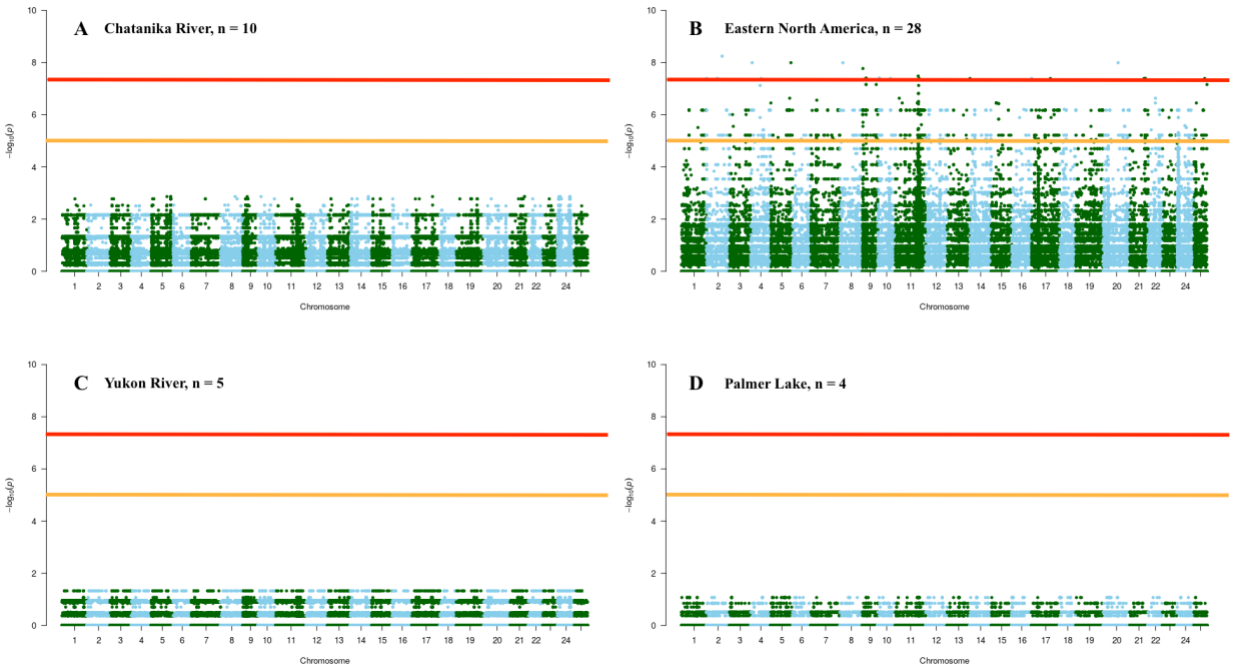



**Figure 12. Manhattan plots for p-values of HWE test within groups.** (A) Chatanika River, (B) Eastern North America, (C) Yukon River at Hootalinqua, (D) and Palmer Lake. (n) indicates the number of pike in each group.

## Tajima's D

Average genome-wide values of Tajima's D were negative for all Northern Pike when analyzed together and as separate populations. Eastern North American Northern Pike had the lowest genome-wide Tajima's D and the smallest standard deviation. Northern Pike hold less genetic variation than expected among and within most populations.

**Table 6. Genome wide summary statistics of Tajima's D.**

| Group | Min. | 1st Quartile | Median | Mean ± SD | 3rd Quartile | Max. | N |
|---|---|---|---|---|---|---|---|
| All Pike | -2.965 | -2.001 | -1.75 | -1.44 ± 0.99 | -1.329 | 5.086 | 47 |
| CR | -2.682 | -1.164 | -0.592 | -0.228 ± 1.25 | 0.551 | 3.535 | 10 |
| YR | -2.17 | -1.11 | -0.69 | -0.34 ± 1.14 | 0.41 | 2.8 | 5 |
| PL | -1.86 | -1.05 | -0.23 | -0.09 ± 1.13 | 1.1 | 2.62 | 4 |
| ENA | -2.981 | -1.952 | -1.696 | -1.609 ± 0.59 | -1.452 | 4.59 | 28 |

Pike were analyzed all together and as separate groups as defined by PCA and DAPC. Min. = minimum. SD = standard deviation. Max. = maximum. N = number of individuals in group. CR = Chatanika River, Alaska. YR = Yukon River at Hootalinqua, Yukon. PL = Palmer Lake, BC. ENA = Eastern North America (including Castlegar, Charlie Lake, Manitoba, New York and New Jersey).

When all Northern Pike are observed together, Tajima's D is negative across much of the genome (Figure 13). An obvious feature of Figure 13 are the peaks of Tajima's D that appear on almost every linkage group genome wide. Some of these peaks overlap with regions of high-loading SNPs identified by DAPC. On a genome-wide level, Tajima's D indicates that forces beyond mutation and drift, such as selection or demography, are influencing the evolution of North American Northern Pike. Additionally, some of these regions are important in defining population structure.

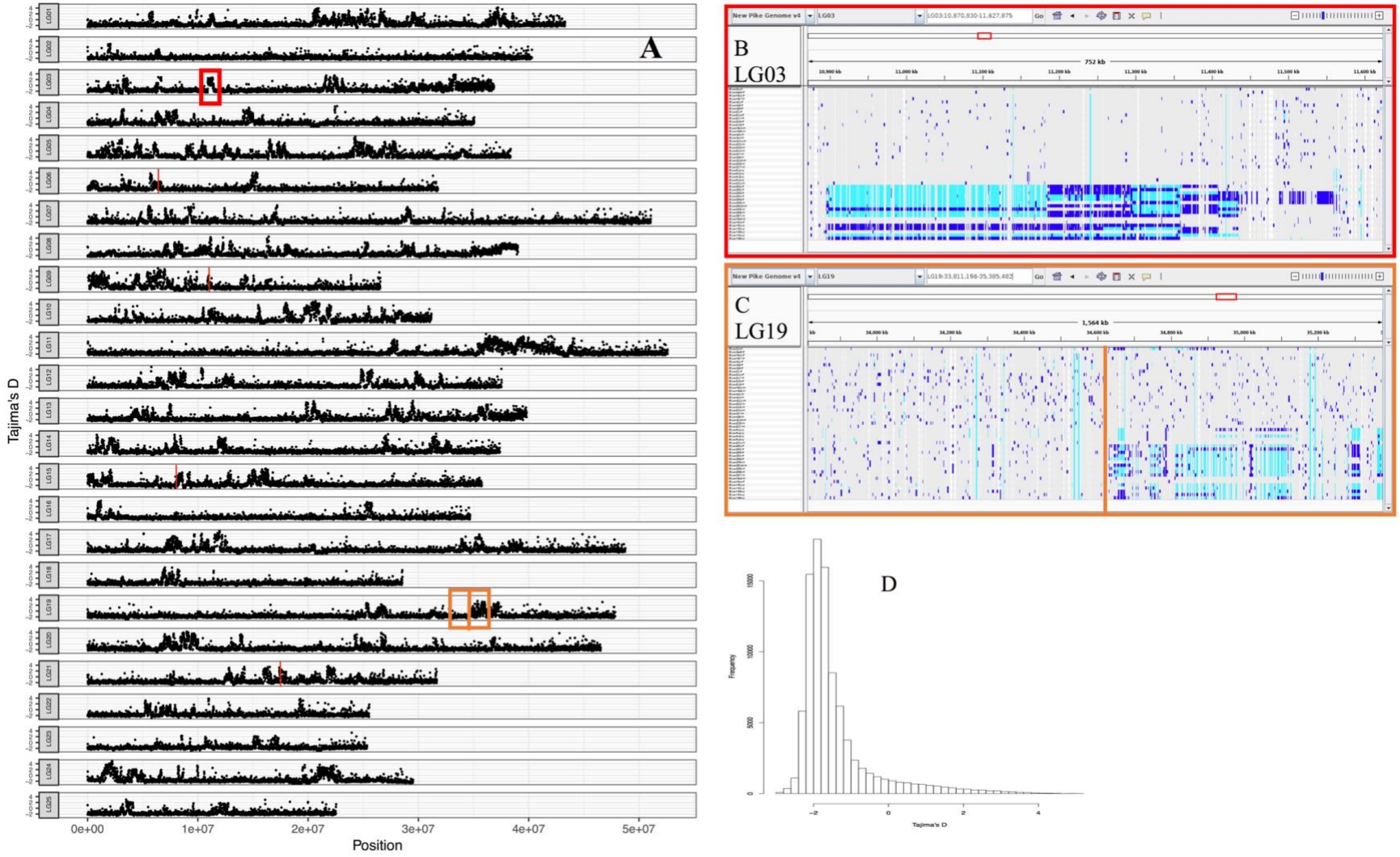



**Figure 13. Tajima' s D by linkage group.** (A). Red lines on LGs 06, 09, 15, and 21 show where high loading SNPs from DAPC analysis overlap with increased values of Tajima' s D. The red inset shows LG03 between 10.8 and 11.6 MB in Integrated Genomic Viewer, where population structured genotypes relate to an increase in Tajima' s D. The orange inset shows LG19 between 33.8 and 35.4 MB, another region of increased Tajima' s D associated with population structure. In both IGV images, dark blue represents heterozygous genotypes and light blue represents homozygous alternate genotypes. Histogram of Tajima' s D values in (B). The positively skewed topology is indicative of population expansion.

## Wright's fixation index

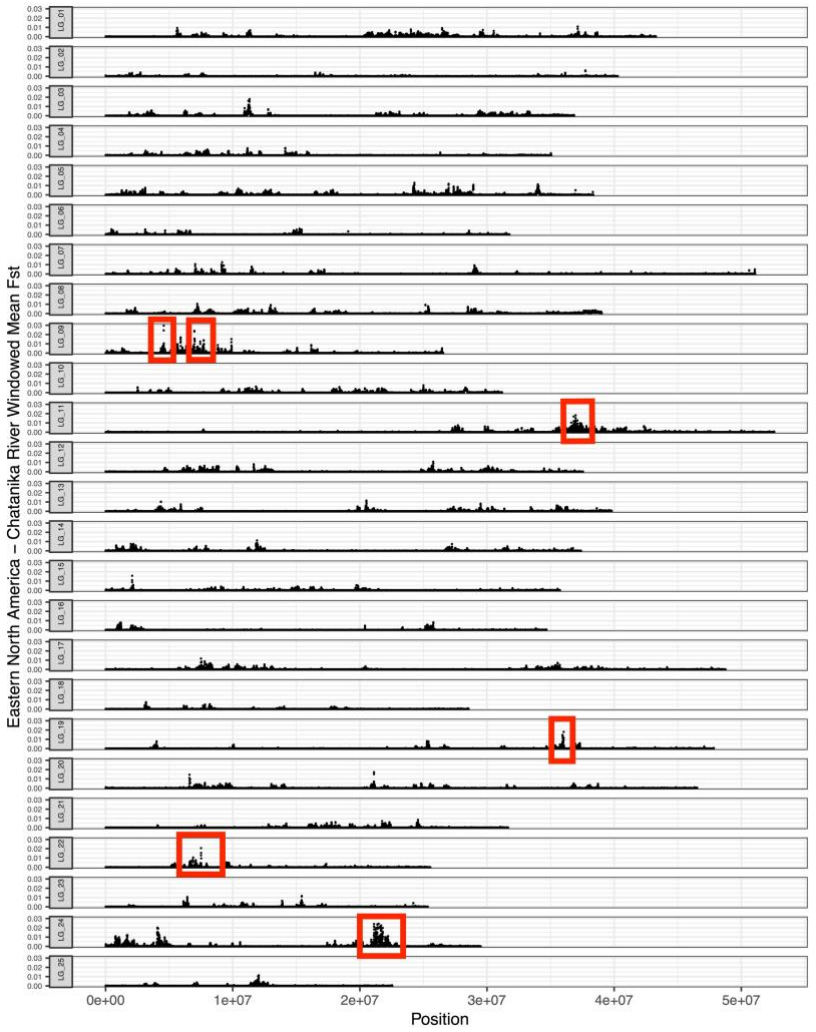
As calculated by VCFTools using the Weir and Cockerham method (Weir and Cockerham, 1984), mean $F_{ST}$s are the greatest between Palmer Lake, B.C., and Chatanika River, Alaska, closely followed by Palmer Lake and Yukon River, and North America and Alaska (Table 7). North American and Palmer Lake pike are about as similar to each other as Chatanika River is to the Yukon River pike, based on these values.

**Table 7. Mean $F_{ST}$ values between groups.**

| Comparison | Mean $F_{ST}$ |
|---|---|
| PL vs CR | 0.22 |
| PL vs H | 0.20 |
| NA vs CR | 0.20 |
| NA vs H | 0.17 |
| A vs H | 0.13 |
| ENA vs PL | 0.12 |

$F_{ST}$ as calculated by the Weir and Cockerham method (1984). Calculations include only variant sites and not all sites in the genome. PL – Palmer Lake, B.C.; CR – Chatanika River, Alaska; H – Yukon River at Hootalinqua, Yukon Territory; ENA – Eastern North American populations (Charlie Lake, Castlegar, Manitoba, New York, New Jersey)

In order to identify discrete genomic locations where populations differed from each other the most, we calculated the mean $F_{ST}$ of all sites across 10,000 base pair windows. We then plotted these means every 5,000 bases. For all linkage groups, the topography of these plots is a flatline around 0 with discrete peaks where populations are stratified. Peaks on $F_{ST}$ plots are consistent with the locations of peaks on Tajima's D plots, supporting the accepted concept that increased values of Tajima's D are associated with population structure. To illustrate, see the windowed-mean plot of $F_{ST}$ for Eastern North America vs Chatanika River, Alaska (Figure 14). Only the results for North America vs Chatanika River are presented as these are the groups with the largest number of individuals as therefore provide the most robust comparisons.




| Linkage Group | Region | Genes in Region | Variants in H | Variants in PL | Functional Themes |
|---|---|---|---|---|---|
| LG24 | 21.2 – 22.0 MB | Y | Y | N | Highly repetitive region/Cell adhesion and basement membrane proteins |
| LG09 | 4.5 MB | Y | Y | N | Unknown – genes are uncharacterized |
| LG09 | 6.9 -7.0 MB | Y | Y | Y | Major Histocompatibility Complex |
| LG22 | 7.5 MB | Y | N | N | Signal transduction and immunity |
| LG19 | 36.0 – 36.1 MB | Y | Y | Y | Immunity and pathogenesis |
| LG11 | 36.5 – 37.0 MB | Y | Y | N | Innate immunity |

**Figure 14. Chatanika River – Eastern North America Fst comparison by linkage group.** The top 6 regions of elevated Fst are highlighted in red and described in the accompanying table. Y = Yes. N = No. H = Hootalinqua. PL = Palmer Lake.

The top six regions of the genome that have elevated $F_{ST}$ are highlighted and the majority of them have functions in immunity. Interestingly, one of the greatest areas of differentiation between Eastern North America and Chatanika River has only uncharacterized genes. These results show that populations east and west of the Continental Divide are mainly differentiated by variation in regions of immunity, and by a region of unknown function.

# Discussion

## Overview

Our results confirm very low levels of observed heterozygosity ($H_o$) in Northern Pike, with the greatest population-level $H_o$ in Alaska (0.32) and the smallest in Eastern North American Pike (0.092). For comparison, average $H_o$ in other freshwater fish is reported to be 0.46 (DeWoody and Avise, 2000), and $H_o$ in Atlantic Salmon has been reported to range from 0.49-0.92 in some eastern United States populations (Spidle et al., 2004). In Northern Pike of North America, an average of one heterozygous SNP every 10,000 bases is observed. In other organisms, an average of one polymorphic SNP every 1000 bases is observed in humans (the 1000 Genomes Project Consortium, 2015), one every 750 bases in salmonids (Koop, 2018), one every 500 bases in Atlantic Cod (Star et al., 2011), and one every 300 bases in herring (Martinez Barrio et al., 2016).Variation in Northern Pike is concentrated in regions of high SNP density and is largely in regions of the genome where immune genes reside. Clustering patterns and phylogenetic analysis suggest that populations east and west of the Continental Divide are isolated from each other, and multiple isolated populations exist west of the Continental Divide. These observations, along with that of SNPs out of HWE across the genome and low Tajima's D values in eastern North America suggest that Pike colonized North America from the Beringia refugium, and are still expanding east of the Rockies.

## Variation

As previous studies indicated, our results confirm a remarkably low level of genetic variation in Northern Pike, averaging about one heterozygous SNP every 10,000 bases. The Castlegar individual had the fewest number of heterozygous SNPs of all Northern Pike analyzed, however, because this individual was sequenced with a different technology than the rest of the

samples, it is possible but unlikely that this low heterozygous count is indicative of a founder effect or is a by-product of lower mapping qualities attributed to the alternate sequencing technology.

In the Northern Pike genome, SNPs are not distributed evenly but rather concentrated in regions of high variability that are associated with multiple copy number variants and immune function (Figure 6 & Figure 10). Observed heterozygosity agrees with ranges summarized in Table 1 (Introduction). Compared to heterozygosity values reported by (DeWoody and Avise, 2000), heterozygosity in Northern Pike falls below their average of 0.46 reported for freshwater fish. When all Northern Pike across North America are observed together, observed mean $H_e$ is substantially lower at 0.08350. Interestingly, this is lower than all mean $H_o$s calculated within populations and is likely due to the different sample sizes among and within populations. Regardless of the way it is measured, observed heterozygosity for Northern Pike is less than what is seen in other freshwater fish, and is on the far low end of reported $H_o$s for Northern Pike, especially in Eastern North America.

Northern Pike genotype frequencies mirror the relationships observed between populations from other analyses such as the individual genotype counts, the phylogenetic tree, PCA and DAPC. Palmer Lake has the highest frequency of variants at any variant site, and the majority of these are homozygous alternate genotypes. This suggests that Palmer Lake was founded by few individuals and that the population is isolated. Alaskan Pike have the highest observed $H_e$ of any group (0.32), and the highest variant frequency genome wide (3.25e-04; Table 5, part B). Older populations generally have greater heterozygosity than newer populations as they have had time to accumulate mutations and spread them through the population. It follows then that Chatanika River is the oldest population in North America, as it has the greatest

number of heterozygous sites genome wide and the greatest heterozygosity (summarized in Table 8). The Hootalinqua population has the greatest mean nucleotide diversity at variant sites, and this could reflect the opportunity for this group to breed with Northern Pike inhabiting the Teslin River and other tributaries of the Yukon River near Hootalinqua as well as the Chatanika pike through the Yukon River. Northern Pike that inhabit the remainder of North America have low $H_o$ (summarized in Table 8). This is indicative of relatively young populations that have not had the time to accumulate variation. However, on a genome-wide level, this group has a slightly greater mean heterozygous frequency per site than Palmer Lake (Table 5, Part B). This could be because the 28 individuals that comprise the Eastern North American group are spread over a vast geographical area, whereas only four pike comprise our Palmer Lake group, and they are isolated to a very small lake. As a result, Eastern North American fish likely contribute more sites with heterozygous genotypes, thus incorporating fewer values of zero in the genome-wide mean than Palmer Lake. The low frequency of homozygous alternate genotypes in Eastern North America reflects that this group is more closely related to the individual used to build the reference genome than any other group, and supports the positioning of the Castlegar Pike in the PCA, phylogeny, and DAPC.

**Table 8. Summary of variation statistics.**

| Group | Mean No. of Heterozygous Loci per Fish | Observed Heterozygosity | Nucleotide Diversity | Tajima's D | N |
|---|---|---|---|---|---|
| All Pike | 93,429 | 0.08350 | 0.1250 | -1.44 | 47 |
| Chatanika River | 181,268 | 0.3228 | 0.2906 | -0.228 | 10 |
| Hootalinqua | 127,012 | 0.3171 | 0.3288 | -0.34 | 5 |
| Palmer Lake | 58,337 | 0.2849 | 0.2135 | -0.09 | 4 |
| Eastern North America | 61,073 | 0.0922 | 0.0915 | -1.609 | 28 |

The greatest densities of SNPs are located in regions that contain mostly immune related genes, such as immunoglobins, tripartite motif – containing genes, and MHC genes present in multiple copies. This observation suggests that maintaining the ability for immunological adaptation is of high importance in Northern Pike, and suggests that the location of the variation, and not the overall amount, may be more important for success.

Negative values of Tajima's D result from an excess of low frequency alleles and are classically interpreted as evidence for a strong bottleneck and/or population expansion, or positive selection. In a bottleneck, the number of individuals in a population is greatly reduced. As such, the pool of genetic variation is fragmented to the subset held within the remaining individuals. The few remaining individuals contain just a portion of the variation that was present before the bottleneck. An allele that was present at a frequency of 0.5 may now be present at a frequency of 0.1 or 0.9, for example, or may become fixed as smaller effective population sizes tend to fix alleles in fewer generations than larger effective population sizes. Expansion further exacerbates the perpetuation of low frequency alleles. Expansion can be thought of as a number of small founding events, which have a similar effect to a bottleneck in that a subset of individuals (and therefore, the gene pool) from the original population are the ones to colonize and populate a new environment. Positive values of Tajima's D are indicative of an excess of intermediate allele frequencies due to population contraction, structure, or balancing selection/heterozygote advantage (Biswas and Akey, 2006; Ramírez-Soriano et al., 2008; Simonsen et al., 1995; Tajima, 1989). From our summary statistics, we see a mean negative value of -1.44 ±0.99 when all Northern Pike are analyzed together, and from Figure 13 and Table 8 we see that the majority of sites have a Tajima's D values of -1.75 or less. These observations

support the notion of a strong population bottleneck or small founding population in Eastern North America, followed by population expansion.

As the field of biology has developed, our understanding of the role of genetic variation has developed as well. Genetic variation is the precursor for differentiation, adaptation, speciation, and is responsible for biodiversity. Generally, it has been accepted that more genetic variation (in terms of the number of polymorphic alleles genome wide, and the percent of the population heterozygous at the polymorphic site) is correlated with the ability of a species to adapt to changing environments. Likewise, it has been demonstrated that a lack of genetic variation is a contributing factor to extinction (Amos and Balmford, 2001; Frankham, 2005). However, as sequencing technology has advanced, as well as our ability to interpret and compare sequence data, we see multiple examples of successful and thriving species with extremely low levels of genetic variation and relatively homogenous genomes (Abadía-Cardoso et al., 2017; Merola, 1994; Milot et al., 2007; Robinson et al., 2016). From this study we can see that it is possible for a species to colonize an immense geographic range despite having low levels of genetic variation. In addition to cheetahs (*Acinonyx jubatus*), albatrosses (*Diomedea exuland* and *Diomedea amsterdamensis*), northern elephant seals (*Mirounga angustirostris*), channel island foxes (*Urocyon littoralis*), and likely others, Northern Pike are an example of a species surviving despite an overall lack of genetic variation.

This begs the question of the role of genetic variation in survival. From our study, we see that regions of concentrated SNPs and heterozygosity affect the major histocompatibility complex and other immune related genes. Northern Pike are apex generalist predators who prey on other fish (including smaller pike), insects, amphibians, small birds, and in turn ingest the pathogens their prey is host to. Thus, the ability of Northern Pike to protect itself against a

repertoire of pathogens may be essential to its survival and success. The MHC is known to be highly polymorphic in vertebrates. Preventing infection is a common challenge for animals as new pathogens are commonly encountered, and heterozygosity in the MHC provides a selective advantage (Penn et al., 2002). Survival of the critically endangered Attwater's prairie-chicken has been linked to heterozygosity in the MHC and other immune related genes and not to genome-wide heterozygosity (Bateson et al., 2016). Aside from the MHC, the existence and location of excess polymorphism is different between species yet seems to be related to aspects of their specific survival strategies. In stickleback, spines and bony plates have demonstrated importance for the ability of the fish to mitigate predation (Marshall and Wund, 2017; Reimchen, 1994). Standing genetic variation has been shown to be elevated in the genomic region that controls the number and morphology of bony plates (Nelson et al., 2019). In the endangered channel island fox, whose genome has been described as a monomorphic flatline, peaks of heterozygosity coincide with olfactory receptor genes (Robinson et al., 2016). Foxes are known to have a keen sense of smell that helps them identify each other and detect food concealed under rocks, ground, and snow. Regions of the human genome with tandem gene copy number variants are reportedly enriched with genes functioning in cell adhesion, sensory perception, and neurophysiology (Redon et al., 2006). Cell adhesion molecules play a major role in cell to cell communication and therefore as a communication channel between stimuli and our nervous response (Li et al., 2009a). As humans, our ability to perceive, make connections, and manipulate external stimuli is fundamental to our survival and success.

These observations suggest a pattern: regions of high variation in the genome are associated with phenotypic variation whose sensitivity may be instrumental to the survival of the species. This may mean that the maintenance of genetic variation is most crucial in gene regions

that affect a species' ability to detect/respond to stimuli that are both fundamental to survival and in flux in their habitat.

## Population genomics

Major trends in our phylogeny, PCA, and DAPC cluster Northern Pike into two major groups separated by the North American Continental Divide. In our phylogeny, the longest branch separates these two major nodes. Populations east of the Continental Divide are remarkably similar to one another and group into one tight cluster. The one exception to this is the sample from Castlegar, BC. This pike groups with those from eastern North America, despite being collected west of the Continental Divide. This is explained by the fact that Northern Pike are invasive in Castlegar and their origins have been traced to east of the Continental Divide in Montana (Mehaffey, K.C., 2018; Vashro, Jim, 2018), where they were transported west of the divide by humans.

The first discriminant axis of our DAPC accounts for the majority of variation in the data and attributes it to SNPs that differentiate Northern Pike from east and west of the Continental Divide. Genotypes at these high loading sites have gone to fixation at alternate alleles in Northern Pike on opposite sides of this barrier. The lack of heterozygosity at these sites could suggest that we have not sampled to saturation to find an individual with a heterozygous genotype, but because these sites are located on every linkage group across the genome, it more likely suggests that populations from Northwest are not breeding with populations from the east. We believe that genetic drift led to alternate sites being fixed in these two groups over time. Our $F_{ST}$ comparisons between Chatanika River and Eastern North America show that large regions associated with the major histocompatibility complex (MHC) and other immune related processes are distinguishing features of these two groups. These analyses highlight the

geographic separation between populations east and west of the Rockies. Further evidence that distinguishes Northwestern populations from eastern North American populations is the prevalence of SNPs out of HWE across the genome when all Northern Pike are analyzed together. Observations similar to these have been used to distinguish new species, such as the identification of a third orangutan species (Nater et al., 2017).

From the first discriminant axis in the DAPC, where most of the variation is accounted for, the 309 highest loading SNPs group Chatanika River and Hootalinqua pike together, and group Palmer Lake with the rest of North America. This is a different clustering pattern than what is seen with genome-wide SNPs that isolate Eastern North America from populations west of the Rocky Mountains described above. Theses SNPs are not evenly spread across the genome but rather occur densely in discrete locations. Of the 309 high loading SNPs, 249 are located on linkage group 21 between 17.2 – 17.6 Mb and fall on the transcribed region for a gene called partitioning defective 3 homolog (*pard3*). It is unclear if this clustering patter signifies a region under selective pressure or is the result of drift. Either way, it is notable that such a small region differentiates two distinct regions – the Yukon Drainage Basin from the rest of North America. The gene *pard3* has been repeatedly shown to have an influence on cell polarity in metazoans (Brajenovic et al., 2004; Joberty et al., 2000; Khazaei and Püschel, 2009). In humans, this gene plays a major role in epithelial cell tight junction formation (Chen et al., 2017), neuronal polarity (Khazaei and Püschel, 2009), and is involved in Schwann cell myelination (Beirowski et al., 2011). The next largest density of high loading SNPs is a group of 14 located on linkage group 15 between 7.972 Mb and 8.022 Mb. They fall on the transcribed region of the gene sodium bicarbonate transporter-like protein II (*slc4a11*). This gene is a unique member of the *slc* family of genes. If borate is present, it is an electrogenic $Na^+$-coupled borate transporter, and has been

shown to affect cell growth and proliferation through the MAPK pathway, and is also pivotal for boron homeostasis (Park et al., 2004). Taken together, Northern Pike from the Yukon River drainage basin can be distinguished from the rest of North American populations functions using variation that affects two small regions of the genome. This region may have selective influence in cell polarity and growth.

Other clustering patterns separate Palmer Lake and Hootalinqua into their own distinct groups. In the phylogeny, the isolation of palmer lake individuals is supported by a bootstrap value of 100 and in the DAPC it is demonstrated in the second discriminant axis. High loading sites segregating Palmer Lake are located on every linkage group across the genome and have a homozygous alternate genotype. We believe these are sites that have gone to fixation at the alternate alleles due to random genetic drift and physical separation over time. As there are no heterozygotes at these sites, it seems to follow that Palmer Lake is an isolated population. The majority of high loading SNPs segregating Palmer Lake are located on linkage groups 9 (56 SNPs) and 11 (50 SNPs) and the remainder of the high loading SNPs are distributed fairly evenly across the genome. On linkage group 9, the high loading SNPs are located between 11.151 Mb and 11.216 Mb and fall in the transcribed region of the gene microtubule associated serine/threonine kinase 1 (*mast1*). It codes for an enzyme that links dystrophin/utrophin network with the microtubule network (Lumeng et al., 1999) and is a physical link in neuromuscular junctions and neuronal postsynaptic densities (Lumeng et al., 1999). On linkage group 11, the 50 high loading SNPs fall in the transcribed region of a gene called BTB/POZ domain containing protein 17 (*btbd17*). The BTB/POZ domain is a protein-protein interaction domain that is evolutionarily conserved, and can have functions in regulating subcellular location, DNA binding, and can impact gene expression by selective binding of co-factors (Collins et al., 2001).

According to Gene Ontology Resource (GO), this gene may have functions in transcriptional responses to viruses, and may negatively regulate viral genome replication.

The final cluster is that of Hootalinqua. Drawing from the position of the Hootalinqua pike in the phylogeny, and their central position in both the PCA and DAPC, we believe this population bridges the Alaskan population to the remainder of the populations in North America. The Yukon River is a large drainage system that collects from many tributaries. The Teslin River joins the Yukon River just a kilometer downstream from Hootalinqua, and opens up the potential for different strains of Northern Pike from each river to cohabitate and interbreed. The occurrence of different strains of Northern Pike at this location could explain the varying levels of heterozygosity seen in the genotype count plot, the topography of Hootalinqua pike in the phylogenetic tree, the relatively loose clustering seen in the PCA cluster, and this population having the highest nucleotide diversity despite having only five individuals. Hootalinqua is segregated by the third discriminant axis of the DAPC, which places them in a genetically central position. The majority of high loading SNPs that isolate Hootalinqua occur in dense clusters in transcribed regions of the genome that have functions in the nervous and muscular systems, are involved in responses to viral infections, DNA damage repair, and embryonic development. Because these high loading SNPs are located in discrete regions of the genome, they may highlight variants that offer a selective advantage. The majority (158) of these SNPs fall in a dense region on linkage group 6 between 6.334 Mb and 6.460 Mb where 3 genes are located: *gvinp1*, *bccip*, and *mmp21*. These genes may have roles in antiviral activity (Haller and Kochs, 2002), DNA damage repair and cell cycle regulation (Su et al., 2016), and tissue development in embryogenesis (Marchenko et al., 2003), respectively. The remainder of densely arranged high loading SNPs fell outside of gene regions.

## Phylogeographical implications

Our results support the assertion that North America was originally colonized by Northern Pike descended from the Yukon River drainage basin, and by extension, from a Beringia refugium. This agrees with d-loop and cytochrome C analysis by Skog et al., (2014). Evidence supporting this comes from the descending number of heterozygous SNPs per individual from Chatanika River to Eastern North America, lower population wide heterozygosity, as well as a genome-wide depression of Tajima's D in eastern North America. Such observations indicate a bottleneck event and/or dispersal initiated by a small number of founders (Dlugosch and Parker, 2008). This is corroborated by an increase in the occurrence of unique SNPs and a decrease in minor allele frequencies across eastern North America. What remains unclear is the number of founding populations that re-colonized eastern North America after the ice age. A similar pattern of colonization has been suggested for the pygmy whitefish (D. S. Witt et al., 2011).

## Limitations and considerations

The low number of Northern Pike in each population was a limiting factor for this study. Our Palmer Lake and Hootalinqua populations consisted of just 4 and 5 fish, respectively. This can make analyses less robust, and we likely have not reached a sampling saturation. Nonetheless, because of the very high resolution of the sequencing method, major trends such as number and location of nucleotide variants is reliable and telling. Another consideration is the geographical origins of our samples. Pike have been extensively introduced to new habitats and documentation of introductions is sparse and incomplete (Harvey, 2009). Although most introductions are extensions of a native range, some have long distance introductions have occurred (Harvey, 2009) and as such, there is the possibility of introductions confusing trends, as

seen with our Castlegar Pike. Two of our populations (New Jersey and Manitoba) were sampled from fish hatcheries. As such, we considered the possibility that a limited number of parents could be responsible for the majority of the genetics at these locations. However, these locations show no less variation than the wild New York population, and clustering analysis is supportive of geographical population stratification. Additional samples from across the Canadian Arctic and United States would be valuable for extending these results.

## Chapter 2 conclusions

The overall heterogeneity of the species is likely explained by its tendency to have a small effective population size and its role as a freshwater generalist apex predator and cannibal, capable of colonizing new habitats with just a few fish. All of these factors are associated with reduced variation (DeWoody and Avise, 2000; Gillespie, 2004; Smith and Fujio, 1982). Despite low overall levels of genetic variation in Northern Pike, the variation that does occur is associated with features known to contribute to evolution and adaptation, such as regions with multiple copy number variants, duplications, elevated standing genetic variation, and immune related functions (Barrett and Schluter, 2008; Dennis and Eichler, 2016).

Our result suggest that Northern Pike in North America form two major groups physically separated by the Rocky Mountains: Northwestern and Eastern North America. Within the northwest, Palmer Lake pike have low levels of heterozygosity and a high number of uniquely fixed alleles, indicating this population is physically isolated from the others and following its own evolutionary trajectory. West of the great divide, Northern Pike in major river systems (Chatanika, Yukon) harbour more variation and greater nucleotide diversity within just 5 – 10 individuals than all Northern Pike east of the Continental Divide from Charlie Lake, British Columbia, to Hackettstown, New Jersey.

Based on patterns of variation, we conclude that concentrated regions of genetic variation which influence immune function are important to the success of Northern Pike, and allow them to flourish despite an overall lack of genetic variation genome-wide. The majority of variation that occurs in Northern Pike is limited to dense regions with functions in immunity. Genomic regions that separate populations have functions related to the nervous system and development. We support propositions that Northern Pike originally colonized North America from a Beringia refugium, and that a very small founding population is responsible for colonization east of the Continental Divide. However, the pattern of re-colonization after the last ice age is still unclear, as small populations surely persisted south of the ice sheets and played a role in recolonization.

# Chapter 3

# Sex Determination in North American Northern Pike

## Summary

Northern Pike are an abundant, wide spread, economically and ecologically valuable species across the Northern Hemisphere. Phylogenetically, they represent the sister group (order: Esociformes) to the heavily studied salmonids (order: Salmoniformes) and serve as a valuable outgroup in salmonid research. In late 2017, a male-specific sex determination gene, *amhby*, was characterised in European strains of Northern Pike (Pan, 2017; Pan et al., 2019). This gene was located on the telomeric end of LG24 and accompanied by flanking regions of linked male-specific heterozygosity. However, *amhby* and linked male specific heterozygosity was not detected in North American populations by RAD-Seq analysis (Pan, 2017). Skewed sex ratios observed in laboratory experiments and natural populations indicate that environmental factors may play a role in sex determination in eastern North American Northern Pike. This high-resolution investigation into sex determination in Northern Pike of North America shows the presence of *amhby* in three North American populations, but it appears to have been lost east of the North American Continental Divide. Analysis revealed that not even a single heterozygous locus was shared among all males or all females in our Northern Pike from across North America (i.e. on a continental level, there are zero sex-specific alleles). This means there are at least two different sex determination mechanisms in Northern Pike. Finally, a comparison is offered between the genetic sex determination systems of Northern Pike and Salmonids.

Detection of 3,555 male specific heterozygous SNPs concentrated in a 500kb region of LG24 indicates a strong GSD signal in Alaskan Pike. All females are homozygous reference in

this region. Male-specific false-mapping of sequence reads to the autosomal copy of *amh* on LG08 (as evidenced by elevated read depth over the exons of the gene) indicates a duplicated copy of *amh* present only in males of this population. An almost identical genetic signal appears in one of our Hootalinqua pike, indicating that this population has the same sex-associated phenotype. This male specific heterozygosity on LG24 and the evidence for a duplicated copy of *amh* is absent from the remainder of our re-sequenced Northern Pike. PCR analysis using a primer set designed from our sequences, as well as *amhby*-specific primers from Pan et al., (2017) confirm the presence of male specific heterozygosity and the presence of *amhby* in Chatanika River and Hootalinqua males, and confirms the loss of *amhby* from all males and females east of the Continental Divide. PCR analysis on additional samples corroborated these results and interestingly, confirmed the presence of these male-specific markers as well as a duplicated copy of *amhby* in the invasive population in Castlegar, B.C.

One-hundred and ninety-two female-specific heterozygous SNPs spanning 350kb on LG10 were detected in our New Jersey population, and this was the only other notable sex-specific signal observed in North American Northern Pike east of the Continental Divide. Genes in this region are immune and estrogen-related, and involved in the TGF-β signalling pathway. We propose that a younger, genetic sex determination system was gained in this population.

# Introduction

Primary sex determination in fish shows much plasticity from ESD to GSD, to a combination of the two. Genes involved in GSD are different among many of the species studied so far, however most involve the TGF-β signalling pathway. Recent work on Northern Pike in Europe shows a duplicated copy of the anti-Müllerian hormone gene (*amh*), called *amhby*, is present only in males and is responsible for male sex development (Pan et al., 2019) in an XX (female) – XY (male) system. However, the additional copy of *amhby* was not detected in North American Northern Pike using RAD-Seq (Pan, 2017).

Genes that control sex determination can lose their function to another gene or stimulus. This is especially common in ectotherms. A master sex determining (MSD) gene is one that initiates the cascade of cellular events that causes gonadal differentiation toward either a male or female phenotype. There are two described ways in which MSD genes can be overturned or lost. The first is through genetic mutations that lead to the loss of function of the MSD gene or lead to gain of function of a new MSD gene (Matsuda, 2018). This is demonstrated in studies of *Oryzias spp.* and is seen in the variety of MSD genes within the genus. In *O. latipes*, *dmy* is the MSD gene. In *O. luzonensis* and *O. dancena*, the MSD genes are *gsdfy* and *sox3y*, respectively. These genes are known players in sex determination pathways in fishes. The ancestral MSD gene in this genus is thought to by *dmy*, and mutations in regulatory regions of *gsdfy* and *sox3y* are theorized to have caused them gain the role of MSD in *O. luzonensis* and *O. dancea* (Matsuda, 2018). In all three of these species, the sex determining gene is on the Y chromosome in an XX-XY sex determination system.

The second possible way that a MSD gene can be lost or lose its function involves sex reversals, such that an individual with the genotype of one sex develops the phenotype of the opposite sex. Such sex reversals are known to be associated with temperature and stress, and can also be manipulated by hormone treatment. Sex reversals can lead to skewed sex ratios and possibly the entire loss of GSD. For example, mating a phenotypic male/genetic female with a genotypical female will result in all XX offspring, therefore eliminating the Y chromosome and male specific variation. This concept has been demonstrated through laboratory experiments with wild-caught sex reversed bearded dragons (Holleley et al., 2015). If stress and temperature extremes persist, the Y chromosome can be purged from the population completely, especially if the population in which sex reversal is occurring is small and isolated.

Here, we analyze North American Northern Pike for evidence of genetic sex determination systems.

# Methods

## Sequencing and SNP discovery

All SNPs used in this analysis were obtained according the methods described in Chapter 2.

## Phenotypic confirmation of sex

Individual fish were sexed either histologically, by dissection, or at the hatchery at the time they were striped of eggs or sperm. The majority of individuals were phenotypically sexed except for the four pike from Palmer Lake and the five from Yukon River, and therefore these groups were not included in any sex-specific analysis.

## Genome wide association study

A genome wide association study (GWAS) was performed using plink v. 1.9b_5.2-x86_64 (Purcell et al., 2007) with the "fisher-midp" option and visualized with the R software qqman v. 0.1.4 (Turner, 2017). For this analysis, the number of males and females was 21 and 17, respectively. We performed the GWAS again on each of our populations separately in attempt to identify regions of the genome that might reflect population specific sex biased variation. The sample sizes for these association studies were as follows: Chatanika River – 5 males, 5 females; Manitoba – 3 males, 3 females; New York – 6 males, 5 females; New Jersey – 3 males, 6 females; and Eastern North America – 16 males, 12 females.

## Sex-specific DAPC

We performed a DAPC based on sex among all Northern Pike and within populations and defined groups (from PCA/DAPC) with the R software Adegenet (Jombart, 2008). Once separated by population and groups, each subset of SNPs was filtered independently such that any SNP site that had a homozygous alternate allele was removed. A combination of homozygous reference, homozygous alternate, and heterozygote SNPs among individuals at a

particular site indicates that both males and females can harbour the SNP and therefore it is not sex-specific. As we are explicitly interested in sex-specific variation, these SNPs are not informative. Removing sites that contained homozygous alternate alleles left behind SNP sites that only contained homozygous reference and heterozygous alleles for analysis. We queried the resulting loadings tables for values that were specific to all-male or all-female variation, and compiled lists of genomic locations where sex-specific SNPs reside in each population. In order to identify regions of dense, sex-specific heterozygosity, we made histograms of sex-specific SNPs on each linkage group for each analysis.

## Sex-specific $F_{ST}$

To further probe genetic differences between males and females, we used VCFtools to calculate $F_{ST}$ values per site between all males and all females, and between males and females within populations. Per-site $F_{ST}$ values were plotted using the R program qqman (Turner, 2017).

## Sex-specific PCR assays

Primers were manually designed around male-specific SNPs such that both sexes would produce a 500 base amplicon, and males would produce a an additional smaller, 250 – 300 base amplicon. This smaller amplicon is generated by a third 'nested' primer whose sequence is based on male specific SNPs. We also used primers designed by Pan et al., (2017) to amplify exons and conserved regions of the *amhby* gene found in European Northern Pike. We used these primers to test for the presence of *amhby* in our re-sequenced samples and expanded the assay to additional samples from North America that were not used in sequencing. PCR conditions are listed in Table 9.

**Table 9. PCR conditions in sex targeted PCR assays**

ConserveAMH1 F1 & R1*

Reaction: 10ul

| Reagents | Stock Concentration | Vol Added (ul) | [] in PCR | Units |
|---|---|---|---|---|
| MgCL2 | 25 | 0.8 | 2 | mM |
| Primer F | 10 | 0.5 | 0.5 | uM |
| Primer R | 10 | 0.5 | 0.5 | uM |
| dNTPs | 10 | 0.2 | 0.2 | mM |
| buffer | 5 | 2 | 1 | x |
| taq | 5 | 0.05 | 0.025 | u/ul |
| water | 100 | 4.95 | 49.5 | % |
| DNA | 50-100 | 1 | 5-10 | ng/ul |

Thermocycler Conditions

| Cycle Name | Temperature ˚C | Time (min) | Number of Cycles |
|---|---|---|---|
| Initial denature | 95 | 5 | 1 |
| cycle denature | 95 | 0.5 | 35 |
| cycle anneal | 54 | 0.5 | |
| cycle extend | 72 | 1 | |
| Final extension | 72 | 10 | 1 |

SeqAMH_1 F4 & R4*

Reaction: 10ul

| Reagents | Stock Concentration | Vol Added (ul) | [] in PCR | Units |
|---|---|---|---|---|
| MgCL2 | 25 | 0.8 | 2 | mM |
| Primer F | 10 | 0.5 | 0.5 | uM |
| Primer R | 10 | 0.5 | 0.5 | uM |
| dNTPs | 10 | 0.2 | 0.2 | mM |
| buffer | 5 | 2 | 1 | x |
| taq | 5 | 0.05 | 0.025 | u/ul |
| water | 100 | 4.95 | 49.5 | % |
| DNA | 50-100 | 1 | 5-10 | ng/ul |

Thermocycler Conditions

| Cycle Name | Temperature ˚C | Time (min) | Number of Cycles |
|---|---|---|---|
| Initial denature | 95 | 5 | 1 |
| cycle denature | 95 | 0.5 | 35 |
| cycle anneal | 52 | 0.5 | |
| cycle extend | 72 | 1 | |
| Final extension | 72 | 10 | 1 |

Primer Set 24.5

Reaction:10ul

| Reagents | Stock Concentration | Vol Added (ul) | [] in PCR | Units |
|---|---|---|---|---|
| MgCL2 | 25 | 0.8 | 2 | mM |
| Primer F | 10 | 0.2 | 0.2 | uM |
| Primer R | 10 | 0.5 | 0.5 | uM |
| Probe | 10 | 0.3 | 0.3 | uM |
| dNTPs | 10 | 0.2 | 0.2 | mM |
| buffer | 5 | 2 | 1 | x |
| taq | 5 | 0.05 | 0.025 | u/ul |
| water | 100 | 4.95 | 49.5 | % |
| DNA | 50 - 100 | 1 | 5-10 | ng/ul |

Thermocycler Conditions

| Cycle Name | Temperature ˚C | Time (min) | Number of Cycles |
|---|---|---|---|
| Initial denature | 95 | 5 | 1 |
| cycle denature | 95 | 0.5 | 35 |
| cycle anneal | 52 | 0.5 | |
| cycle extend | 72 | 0.5 | |
| Final extension | 72 | 10 | 1 |

| Primer Name | Sequence 5' → 3' |
|---|---|
| SeqAMH1Fw4* | CAACATGGTGGCAACTAAGTG |
| SeqAMH1Rev4* | GGTAATATTTGTGCCCTGTG |
| ConserveAMH1_F1* | GTTACTTTTCTGCCTAGCGTGA |
| ConserveAMH1_R1* | CTATTACTAGTGTGGATAAGGCCG |
| 24.5 F | AATTACAGACCTCTACATGCT |
| 24.5 R | GATAGTCCCATAGATTGTAGA |
| 24.5 Probe | GCAAATGACGGCGCACTGTT |

Asterisks (*) denote primers from Pan (2017).

# Results

## Genome wide association study

Our GWAS for all sexed Northern Pike (38 in total, 17 female, 21 male) returned no significant results, indicating that a robust GSD mechanism is not shared across North American Pike (Figure 15). This is in contrast to what Pan et al., (2019) described for European Northern Pike, where a duplicated and differentiated *amhby* was identified on linkage group 24 in the sub-telomeric region. Attempts to perform GWAS analysis within populations were fruitless because of the low sample sizes when individuals were split into separate populations, and failed to return any significant results.

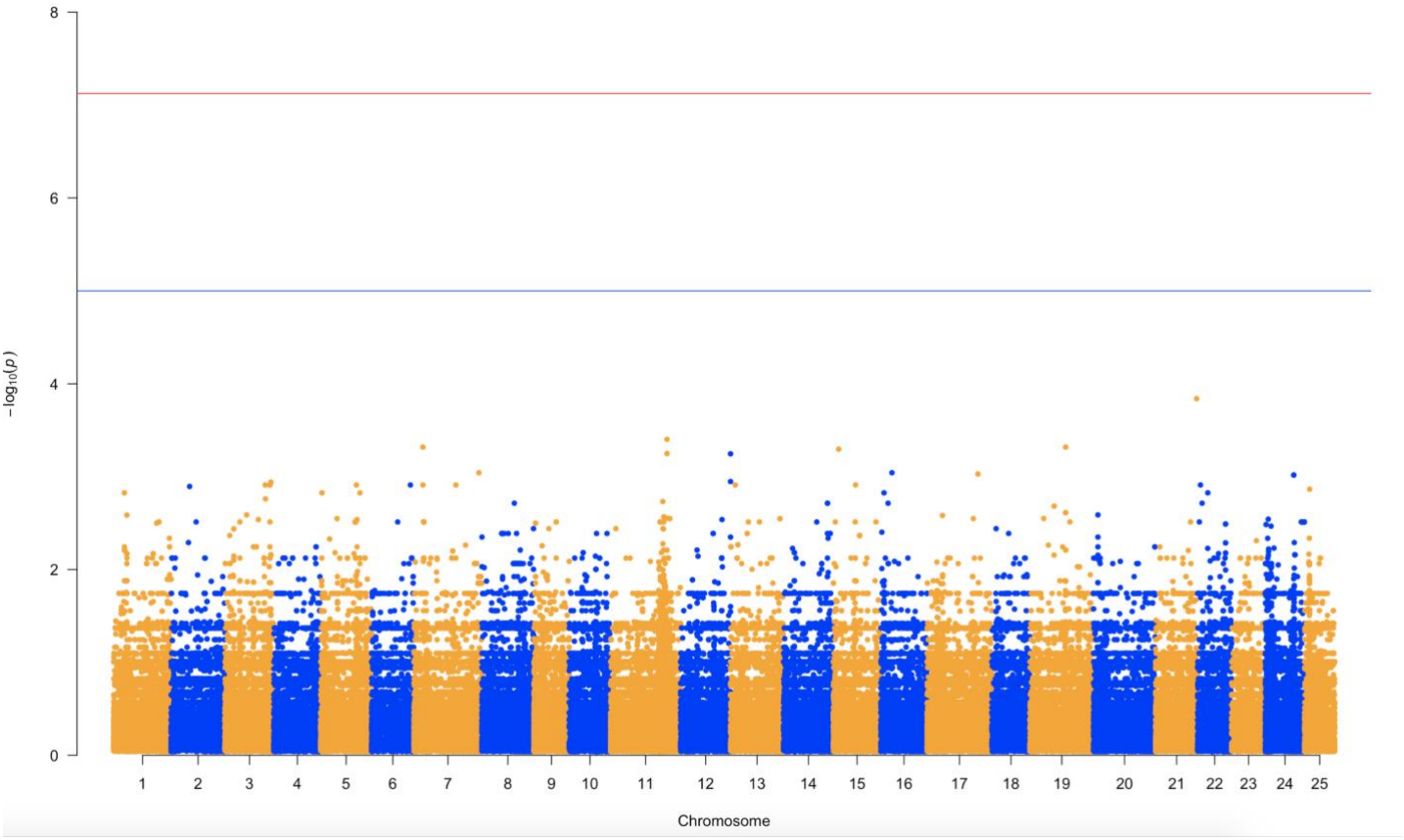



**Figure 15. Male - Female GWAS results.** Red line is the genome wide significance level (Bonferonni corrected). Blue line is chromosome-level significance.

## Sex-specific DAPC

In further attempts to identify a sex-specific region, we performed DAPCs with all males and females, with males and females from the Eastern North American group as defined by PCA, and with males and females within populations. When all sexed Northern Pike were analyzed together, not even one heterozygous locus in the genome was specific exclusively to males or females across North America. Likewise, when the Alaskan population was dropped from the analysis, we again found that zero heterozygous loci were associated exclusively with either sex throughout Eastern North America. Within populations, DAPC analysis was able to group males and females into distinct clusters for the Alaskan population only. For the remainder of the analyses, we ran the DAPC based on manually assigned sexes as opposed to groups identified by the software. Looking within each population, we were able to identify sex-specific SNPs in each population (Table 10). All populations showed hits of sex-specific SNPs across the genome, however the number of sex-specific SNPs varied in each population from 24 – 4,728. The location and density of sex-specific SNPs also varied in each population. Looking into each one of these SNPs individually is beyond the scope of this work, so we focussed on 50kb regions with the maximum number of sex-specific SNPs in each population. Density peaks of sex-specific SNPs were located on linkage group 24 in the Alaskan population in the same region identified by Pan et al., (2019), and on linkage group 10 in the New Jersey population (Figures 16 and 17).

**Table 10. Summary of populations/groups tested for the presence of sex-specific SNPs.**

| Population/Group | N Male | N Female | Sex-specific SNPs Found | Total Count of Sex-specific SNPs | Max Number of SNPs per 50kb Bin | Linkage Group | Heterozygous Sex | DNA Features where SNPs occur |
|---|---|---|---|---|---|---|---|---|
| All pike | 21 | 17 | N | 0 | - | - | - | |
| All Eastern NA pike | 16 | 12 | N | 0 | - | - | - | |
| Chatanika River | 5 | 5 | Y | 4728 | 773 | LG24 | M | exons, introns, non-coding |
| Manitoba | 3 | 3 | Y | 547 | 14 | LG22 | M | intergenic |
| New York | 6 | 5 | Y | 24 | 7 | LG11 | M | intergenic |
| New Jersey | 6 | 3 | Y | 626 | 54 | LG10 | F | exons, introns, non-coding |

All Northern Pike together do not share one common site where all males or all females are heterozygous. All Eastern NA Norhtern Pike include Manitoba, New York, and New Jersey populations. N = No. Y =Yes. M = Male. F = Female. Heterozygous Sex denotes the sex of the pike that have the max number of sex-specific SNPs per 50kb bin.

Alaskan pike have a dense, localized, male-specific region of heterozygous SNPs with 3,555 markers spanning a 500 kb region (Figure 16). This region is located on linkage group 24 between 650 – 1150 KB. Male specific SNPs here occur in introns, exons, and non-coding regions between genes. The affected genes in this region include disks large homolog 4-like (*dlg4*), two low affinity immunoglobulin gamma Fc receptor III-like genes (*fcgr3a*), Fc receptor-like protein 2 (*fcrl2*), two Fc receptor-like protein 5 genes (*fcrl5*), two butyrophilin subfamily 1 member A1-like genes (*btn1a1*), H3 lysine-79 specific histone-lysine N-methyltransferase (*dot1*), an uncharacterized gene (LOC105006602), and two non-coding RNAs (LOC114830354 & LOC114830343). We also detected male specific SNPs on the autosomal copy of *amh*, and observed that the read depth was greater for males across this gene. Our results confirm populations east of the great divide do not share this genetic sex determining mechanism (Figure16).

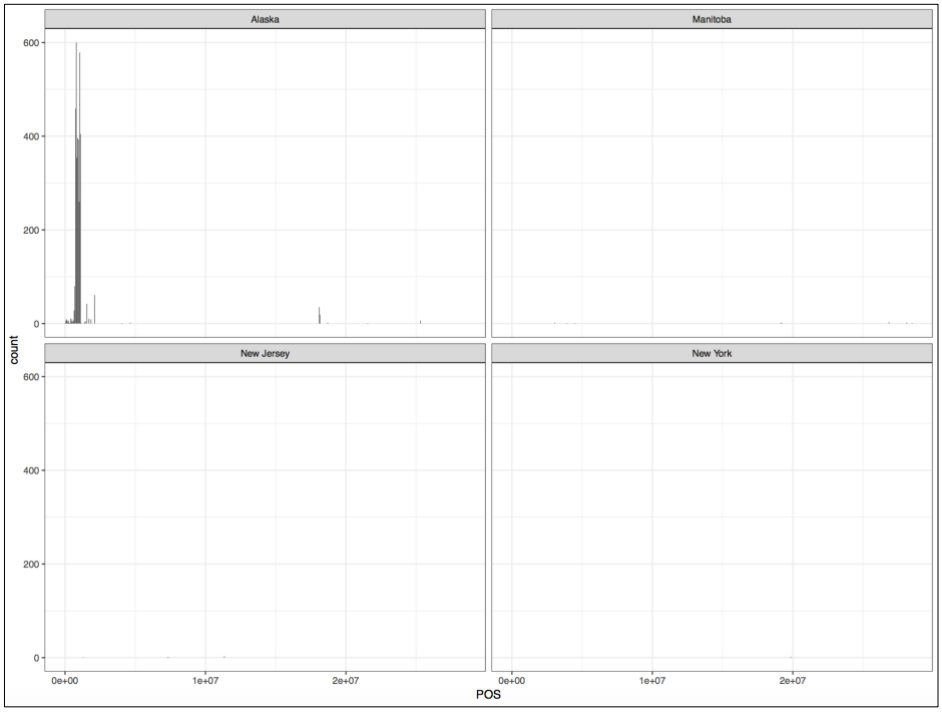



**Figure 16. Density of sex-specific SNPs on linkage group 24 in sexed populations.** The X-axis denotes the position of 50KB bins. The y-axis is the SNP count per 50KB bin.

New Jersey showed a weaker, yet well-defined signal on linkage group 10 between 17.98 – 18.10 MB, with 192 female specific heterozygous SNPs spanning 350 KB. Although this signal is much weaker than that seen on LG24 in Alaskan pike, it is the only region in the genome where a dense region of sex-specific SNPs occurs, and is therefore the most likely region to be responsible for genetic sex determination, if proven to be robust. SNPs occur in exons, introns, and non-coding DNA sequence. The affected genes include an uncharacterised protein (LOC 105012673), peroxisomal biogenesis factor II beta (*pex11b*), Ras-like protein without CAAX1 (*rit1*), snynaptotagmin-11 (*syt11*), mothers against decapentaplegic homolog 4 – like (*smad4*), SNARE associated protein (*snapin*), chromatin target of PMRT1 (*chtop1*), and interleukin enhancer binding factor 2 (*ilf2*).

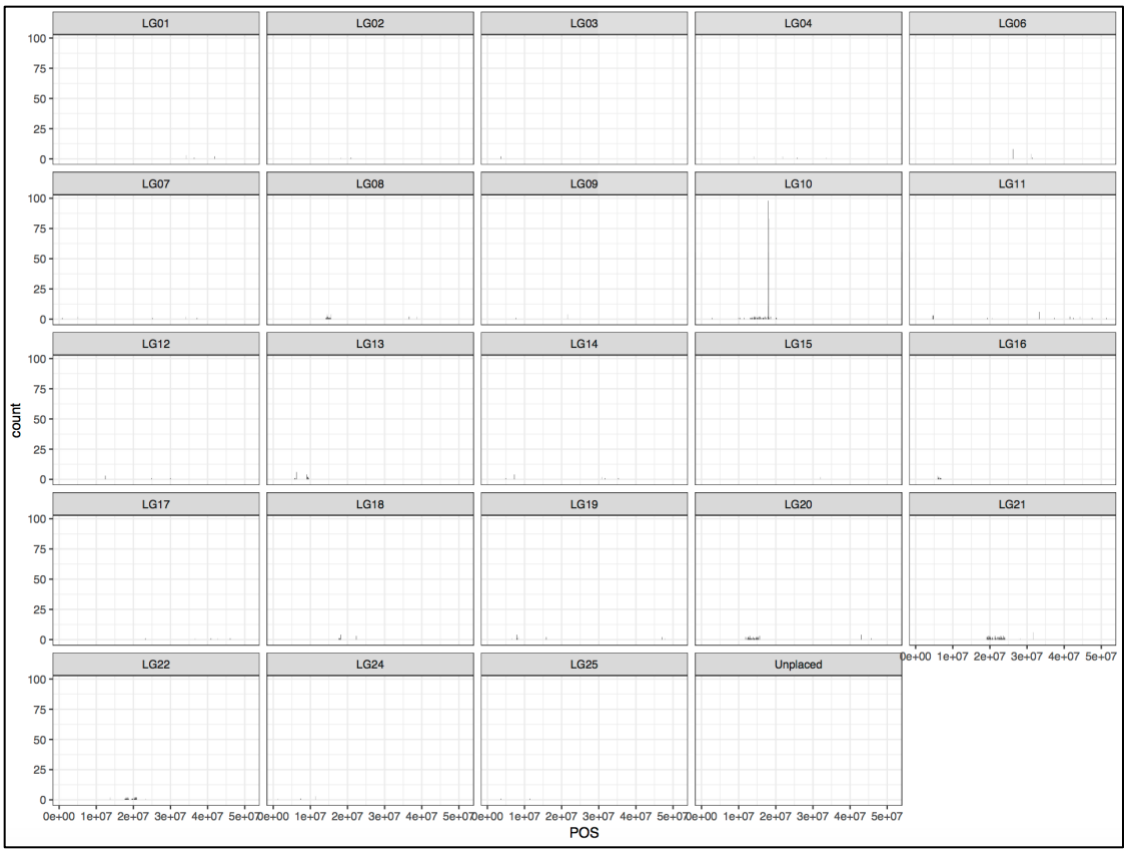



**Figure 17. Density of sex-specific SNPs in New Jersey pike across all linkage groups.** The x-axis is the position of 50KB bins. The y-axis is the count of SNPs per 50KB bin.

Populations from Manitoba and New York had a maximum of 14 and 7 SNPs per 50kb bin, respectively. In both populations, these SNPs occurred in intergenic regions. Although a total of 547 male specific heterozygous SNPs were identified the Manitoba population, most of them occurred as a single SNP in a 50Kb bin. As there were only three males and three females in this population, further investigations on additional individuals would be required to eliminate false positive identification of sex-specific SNPs. In New York, a total of 24 male specific SNPs were identified across the genome. This is a remarkably low level of differentiation between sexes.

## PCR assays

The primer sets tested show consistent results across populations. All three sets used positively identify the presence of *amhby* in male pike from Chatanika River, Alaska, and Castlegar, BC. These primer sets also detected *amhby* in one of five pike from the Yukon River, although these fish were not sexed so we cannot confirm the phenotypic sex. In all females, and males from the remainder of North America *amhby* was not detected (Table 11). In Palmer Lake, *amhby* is not detected. However, these fish were not sexed, and therefore could potentially all be female. Because of this, we cannot confirm the presence or absence of *amhby* in Palmer Lake.

**Table 11. Positive detections of *amhby* and the LG 24 sex determining region.**

| Primer Pair name | Region amplified | Chatanika River | | Castlegar | | Eastern NA | | Hootalinqua | Palmer Lake |
|---|---|---|---|---|---|---|---|---|---|
| | | Males | Females | Males | Females | Males | Females | Not sexed | Not sexed |
| Amhby_conserve F1 & R1* | Amhby: partial exon 2 | 5/5 | 0/5 | 3/3 | 0/3 | 0/16 | 0/11 | 1/5 | 0/4 |
| SeqAMH_1 F4 & R4* | Amhby: partial exon 7 | 5/5 | 0/5 | 3/3 | 0/3 | 0/16 | 0/11 | 1/5 | 0/4 |
| 24.5 | LG24: 996,878 - 997,339 | 5/5 | 0/5 | 3/3 | 0/3 | 0/16 | 0/11 | 1/5 | 0/4 |

Numerators denote the number of Northern Pike in which *amhby* was detected and denominators denote the total number of Northern Pike of the noted sex. The asterisk (*) indicates primers developed by Pan (2017). Eastern NA includes Charlie Lake (1M), Manitoba (3M, 3F), New York (6M, 5F), and New Jersey (6M, 3F).

## Sex-specific $F_{ST}$

Genome-wide mean male-female $F_{ST}$ values are listed in Table X. Means were computed using SNP sites. Only Chatanika River and New Jersey populations returned positive genome-wide means of male-female $F_{ST}$, indicating that these are the only populations that have any measurable genetic segregation between males and females. Within populations, negative values of $F_{ST}$ are effectively zero. Therefore, in New York and Manitoba populations, and in all sexed pike when analyzed together, males and females share genetic variation and SNPs are not harboured exclusively in one sex, on average.

**Table 12. Genome wide mean male-female $F_{ST}$s by population.**

| Population | Male-Female $F_{ST}$ |
|---|---|
| All pike | -0.0071 |
| CR | 0.0016 |
| MB | -0.0169 |
| NY | -0.0021 |
| NJ | 0.0036 |

"All pike" include only those that have been sexed. Means include SNP sites only.

To look at how $F_{ST}$ measurements are distributed along chromosomes, we plotted raw values along each linkage group for each sexed population and for all sexed pike. Plotted below is male-female $F_{ST}$ comparison for Chatanika River across linkage groups (Figure 18). Chatanika River is used as an example here because this population produced a sex-specific signal on linkage group 24 between 0.2 – 1 MB, detectable through DAPC analysis with heterozygous SNPs (this thesis). The male-female $F_{ST}$ plot below shows there is a density of SNPs at that region, but it is not the highest peak across all linkage groups. There are peaks of greater amplitude on linkage groups four, eight, eleven, seventeen, as well as a different location on LG24. When we trace these peaks back to the associated genotypes, we see that peaks of $F_{ST}$ can be associated with regions where males and females are truly genetically segregated, but also with areas where there is an uneven mix of heterozygous, homozygous alternate, and homozygous reference genotypes across the sexes. Male-female $F_{ST}$ analysis was not a clearly reliable method of detecting regions of sex-specific variation in our populations.

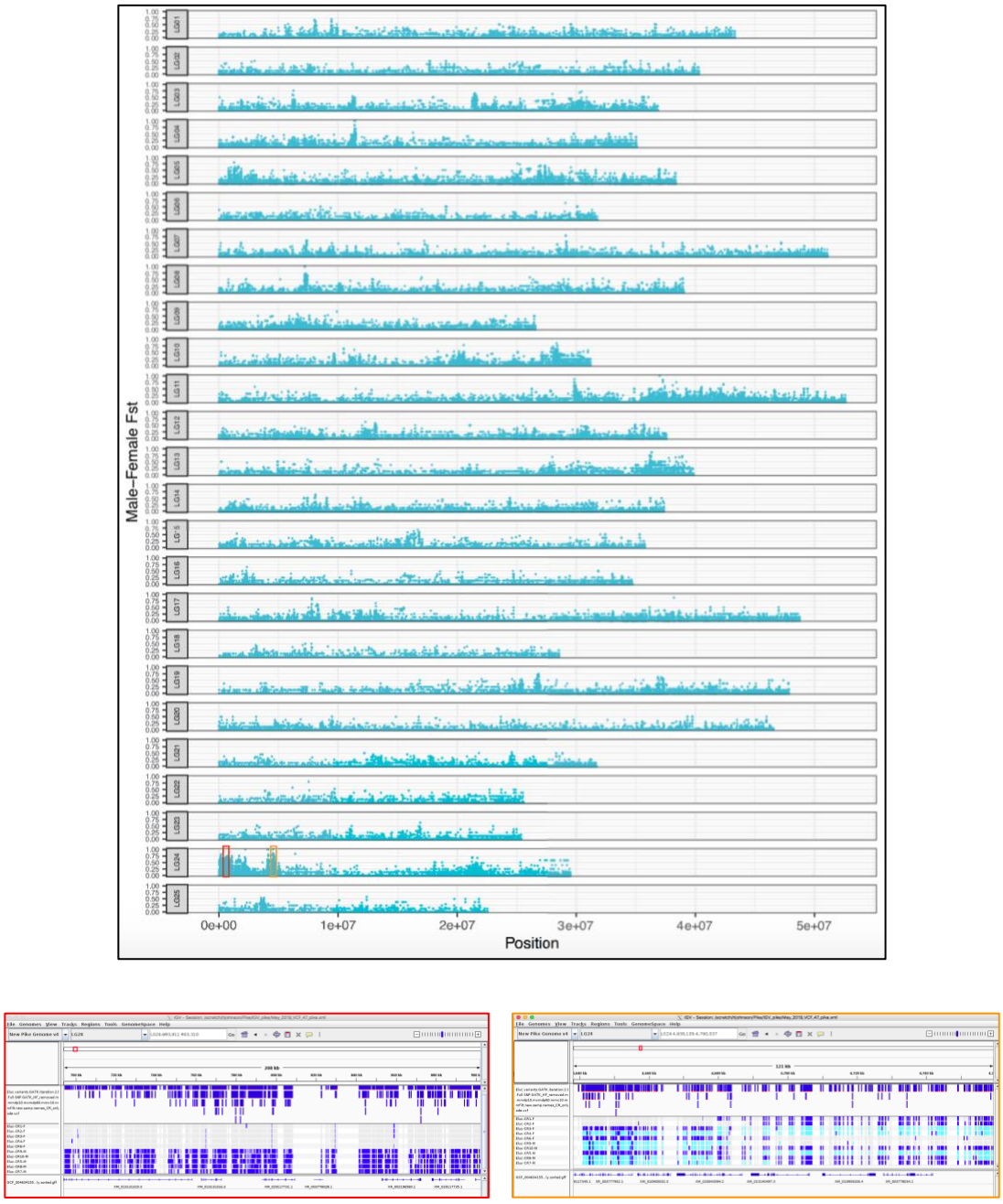



**Figure 18. Male-female FST along linkage groups in Chatanika River pike.** Negative $F_{ST}$ values were set to zero for this plot to ease visualization. (A) is the genome-wide distribution of $F_{ST}$ values, organized by linkage group. (B) and (C) are expanded images from Integrated Genomics Viewer of two separate regions on LG24 where $F_{ST}$ values peak. Notice that the peaks are of similar amplitude but the peak at 700 KB is associated with male specific heterozygous SNPs (dark blue), but the peak at 4.8MB is not.

# Discussion

## *amhby* is present in Northern Pike populations west of the Continental Divide and absent in population east of the Continental Divide

European Northern Pike use an ancient duplicated *amhby* as the master sex determination gene (Pan et al., 2019). Here, we confirm that at least three populations of Northern Pike in North America share this sex determining gene (Chatanika River, Hootalinqua, and Castlegar pike). These populations are all located west of the Continental Divide. Intriguingly, this sex determining mechanism was absent in all populations east of the Continental Divide, and not a single sex-specific SNP was shared among all males or all females across North America. This result was replicated by a colleague in our lab who performed a k-mer analysis on the same data (Whitehead, 2019). In a k-mer analysis, sex-specific signal detection does not rely on aligning reads to a reference genome and therefore eliminates the possibility of failed detection from non-mappable reads, thus lending additional support to the loss of *amhby* and associated male specific heterozygosity east of the Continental Divide.

In the previous chapter, we found that populations east of the Continental Divide harbour significantly less heterozygosity than populations in the Yukon drainage basin and that Northern Pike from each region are reproductively isolated. Genetics of the Hootalinqua population suggested that the Yukon River and associated tributaries may have at one time been a link between eastern and western North America. We suggested that at some time during the last ice age (120 – 12 thousand years ago), Northern Pike crossed from Europe to Alaska through Beringia, and that a small founding population colonized the remainder of North America. We believe that the loss of *amhby* is related to this founding event. Indeed, *amhby* and the TGF-β pathway are influenced by temperature and stress (Baroiller et al., 2009; Goikoetxea et al.,

2017), both of which may have been relevant given the geological dynamic during the last ice age. Cortisol is believed to down-regulate genes involved in female development and up-regulate *amh* producing a masculinizing effect that can cause XX individuals to undergo environmental sex reversal and develop as males [reviewed in (Goikoetxea et al., 2017)]. If such a sex reversed individual mates with a typical XX female, all (genotypically) female offspring would result. If the stressor persisted over generations, the effect of this would be to weaken or eliminate the GSD mechanism and decrease genome wide variation. This is the pattern we see in North American Northern Pike.

The absence of a detectable signal in central North American populations (Manitoba, New York) may be due to a few possible factors. Because we have low sample sizes of males and females within populations, we do not have the ability to detect a GSD that may only involve one or a few SNPs, as is characteristic of a newly acquired GSD mechanism. We would need data from many more males and females from each population to achieve a significant detection if a solely GSD mechanism exists. It is also possible that these populations use a mechanism of sex determination that cannot be detected by our methods such as epigenetic or post-translational modifications. Epigenetic processes include DNA methylation and histone modification, both of which act to affect gene expression and regulation (Gibney and Nolan, 2010) and cannot be detected by the whole genome resequencing approach taken in this thesis. Post-translational modifications, such as alternative splicing or activity of non-coding RNAs, can affect gene expression and/or protein function, and cannot be detected by our resequencing approach either. Possibly, ESD or a combination of ESD and GSD is at play. Observations of skewed sex ratios in the St. Lawrence River (Huffman et al., 2014) and Gilbert Lake, Wisconsin (Priegel and Krohn, 1975), as well as an absence of a detectable GSD lends credibility to the prospect of ESD

in these populations. Climate change has been linked to highly skewed sex ratios in sea turtles (Jensen et al., 2018), and this effect is expected to become more frequent in ectotherms (Massey et al., 2019). An ESD could make Northern Pike populations vulnerable to the warming climate (Bókony et al., 2017). In any case, the loss of a GSD in these populations provides a snapshot of a sex determination system in transition in Northern Pike.

The presence of *amhby* in invasive Castlegar males is puzzling, as this the representative fish from this population clustered tightly with eastern North American populations and is thought to have originated east of the Continental Divide. The presence of *amhby* draws into question the possibility of another refugia that may have maintained this GSD. Alternately, it may hint at an introduction of Northwestern Northern Pike to the southwestern region of their range.

## A possible female heterozygous GSD mechanism in New Jersey pike

We identified a region of dense female specific heterozygous SNPs in New Jersey pike on linkage group 10. Together, the reported function of affected genes lends support to the involvement of this region in sex determination. Broadly, these genes have functions relating to estrogen signalling (*chtop1*) (van Dijk et al., 2010; Fanis et al., 2012), the TGF-β signalling pathway (*smad4*) (Du et al., 2018; Zhao et al., 2018), synaptic vesicle docking and fusion (*snapin*) (Granata et al., 2008; Ilardi et al., 1999), blockage of neuronal signalling via inhibition of endocytosis and vesicle recycling (*syt11*) (Wang et al., 2016, 2018), peroxisome proliferation (*pex11b*) (Thoms and Erdmann, 2005), regulation of stress-related signalling pathways and promotion of cell survival, neuronal development and regeneration (*rit1*) (Shi et al., 2011), and influencing the transcription and activity of interleukin 2 and 3 (*ilf2*) (Ryff and Pestka, 2013).

Estrogen is a known sex hormone, and *chtop1* encodes a protein that is crucial to the activation of estradiol-dependent transcription (van Dijk et al., 2010), and plays a major role in cell processes such as differentiation, growth and proliferation, and apoptosis (Fanis et al., 2012; Izumikawa et al., 2018). Genes associated with estrogen signalling have been nominated as master sex determining genes in other ZW systems (Kawase et al., 2018; Koyama et al., 2019; Purcell et al., 2018).

The TGF-β signalling pathway seems to be the shared commonality among identified sex determination genes in teleosts. *Smad4* is a known modulator of the TGF-β signalling pathway and has been shown to be a 'switch' for TGF-β function. It is also susceptible to extensive regulation by pathways such as MAPK, PI3K/AKT, and WNT/β-cantenin (Zhao et al., 2018), some of which are recognized pathways in sex determination systems (Baetens et al., 2019; Bogani et al., 2009; Navara, 2018; Zheng et al., 2018). Feedback of *smad4* in the TGF-β signalling patway has been shown to influence apoptosis in mammalian granulosa cells (Du et al., 2018). Thus, the presence of genes dense with sex-specific SNPs known to be involved in the TGF-β signalling pathway lends support to the possibility of this region being a sex determining one.

The presence of an immune related gene in this region (*ilf2*) that is dense with sex-specific heterozygous SNPs suggests that it has a functional role in sex determination. *ilf2* regulates the transcription of interleukin 2. Interleukin 2 has been shown to have functions in preventing the body from locally attacking its own cells and to be involved in the autoimmune response (Ryff and Pestka, 2013), It seems reasonable then, to suggest that this gene may have a role in either preventing or allowing attack of developing male or female gonadal cells.

The genes mentioned above have functions in teleost sex determination pathways, neuronal signalling, apoptosis, and cell degradation, suggesting that if this region plays a role in genetic sex determination, it may be through estradiol and/or TGF-β dependent degradation or repression of target cells, and would signify a transition from an XX/XY system in western North America, to a ZZ/ZW system in this eastern population.

## Commonalities between Northern Pike GSD and Salmonid GSD

Commonalities exist between the GSDs discussed here in Northern Pike, and the sex determining gene SDY in salmonids. All three of these systems are linked to immunity. In Northern Pike with *amhby*, the sex determining region contains multiple copies of immune system genes (Fc-like receptors) that are heavy with male specific SNPs. In New Jersey Pike, *ilf2* is rich with female specific SNPs. *sdy* itself is recognized as an immune related gene (Yano et al., 2012).

*Sdy* is a unique protein with a domain similar to *Irf3* and possesses DNA binding ability. In 2003, a study detailed the similarity of *Irf3* to the SMAD/FHA (forkhead associated) superfamily (Qin et al., 2003). SMAD proteins are directly involved in the TGF-β signalling pathway, which is a common pathway of teleost sex determination. This suggests that the two genetic sex determination systems in Northern Pike discussed above, as well as the *sdy* system in salmonids are all linked to the TGF-β signalling pathway. Similarities are summarized in Table 13.

It has been shown that *irf3*-dependent interferon response is evolutionarily conserved among vertebrates, although that response may be obtained through different pathways in fish than other vertebrates (Sun et al., 2010). The interferon response is well characterised and is known to protect against viral infection by warning the body of a foreign entity and signalling cells

**Table 13. Comparison of Northern Pike and Salmonid genetic sex determination systems.**

| Group: | *E. lucius* LG24 (Europe, northwestern N.A.) | *E. lucius* LG10 proposed (New Jersey) | Salmonids |
|---|---|---|---|
| Heterozygous Sex: | Males | Females | Males |
| SD gene: | *amhby* | NOT DETERMINED | *sdy* |
| General summary of function: | Contains TGF-β domains. TGF-β pathway known to be involved in SD in many fish. Controls proliferation, differentiation, and can modulate cell invasion and immune regulation. | Mediates TGFβ signalling (*smad4*), activates estradiol dependent transcription (*chtop1*), prevents autoimmune attack (*ilf2*). | Funtional domains related to *Irf3* and part of the SMAD/FHA Superfamily. SMAD known to interact with TGF-β pathways. I*rf3*-dependent interferon response protects against viral infection (prevents transcription, promotes apoptosis). |
| Other affected genes (i.e. other genes in area with sex-specific SNPs) | *dlg4*, multiple Fc receptor variants, *btn1a1*, *dot1*, an uncharacterized gene and two uncharacterized ncRNAs. | *pex11b*, *rit1*, *syt11*, *snapin*, and an uncharacterized gene | Varies among salmonids; is non-syntenic. |

to recognize and destroy target particles, prevent transcription, and self-destruct through apoptosis (Collet, 2014; Haller and Kochs, 2002). Given this, and our knowledge that SMAD proteins are involved in the TGF-β signalling pathway, it seems reasonable to suggest that the salmonid sex determination gene *sdy* functions by either causing or preventing attack on developing gonadal cells. Furthermore, interleukin 2 (NJ pike) has the effect of preventing attack, whereas the interferon response (salmonids) has the effect of initiating attack. Given that New Jersey pike are female heterozygous, and salmonids are male heterozygous at the sex determining region, this might suggest that female gonadal tissue could be a target of a sex-determination related immune response. Perhaps female heterozygosity in has the effect of preventing female gonadal tissue from being destroyed in New Jersey Pike, whereas male heterozygosity in salmonids and *amhby*-pike have the effect of attacking female gonadal tissue during the sex determination period of development.

## Limitations and considerations

Our ability to identify sex-specific regions of the Northern Pike genome was compromised by the low numbers of males and female surveyed per population. Additional samples from Northern Pike's natural range would help elucidate where and when *amhby* was lost, and help clarify the recolonization route of North America after the retreat of the Pleistocene glaciers.

## Chapter 3 conclusions

At least two sex determination mechanisms exist in Northern Pike in North America. In a population from Chatanika River, Alaska, we identified *amhby* (as described by Pan et al., (2019)) to be perfectly associated with males in an XX/XY GSD mechanism. This system was also detected in a Yukon River population, as well as in Northern Pike that are invasive to the Columbia River in Castlegar, B.C. East of the Continental Divide, this sex determination mechanism and associated male heterozygosity was absent, indicating a loss of this GSD. We were unable to detect a sex-specific signal from southern Manitoba or St. Lawrence River populations, suggesting that ESD might be occurring (as evidenced by skewed sex ratios), or that just one or a few SNPs are currently associated with sex determination in these populations. On the eastern coast of North America in Hackettstown, New Jersey, we found evidence of a possible ZZ/ZW female heterozygous sex-determination system, where female-specific SNPs occur in genes involved in estrogen signalling, the TGF-β signalling pathway, and immunity. Additional samples from New Jersey would be required to verify this proposition.

# Chapter 4

# Conclusions

In this thesis, it is confirmed that Northern Pike have a low level of genetic variation genome-wide, with an average of just one heterozygous SNP every 10,000 bases. Genetic variation is not spread evenly throughout the genome but occurs in dense regions separated by lengthy spans of homozygosity. Concentrated regions of variation contain genes which influence immunity and cell adhesion/basement membrane structure, and are rich in multiple copy number variants. Such a pattern suggests that in Northern Pike, phenotypic variety in immunity and the structure and nature of tissue permeability is crucial. SNPs that differentiate populations occur in genes that have functions in the nervous system and development. On a broader scale, this study demonstrates that population expansion and persistence is not necessarily dependent on overall heterozygosity or the total amount of polymorphic loci in the genome.

At least two different sex determination mechanisms are present in North American Pike. The male heterozygous XX/XY system described by Pan et al., (2019) was identified in three populations, including the invasive Northern Pike in Castlegar, B.C. We were unable to identify a sex-specific signal in Manitoba and St. Lawrence River pike. In these populations, GSD may be at a very early stage of differentiation or ESD may be occurring. We were not able to make any conclusions about a sex determination mechanism in Palmer Lake as these fish were not sexed. In New Jersey pike, we found evidence for a female heterozygous ZZ/ZW GSD system involving pathways known to be prominent in sex determination.

Genome-wide patterns of genetic variation as well as the presence of the *amhby* sex determination mechanism in Northern Pike from the Yukon River drainage basin provides evidence that Northern Pike colonized North America from Europe through Beringia. The reduction in the number of heterozygous SNPS and minor allele frequencies in Northern Pike east of the Continental Divide, along with the loss of *amhby* and all male specific variation indicates that Northern Pike in this geographic area are much younger and likely descended from a small number of individuals. Northern Pike from Palmer Lake in Northwestern B.C. have a unique genetic signal and are isolated from other populations studied here.

Although Castlegar and New Jersey pike cluster tightly with Manitoban and St. Lawrence populations, the presence of *amhby* in Castlegar, and evidence of female heterozygous sex determination system in New Jersey confuse the notion of Northwest to Southeast dispersal across North America. Evidence presented here supports the hypothesis that Northern Pike originally colonized North America from the Yukon drainage. However, multiple ice sheet advances and retreats could have left small isolated refugia south of the glacial boundary, which extended across what is now the northern United States. Additional sampling of Northern Pike along the extent of their southern and northern boundaries would be helpful in disentangling re-colonization origins after the last glacial retreat.

# Literature Cited


Abadía-Cardoso, A., Freimer, N.B., Deiner, K., and Garza, J.C. (2017). Molecular Population Genetics of the Northern Elephant Seal Mirounga angustirostris. J Hered *108*, 618–627.

Abbott, J.K., Nordén, A.K., and Hansson, B. (2017). Sex chromosome evolution: historical insights and future perspectives. Proc Biol Sci *284*.

Amos, W., and Balmford, A. (2001). When does conservation genetics matter? Heredity *87*, 257–265.

Ancona, S., Dénes, F.V., Krüger, O., Székely, T., and Beissinger, S.R. (2017). Estimating adult sex ratios in nature. Phil. Trans. R. Soc. B *372*, 20160313.

Baetens, D., Verdin, H., De Baere, E., and Cools, M. (2019). Update on the genetics of differences of sex development (DSD). Best Practice & Research Clinical Endocrinology & Metabolism.

Bao, L., Tian, C., Liu, S., Zhang, Y., Elaswad, A., Yuan, Z., Khalil, K., Sun, F., Yang, Y., Zhou, T., et al. (2019). The Y chromosome sequence of the channel catfish suggests novel sex determination mechanisms in teleost fish. BMC Biol *17*.

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

Barrett, R.D.H., and Schluter, D. (2008). Adaptation from standing genetic variation. Trends in Ecology & Evolution *23*, 38–44.

Bateson, Z.W., Hammerly, S.C., Johnson, J.A., Morrow, M.E., Whittingham, L.A., and Dunn, P.O. (2016). Specific alleles at immune genes, rather than genome-wide heterozygosity, are related to immunity and survival in the critically endangered Attwater's prairie-chicken. Molecular Ecology *25*, 4730–4744.

Beirowski, B., Gustin, J., Armour, S.M., Yamamoto, H., Viader, A., North, B.J., Michán, S., Baloh, R.H., Golden, J.P., Schmidt, R.E., et al. (2011). Sir-two-homolog 2 (Sirt2) modulates peripheral myelination through polarity protein Par-3/atypical protein kinase C (aPKC) signaling. Proc. Natl. Acad. Sci. U.S.A. *108*, E952-961.

Bekkevold, D., Jacobsen, L., Hemmer-Hansen, J., Berg, S., and Skov, C. (2015). From regionally predictable to locally complex population structure in a freshwater top predator: river systems are not always the unit of connectivity in Northern Pike *Esox lucius*. Ecology of Freshwater Fish *24*, 305–316.

Benestan, L., Moore, J.-S., Sutherland, B.J.G., Luyer, J.L., Maaroufi, H., Rougeux, C., Normandeau, E., Rycroft, N., Atema, J., Harris, L.N., et al. (2017). Sex matters in massive

parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. Molecular Ecology *26*.

Bianco, P.G., and Delmastro, G.B. (2011). Recenti novità tassonomiche riguardanti i pesci d'acqua dolce autoctoni in Italia e descrizione di una nuova specie di luccio (S.l.: IGF Publishing).

Biswas, S., and Akey, J.M. (2006). Genomic insights into positive selection. Trends in Genetics *22*, 437–446.

Bogani, D., Siggers, P., Brixey, R., Warr, N., Beddow, S., Edwards, J., Williams, D., Wilhelm, D., Koopman, P., Flavell, R.A., et al. (2009). Loss of Mitogen-Activated Protein Kinase Kinase Kinase 4 (MAP3K4) Reveals a Requirement for MAPK Signalling in Mouse Sex Determination. PLOS Biology *7*, e1000196.

Bókony, V., Kövér, S., Nemesházi, E., Liker, A., and Székely, T. (2017). Climate-driven shifts in adult sex ratios via sex reversals: the type of sex determination matters. Phil. Trans. R. Soc. B *372*, 20160325.

Bradford, M.J., Tovey, C.P., and Herborg, L.-M. Biological Risk Assessment for Northern Pike (Esox lucius), Pumpkinseed (Lepomis gibbosus), and Walleye (Sander vitreus) in British Columbia. 54.

Brajenovic, M., Joberty, G., Küster, B., Bouwmeester, T., and Drewes, G. (2004). Comprehensive proteomic analysis of human Par protein complexes reveals an interconnected protein network. J. Biol. Chem. *279*, 12804–12811.

Broad Institute (2017). Picard Tools - By Broad Institute.

Campbell, M.A., Alfaro, M.E., Belasco, M., and López, J.A. (2017). Early-branching euteleost relationships: areas of congruence between concatenation and coalescent model inferences. PeerJ *5*, e3548.

Casselman (1974). External Sex Determination of Northern Pike, Esox lucius Linnaeus. Transactions of the American Fisheries Society *103*, 343–347.

Casselman, and Lewis, C. (1996). Habitat requirements of northern pike (Essox lucius). Can. J. Fish. Aquat. Sci. *53*, 161–174.

Chambers, J.M., Freeny, A.E., Heiberger, R.M., Freeny, A.E., and Heiberger, R.M. (2017). Analysis of Variance; Designed Experiments.

Chen, X., An, Y., Gao, Y., Guo, L., Rui, L., Xie, H., Sun, M., Lam Hung, S., Sheng, X., Zou, J., et al. (2017). Rare Deleterious PARD3 Variants in the aPKC-Binding Region are Implicated in the Pathogenesis of Human Cranial Neural Tube Defects Via Disrupting Apical Tight Junction Formation. Hum. Mutat. *38*, 378–389.

Collet, B. (2014). Innate immune responses of salmonid fish to viral infections. Developmental & Comparative Immunology *43*, 160–173.

Collins, T., Stone, J.R., and Williams, A.J. (2001). All in the Family: the BTB/POZ, KRAB, and SCAN Domains. Molecular and Cellular Biology *21*, 3609–3615.

Craig, J. (2008). A short review of pike ecology. Hydrobiologia *601*, 5–16.

Crossman, E.J. (1991). Introduced Freshwater Fishes: A Review of the North American Perspective With Emphasis on Canada. Can. J. Fish. Aquat. Sci. *48*, 46–57.

D. S. Witt, J., J. Zemlak, R., and Taylor, E. (2011). Phylogeography and the origins of range disjunctions in a north temperate fish, the pygmy whitefish (Prosopium coulterii), inferred from mitochondrial and nuclear DNA sequence analysis. Journal of Biogeography *38*, 1557–1569.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Davidson, W.S., Huang, T.-K., Fujiki, K., von Schalburg, K.R., and Koop, B.F. (2009). The sex determining loci and sex chromosomes in the family salmonidae. Sex Dev *3*, 78–87.

Davies, J. (2016). The birth of the Anthropocene (Oakland, California: University of California Press).

Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Current Opinion in Genetics & Development *41*, 44–52.

Denys, G.P.J., Dettai, A., Persat, H., Hautecœur, M., and Keith, P. (2014). Morphological and molecular evidence of three species of pikes Esox spp. (Actinopterygii, Esocidae) in France, including the description of a new species. Comptes Rendus Biologies *337*, 521–534.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Devlin, R.H., and Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture *208*, 191–364.

DeWoody, J.A., and Avise, J.C. (2000). Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. Journal of Fish Biology *56*, 461–473.

DFO (2011). Science advice from a risk assessment of northern pike (Esox lucius) in British Columbia (Department of Fisheries and Oceans).

van Dijk, T.B., Gillemans, N., Stein, C., Fanis, P., Demmers, J., van de Corput, M., Essers, J., Grosveld, F., Bauer, U.-M., and Philipsen, S. (2010). Friend of Prmt1, a Novel Chromatin Target of Protein Arginine Methyltransferases. Mol Cell Biol *30*, 260–272.

Dlugosch, K.M., and Parker, I.M. (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Molecular Ecology *17*, 431–449.

Du, X., Pan, Z., Li, Q., Liu, H., and Li, Q. (2018). SMAD4 feedback regulates the canonical TGF-β signaling pathway to control granulosa cell apoptosis. Cell Death & Disease *9*, 151.

Ehlers, J., and Gibbard, P. (2011). Quaternary Glaciation. In Encyclopedia of Snow, Ice and Glaciers, V.P. Singh, P. Singh, and U.K. Haritashya, eds. (Dordrecht: Springer Netherlands), pp. 873–882.

Fanis, P., Gillemans, N., Aghajanirefah, A., Pourfarzad, F., Demmers, J., Esteghamat, F., Vadlamudi, R.K., Grosveld, F., Philipsen, S., and Dijk, T.B. van (2012). Five Friends of Methylated Chromatin Target of Protein-Arginine-Methyltransferase[Prmt]-1 (Chtop), a Complex Linking Arginine Methylation to Desumoylation. Molecular & Cellular Proteomics *11*, 1263–1273.

Forsman, A., Tibblin, P., Berggren, H., Nordahl, O., Koch-Schmidt, P., and Larsson, P. (2015). Pike Esox lucius as an emerging model organism for studies in ecology and evolutionary biology: a review. J Fish Biol *87*, 472–479.

Fowler, B.L.S., and Buonaccorsi, V.P. (2016). Genomic characterization of sex-identification markers in Sebastes carnatus and Sebastes chrysomelas rockfishes. Mol Ecol *25*, 2165–2175.

Frankham, R. (2005). Genetics and extinction. Biological Conservation *126*, 131–140.

Gharbi, K., Gautier, A., Danzmann, R.G., Gharbi, S., Sakamoto, T., Høyheim, B., Taggart, J.B., Cairney, M., Powell, R., Krieg, F., et al. (2006). A Linkage Map for Brown Trout (Salmo trutta): Chromosome Homeologies and Comparative Genome Organization With Other Salmonid Fish. Genetics *172*, 2405–2419.

Gibney, E.R., and Nolan, C.M. (2010). Epigenetics and gene expression. Heredity *105*, 4–13.

Gillespie, J.H. (2004). Population genetics: a concise guide (Baltimore [MD] ; London: Johns Hopkins University Press).

Goikoetxea, A., Todd, E.V., and Gemmell, N.J. (2017). Stress and sex: does cortisol mediate sex change in fish? Reproduction *154*, R149–R160.

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

Government of Canada, F. and O.S.S. (2016). 2010 Survey of Recreational Fishing in Canada | Fisheries and Oceans Canada.

Granata, A., Watson, R., Collinson, L.M., Schiavo, G., and Warner, T.T. (2008). The Dystonia-associated Protein TorsinA Modulates Synaptic Vesicle Recycling. J. Biol. Chem. *283*, 7568–7579.

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

Grande, Laten, H., López, J.A., and Quattro, J.M. (2004). Phylogenetic Relationships of Extant Esocid Species (Teleostei: Salmoniformes) Based on Morphological and Molecular Characters. Copeia *2004*, 743–757.

Graur, D., and Li, W.-H. (2000). Fundamentals of molecular evolution (Sinauer Associates).

Guillerault, N., Loot, G., Blanchet, S., and Santoul, F. (2018). Catch-related and genetic outcome of adult northern pike Esox lucius stocking in a large river system. Journal of Fish Biology *93*, 1107–1112.

Haller, O., and Kochs, G. (2002). Interferon-Induced Mx Proteins: Dynamin-Like GTPases with Antiviral Activity. Traffic *3*, 710–717.

Hansen, M.M., Taggart, J.B., and Meldrup, D. (1999). Development of new VNTR markers for pike and assessment of variability at di- and tetranucleotide repeat microsatellite loci. Journal of Fish Biology *55*, 183–188.

Harvey, B. (2009). A Biological Synopsis of Northern Pike (Esox Lucius). Canadian Manuscript Report of Fisheries and Aquatic Sciences 2885 *v + 31p.*, 31.

Hattori, R.S., Murai, Y., Oura, M., Masuda, S., Majhi, S.K., Sakamoto, T., Fernandino, J.I., Somoza, G.M., Yokota, M., and Strüssmann, C.A. (2012). A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. PNAS *109*, 2955–2959.

Healy, J.A., and Mulcahy, M.F. (1980). A biochemical genetic analysis of populations of the northern pike, Esox lucius L., from Europe and North America. Journal of Fish Biology *17*.

Heule, C., Salzburger, W., and Böhne, A. (2014). Genetics of sexual development: an evolutionary playground for fish. Genetics *196*, 579–591.

Höglund, J. (2009). Evolutionary conservation genetics (Oxford University Press).

Holleley, C.E., O'Meally, D., Sarre, S.D., Marshall Graves, J.A., Ezaz, T., Matsubara, K., Azad, B., Zhang, X., and Georges, A. (2015). Sex reversal triggers the rapid transition from genetic to temperature-dependent sex. Nature *523*, 79–82.

Huffman, K., Farrell, J., and Whipps, C. (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. p.

Ilardi, J.M., Mochida, S., and Sheng, Z.-H. (1999). Snapin: a SNARE–associated protein implicated in synaptic transmission. Nature Neuroscience *2*, 119.

Ishiguro, N.B., Miya, M., and Nishida, M. (2003). Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii." Molecular Phylogenetics and Evolution *27*, 476–488.

Izumikawa, K., Ishikawa, H., Simpson, R.J., and Takahashi, N. (2018). Modulating the expression of Chtop, a versatile regulator of gene-specific transcription and mRNA export. RNA Biology *15*, 849–855.

Jacobsen, B.H., Hansen, M.M., and Loeschcke, V. (2005). Microsatellite DNA analysis of northern pike (Esox lucius L.) populations: insights into the genetic structure and demographic history of a genetically depauperate species. Biological Journal of the Linnean Society *84*, 91–101.

Jensen, M.P., Allen, C.D., Eguchi, T., Bell, I.P., LaCasella, E.L., Hilton, W.A., Hof, C.A.M., and Dutton, P.H. (2018). Environmental Warming and Feminization of One of the Largest Sea Turtle Populations in the World. Current Biology *28*, 154-159.e4.

Joberty, G., Petersen, C., Gao, L., and Macara, I.G. (2000). The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42. Nat. Cell Biol. *2*, 531–539.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics *24*, 1403–1405.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics *11*, 94.

Jones, T.R., Roberts, W.H.G., Steig, E.J., Cuffey, K.M., Markle, B.R., and White, J.W.C. (2018). Southern Hemisphere climate variability forced by Northern Hemisphere ice-sheet topography. Nature *554*, 351.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., et al. (2012). A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). PLOS Genetics *8*, e1002798.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., et al. A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, Takifugu rubripes (fugu). PLoS Genetics *8*, e1002798.

Kawase, J., Aoki, J., Hamada, K., Ozaki, A., and Araki, K. (2018). Identification of Sex-associated SNPs of Greater Amberjack (Seriola dumerili). J Genomics *6*, 53–62.

Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Res *12*, 656–664.

Khazaei, M.R., and Püschel, A.W. (2009). Phosphorylation of the par polarity complex protein Par3 at serine 962 is mediated by aurora a and regulates its function in neuronal polarity. J. Biol. Chem. *284*, 33571–33579.

Koed, A., Balleby, K., Mejlhede, P., and Aarestrup, K. (2006). Annual movement of adult pike (Esox lucius L.) in a lowland river. Ecology of Freshwater Fish *15*.

Koop, B. (2018). Salmonid polymorphism. Unpublished data.

Koop, B.F., von Schalburg, K.R., Leong, J., Walker, N., Lieph, R., Cooper, G.A., Robb, A., Beetz-Sargent, M., Holt, R.A., Moore, R., et al. (2008). A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. BMC Genomics *9*, 545.

Koyama, T., Nakamoto, M., Morishima, K., Yamashita, R., Yamashita, T., Sasaki, K., Kuruma, Y., Mizuno, N., Suzuki, M., Okada, Y., et al. (2019). A SNP in a Steroidogenic Enzyme Is Associated with Phenotypic Sex in Seriola Fishes. Current Biology *29*, 1901-1909.e8.

Lande, R. (1988). Genetics and demography in biological conservation. Science *241*, 1455–1460.

Larsen, P.F., Hansen, M.M., Nielsen, E.E., Jensen, L.F., and Loeschcke, V. (2005). Stocking impact and temporal stability of genetic composition in a brackish northern pike population ( Esox lucius L.), assessed using microsatellite DNA analysis of historical and contemporary samples. Heredity *95*, 136–143.

Lee, T.-H., Guo, H., Wang, X., Kim, C., and Paterson, A.H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics *15*, 162.

Leong, J.S., Jantzen, S.G., von Schalburg, K.R., Cooper, G.A., Messmer, A.M., Liao, N.Y., Munro, S., Moore, R., Holt, R.A., Jones, S.J., 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*, 279.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, Y.-H., and Wang, H.-P. (2017). Advances of genotyping-by-sequencing in fisheries and aquaculture. Rev Fish Biol Fisheries *27*, 535–559.

Li, C.-Y., Liu, Q.-R., Zhang, P.-W., Li, X.-M., Wei, L., and Uhl, G.R. (2009a). OKCAM: an ontology-based, human-centered knowledgebase for cell adhesion molecules. Nucleic Acids Res *37*, D251–D260.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009b). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, M., Sun, Y., Zhao, J., Shi, H., Zeng, S., Ye, K., Jiang, D., Zhou, L., Sun, L., Tao, W., et al. (2015). A Tandem Duplicate of Anti-Müllerian Hormone with a Missense SNP on the Y Chromosome Is Essential for Male Sex Determination in Nile Tilapia, Oreochromis niloticus. PLOS Genetics *11*, e1005678.

Lucentini, L., Puletti, M.E., Ricciolini, C., Gigliarelli, L., Fontaneto, D., Lanfaloni, L., Bilò, F., Natali, M., and Panara, F. (2011). Molecular and Phenotypic Evidence of a New Species of Genus Esox (Esocidae, Esociformes, Actinopterygii): The Southern Pike, Esox flaviae. PLOS ONE *6*, e25218.

Luczynski, M., Glogowski, J., Kucharczyk, D., Łuczyński, M., and Demska-Zakes, K. (1997). Gynogenesis in northern pike (Esox lucius L.) induced by heat shock - Preliminary data. Polskie Archiwum Hydrobiologii *44*, 25–32.

Lumeng, C., Phelps, S., Crawford, G.E., Walden, P.D., Barald, K., and Chamberlain, J.S. (1999). Interactions between β2-syntrophin and a family of microtubule-associated serine/threonine kinases. Nature Neuroscience *2*, 611–617.

Macqueen, D.J., and Johnston, I.A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc. R. Soc. B *281*, 20132881.

Mank, J.E., Promislow, D.E.L., and Avise, J.C. (2006). Evolution of alternative sex-determining mechanisms in teleost fishes. Biological Journal of the Linnean Society *87*, 83–93.

Marchenko, G.N., Marchenko, N.D., and Strongin, A.Y. (2003). The structure and regulation of the human and mouse matrix metalloproteinase-21 gene and protein. Biochemical Journal *372*, 503–515.

Marsden, J.E., Kassler, T., and Philipp, D. (1995). Allozyme Confirmation That North American Yellow Perch (Perca flavescens) and Eurasian Yellow Perch (Perca fluviatilis) Are Separate Species. Copeia *1995*, 977–981.

Marshall, C., and Wund, M. (2017). The evolution of correlations between behavioural and morphological defence in Alaskan threespine stickleback fish (Gasterosteus aculeatus): Evidence for trait compensation and co-specialization. Evolutionary Ecology Research *18*, 305–322.

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., et al. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. ELife *5*, e12081.

Massey, M.D., Holt, S.M., Brooks, R.J., and Rollinson, N. (2019). Measurement and modelling of primary sex ratios for species with temperature-dependent sex determination. Journal of Experimental Biology *222*, jeb190215.

Matsuda, M. (2018). Genetic Control of Sex Determination and Differentiation in Fish. In Reproductive and Developmental Strategies: The Continuity of Life, K. Kobayashi, T. Kitano, Y. Iwao, and M. Kondo, eds. (Tokyo: Springer Japan), pp. 289–306.

Matsuda, M., Nagahama, Y., Shinomiya, A., Sato, T., Matsuda, C., Kobayashi, T., Morrey, C.E., Shibata, N., Asakawa, S., Shimizu, N., et al. (2002). DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature *417*, 559.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mehaffey, K.C. (2018). DNA Shows Northern Pike Likely Planted Above Lake Roosevelt. NW Fishletter.

Merola, M. (1994). A Reassessment of Homozygosity and the Case for Inbreeding Depression in the Cheetah, Acinonyx jubatus: Implications for Conservation. Conservation Biology *8*, 961–971.

Miller, L.M., and Kapuscinski, A.R. (1996). Notes: Microsatellite DNA Markers Reveal New Levels of Genetic Variation in Northern Pike. Transactions of the American Fisheries Society *125*, 971–977.

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

Miller, L.M., and Senanan, W. (2003). A Review of Northern Pike Population Genetics Research andIts Implications for Management. North American Journal of Fisheries Management *23*, 297–306.

Milot, E., Weimerskirch, H., Duchesne, P., and Bernatchez, L. (2007). Surviving with low genetic diversity: the case of albatrosses. Proc Biol Sci *274*, 779–787.

Mohanty-Hejmadi, P., Dutta, S.K., Dey, D., and Rath, D.P. (1999). Temperature-dependent sex determination in the salt-water crocodile, Crocodylus porosus Schneider. Current Science *76*, 695–696.

Morikawa, M., Derynck, R., and Miyazono, K. (2016). TGF-β and the TGF-β Family: Context-Dependent Roles in Cell and Tissue Physiology. Cold Spring Harb Perspect Biol *8*, a021873.

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., and Sakaizumi, M. (2012). Tracing the emergence of a novel sex-determining gene in medaka, Oryzias luzonensis. Genetics *191*, 163–170.

Nater, A., Mattle-Greminger, M.P., Nurcahyo, A., Nowak, M.G., de Manuel, M., Desai, T., Groves, C., Pybus, M., Sonay, T.B., Roos, C., et al. (2017). Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. Current Biology *27*, 3487-3498.e10.

Navara, K.J. (2018). Choosing Sexes: Mechanisms and Adaptive Patterns of Sex Allocation in Vertebrates (Springer International Publishing).

Nelson, T.C., Crandall, J.G., Ituarte, C.M., Catchen, J.M., and Cresko, W.A. (2019). Selection, Linkage, and Population Structure Interact To Shape Genetic Variation Among Threespine Stickleback Genomes. Genetics genetics.302261.2019.

nterhoeven (2017). convert a blast output to a bed file. Contribute to nterhoeven/blast2bed development by creating an account on GitHub.

Ouellet-Cauchon, G., Normandeau, E., Mingelbier, M., and Bernatchez, L. (2014). EST-based microsatellites for northern pike (Esox lucius) and cross-amplification across all Esox species. Conservation Genetics Resources *2*, 451–454.

Pan, Q. (2017). To use or not use a Master sex determining gene: evolution of sex determination system in the Esociformes. Doctoral dissertation. Bretagne Loire University.

Pan, Q., Feron, R., Yano, A., Guyomard, R., Jouanno, E., Vigouroux, E., Wen, M., Busnel, J.-M., Bobe, J., Concordet, J.-P., et al. (2019). Identification of the master sex determining gene in Northern pike (Esox lucius) reveals restricted sex chromosome differentiation. BioRxiv 549527.

Park, M., Li, Q., Shcheynikov, N., Zeng, W., and Muallem, S. (2004). NaBC1 Is a Ubiquitous Electrogenic Na+-Coupled Borate Transporter Essential for Cellular Boron Homeostasis and Cell Growth and Proliferation. Molecular Cell *16*, 331–341.

Penn, D.J., Damjanovich, K., and Potts, W.K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. PNAS *99*, 11260–11264.

Pfennig, F., Standke, A., and Gutzeit, H.O. (2015). The role of Amh signaling in teleost fish – Multiple functions not restricted to the gonads. General and Comparative Endocrinology *223*, 87–107.

Priegel, G.R., and Krohn, D.C. (1975). Characteristics of a northern pike spawning population (Wisconsin Department of Natural Resources).

Purcell, C.M., Seetharam, A.S., Snodgrass, O., Ortega-García, S., Hyde, J.R., and Severin, A.J. (2018). Insights into teleost sex determination from the Seriola dorsalis genome assembly. BMC Genomics *19*, 31.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet *81*, 559–575.

Qin, B.Y., Liu, C., Lam, S.S., Srinath, H., Delston, R., Correia, J.J., Derynck, R., and Lin, K. (2003). Crystal structure of IRF-3 reveals mechanism of autoinhibition and virus-induced phosphoactivation. Nature Structural & Molecular Biology *10*, 913.

R Core Team (2018). R: A language and environment for statistical computing. https://www.r-project.org/.

Rambaut, A. (2007). FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

Ramírez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F., and Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics *179*, 555–567.

Ramsden, S.D., Brinkmann, H., Hawryshyn, C.W., and Taylor, J.S. (2003). Mitogenomics and the sister of Salmonidae. Trends in Ecology & Evolution *18*, 607–610.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Reed, D.H., and Frankham, R. (2003). Correlation between Fitness and Genetic Diversity. Conservation Biology *17*, 230–237.

Reimchen, T. (1994). Predators and morphological evolution in threespine stickleback. In The Evolutionary Biology of the Threespine Stickleback, M. A. Bell, and S.A. Foster, eds. (New York, NY: Oxford University Press), pp. 240–276.

Robinson, J.A., Vecchyo, D.O.-D., Fan, Z., Kim, B.Y., vonHoldt, B.M., Marsden, C.D., Lohmueller, K.E., and Wayne, R.K. (2016). Genomic Flatlining in the Endangered Island Fox. Current Biology *26*, 1183–1189.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer.

Rondeau, E.B., Messmer, A.M., Sanderson, D.S., Jantzen, S.G., von Schalburg, K.R., Minkley, D.R., Leong, J.S., Macdonald, G.M., Davidsen, A.E., Parker, W.A., et al. (2013). Genomics of sablefish (Anoplopoma fimbria): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. BMC Genomics *14*, 452.

Rondeau, E.B., Minkley, D.R., Leong, J.S., Messmer, A.M., Jantzen, J.R., von Schalburg, K.R., Lemon, C., Bird, N.H., and Koop, B.F. (2014). The Genome and Linkage Map of the Northern Pike (Esox lucius): Conserved Synteny Revealed between the Salmonid Sister Group and the Neoteleostei. PLoS One *9*, e102089.

Ryff, J.-C., and Pestka, S. (2013). Interferons and Interleukins. In Pharmaceutical Biotechnology: Fundamentals and Applications, D.J.A. Crommelin, R.D. Sindelar, and B. Meibohm, eds. (New York, NY: Springer New York), pp. 413–437.

Sealy, J., Lee-Thorp, J., Loftus, E., Faith, J.T., and Marean, C.W. (2016). Late Quaternary environmental change in the Southern Cape, South Africa, from stable carbon and oxygen isotopes in faunal tooth enamel from Boomplaas Cave. Journal of Quaternary Science *31*, e2916.

Seeb, J.E., Seeb, L.W., Oates, D.W., and Utter, F.M. (1987). Genetic Variation and Postglacial Dispersal of Populations of Northern Pike (Esox lucius) in North America. Can. J. Fish. Aquat. Sci. *44*, 556–561.

Sellis, D., Callahan, B.J., Petrov, D.A., and Messer, P.W. (2011). Heterozygote advantage as a natural consequence of adaptation in diploids. PNAS *108*, 20666–20671.

Senanan, W., and Kapuscinski, A.R. (2000). Genetic relationships among populations of northern pike (Esox lucius). Canadian Journal of Fisheries and Aquatic Sciences; Ottawa *57*, 391–404.

Senay, C., Harvey-Lavoie, S., Macnaughton, C. j., Bourque, G., and Boisclair, D. (2017). Morphological differentiation in northern pike (Esox lucius): the influence of environmental conditions and sex on body shape. Can. J. Zool. *95*, 383–391.

Shi, G.-X., Jin, L., and Andres, D.A. (2011). A Rit GTPase-p38 Mitogen-Activated Protein Kinase Survival Pathway Confers Resistance to Cellular Stress. Molecular and Cellular Biology *31*, 1938–1948.

Simonsen, K.L., Churchill, G.A., and Aquadro, C.F. (1995). Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. Genetics *141*, 413–429.

Skog, A., Vøllestad, L.A., Stenseth, N.C., Kasumyan, A., and Jakobsen, K.S. (2014). Circumpolar phylogeography of the northern pike (Esox lucius) and its relationship to the Amur pike (E. reichertii). Frontiers in Zoology *11*, 67.

Skov, C., and Nilsson, P.A. (2018). Biology and ecology of pike (Boca Raton: CRC Press).

Skov, C., Koed, A., Baastrup-Spohr, L., and Arlinghaus, R. (2011). Dispersal, Growth, and Diet of Stocked and Wild Northern Pike Fry in a Shallow Natural Lake, with Implications for the Management of Stocking Programs. North American Journal of Fisheries Management *31*, 1177–1186.

Smith, P.J., and Fujio, Y. (1982). Genetic variation in marine teleosts: High variability in habitat specialists and low variability in habitat generalists. Mar. Biol. *69*, 7–20.

Spidle, A.P., King, T.L., and Letcher, B.H. (2004). Comparison of genetic diversity in the recently founded Connecticut River Atlantic salmon population to that of its primary donor stock, Maine's Penobscot River. Aquaculture *236*, 253–265.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. Nature *477*, 207–210.

Star, B., Tørresen, O.K., Nederbragt, A.J., Jakobsen, K.S., Pampoulie, C., and Jentoft, S. (2016). Genomic characterization of the Atlantic cod sex-locus. Scientific Reports *6*, srep31235.

Su, J., Sui, Y., Ding, J., Li, F., Shen, S., Yang, Y., Lu, Z., Wang, F., Cao, L., Liu, X., et al. (2016). Human INO80/YY1 chromatin remodeling complex transcriptionally regulates the BRCA2- and CDKN1A-interacting protein (BCCIP) in cells. Protein Cell *7*, 749–760.

Sun, F., Zhang, Y.-B., Liu, T.-K., Gan, L., Yu, F.-F., Liu, Y., and Gui, J.-F. (2010). Characterization of Fish IRF3 as an IFN-Inducible Protein Reveals Evolving Regulation of IFN Response in Vertebrates. The Journal of Immunology *185*, 7573–7582.

Suzuki, T., Mizusaki, H., Kawabe, K., Kasahara, M., Yoshioka, H., and Morohashi, K.-I. (2002). Concerted Regulation of Gonad Differentiation by Transcription Factors and Growth Factors. In The Genetics and Biology of Sex Determination, D.C. Organizer, and J. Goode, eds. (John Wiley & Sons, Ltd), pp. 68–78.

Swain, A., and Lovell-Badge, R. (1999). Mammalian sex determination: a molecular drama. Genes Dev. *13*, 755–767.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics *123*, 585–595.

Takehana, Y., Matsuda, M., Myosho, T., Suster, M.L., Kawakami, K., Shin-I, T., Kohara, Y., Kuroki, Y., Toyoda, A., Fujiyama, A., et al. (2014). Co-option of *Sox3* as the male-determining factor on the Y chromosome in the fish *Oryzias dancena*. Nature Communications *5*, 4157.

the 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Thoms, S., and Erdmann, R. (2005). Dynamin-related proteins and Pex11 proteins in peroxisome division and proliferation. The FEBS Journal *272*, 5169–5181.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178–192.

Turner, S. (2017). qqman: Q-Q and Manhattan Plots for GWAS Data.

Valenzuela, N., and Lance, V. (2004). Temperature Dependent Sex Determination in Vertebrates.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics *43*, 11.10.1-33.

Vashro, Jim (2018). Water Wolves. Montana Outdoors 37–39.

Wang, C., Wang, Y., Hu, M., Chai, Z., Wu, Q., Huang, R., Han, W., Zhang, C.X., and Zhou, Z. (2016). Synaptotagmin-11 inhibits clathrin-mediated and bulk endocytosis. EMBO Rep. *17*, 47–63.

Wang, C., Kang, X., Zhou, L., Chai, Z., Wu, Q., Huang, R., Xu, H., Hu, M., Sun, X., Sun, S., et al. (2018). Synaptotagmin-11 is a critical mediator of parkin-linked neurotoxicity and Parkinson's disease-like pathology. Nature Communications *9*, 81.

Wang, J., Wang, C., Qian, L., Ma, Y., Yang, X., Jeney, Z., and Li, S. (2011). Genetic characterization of 18 novel microsatellite loci in northern pike (Esox lucius L.). Genetics and Molecular Biology *34*, 169–172.

Wayne, M.L., and Miyamota, M.M. (2006). Evolutionary genetics: concepts and case studies (Oxford, UK ; New York: Oxford University Press).

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

Whitehead, J. (2019). Esox lucius k-mer analysis. Unpublished report.

Wilson, M.V.H. (1980). Oldest known Esox (Pisces: Esocidae), part of a new Paleocene teleost fauna from western Canada. Can. J. Earth Sci. *17*, 307–312.

Wilson, M.V.H., Brinkman, D.B., and Neuman, A.G. (1992). Cretaceous Esocoidei (Teleostei): Early Radiation of the Pikes in North American Fresh Waters. Journal of Paleontology *66*, 839–846.

Wu, G.-C., and Chang, C.-F. (2013). The switch of secondary sex determination in protandrous black porgy, Acanthopagrus schlegeli. Fish Physiol Biochem *39*, 33–38.

Yano, A., Guyomard, R., Nicol, B., Jouanno, E., Quillet, E., Klopp, C., Cabau, C., Bouchez, O., Fostier, A., and Guiguen, Y. (2012). An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, Oncorhynchus mykiss. Current Biology *22*, 1423–1428.

Yano, A., Nicol, B., Jouanno, E., Quillet, E., Fostier, A., Guyomard, R., and Guiguen, Y. (2013). The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. Evol Appl *6*, 486–496.

Zeileis, A., and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software *14*, 1–27.

Zhao, M., Mishra, L., and Deng, C.-X. (2018). The role of TGF-β/SMAD4 signaling in cancer. Int J Biol Sci *14*, 111–123.

Zheng, J., Jia, Y., Liu, S., Jiang, W., Chi, M., Cheng, S., and Gu, Z. (2018). Transcriptome analysis of Culter alburnus gonad tissues for discovery of sex-related genes. BioRxiv 351759.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

(2018). RTG Tools: Utilities for accurate VCF comparison and manipulation: RealTimeGenomics/rtg-tools (Real Time Genomics).

# Glossary

| Term | Definition |
|---|---|
| Allele | Gene copies that are slightly different from one another due to mutations in DNA sequence. In a population there may be many different alleles for a particular gene. Within a diploid individual, there are two alleles, which may be the same or different. |
| BAM File | "Binary Alignment Map". The binary form of a SAM file. |
| Bonferroni correction | A correction for multiple testing where the set significance level is divided by the number of tests. |
| Fisher strand bias | A SNP filtering technique used in GATK that tests if a SNP is biased on the positive or negative strand. |
| GATK | "Genome Analysis Tool Kit". A bioinformatics pipeline designed to process resequencing and other data for the purpose of SNP discovery. |
| Locus | A distinct position on a chromosome where a genetic feature (e.g. gene, marker, SNP) is located. Plural is loci. |
| Mapping quality rank sum test | Tests if the alternate allele has higher or lower mapping quality scores than the reference allele. Ideally the test result is 0, indicating no bias. Informs SNP filtering in GATK. |
| Master sex determining gene | MSD gene. A gene that initiates processes leading to sex determination. |
| Maximum mean depth | A filtering parameter in VCFtools that removes SNPs whose average depth is above the specified value. |
| Minor allele frequency | Frequency of the second most common allele. |
| Minimum mean depth | A filtering parameter in VCFtools that removes SNPs whose average depth is below the specified value. |
| Minimum quality | A filtering parameter in VCFtools that removes SNPs with a quality score below the specified file |

| | |
|---|---|
| Quality by depth | A SNP filtering technique used in GATK that judges the validity of a SNP based on assigned quality score divided by the allele depth. |
| Resequencing | Sequencing part or all of an organism's genome for comparison to a reference genome in order to identify differences (SNPs). |
| Read position rank sum test | Tests if the alternate allele is positioned at the distal ends of reads more frequently than the reference allele. Ideally the test result is 0, indicating no bias. Informs SNP filtering in GATK. |
| Root mean square mapping quality | A SNP filtering technique used in GATK that judges the validity a SNP based on the average mapping quality plus the standard deviation of the mapping qualities. |
| SAM file | "Sequence Alignment Map". A standardized text file format that stores the genomic location of where a sequence read maps to a specified reference genome. |
| Sex determination | The moment in embryonic development when a cue (environmental and/or genetic) causes gonadal tissue to differentiate into ovaries (female) or testes (male). |
| Single nucleotide polymorphism | A single nucleotide polymorphism (SNP) is a mutation in DNA that causes nucleotide sequences between individuals to differ from each other. In diploid species, a SNP can be present in heterozygous (e.g A/G) or homozygous (e.g. A/A or G/G) forms. Homozygous SNPs are detectable when multiple individuals are examined. |
| VCF | "Variant Call File". A standardized text file format that stores SNP data for all samples, including genomic location, genotype, and supporting score data. Additional meta-information is detailed in the header lines. |
| Yukon Drainage Basin | Large area covering a portion of Northwestern B.C., Yukon Territory, and Alaska that is drained by the Yukon River, flowing east to west and into the Bering Sea. |