Conserved synteny in the genomes of teleost fish aids in the rapid development of genomic tools to query fundamental biological and evolutionary questions.

by

Eric Rondeau
Bachelor of Science, University of Victoria, 2009

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

Doctor of Philosophy

in the Department of Biology

© Eric Rondeau, 2017
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy
or other means, without the permission of the author.

Supervisory Committee

Conserved synteny in the genomes of teleost fish aids in the rapid development of genomic tools to query fundamental biological and evolutionary questions.

by

Eric Rondeau
Bachelor of Science, University of Victoria, 2009

Supervisory Committee

Dr. Ben F. Koop, Department of Biology
Supervisor

Dr. Steve J. Perlman, Department of Biology
Departmental Member

Dr. R. John Nelson, Department of Biology
Departmental Member

Dr. Stephen F. Cross, Department of Geography
Outside Member

Abstract

Supervisory Committee
Dr. Ben F. Koop, Department of Biology
Supervisor
Dr. Steve J. Perlman, Department of Biology
Departmental Member
Dr. R. John Nelson, Department of Biology
Departmental Member
Dr. Stephen F. Cross, Department of Geography
Outside Member

As two species diverge, much of their genomes begin to differentiate. In many lineages, however, the genomic structure remains remarkably intact, with orthologous gene content maintained across millions of years and significant changes to their biological characteristics. The maintenance of gene content is defined as conserved synteny while the preservation of gene order is defined as conserved linkage; the conservation of both can be incredibly informative when interrogating and comparing two genomes. In non-model organisms, linkage conservation to a well-developed model allows informed, cost-effective and rapid answers to fundamental biological questions without generation of equivalent resources. With the development of new model organisms, we can begin to discuss more fundamental evolutionary concepts, such as the maintenance of chromosomal gene content across larger evolutionary time-scales, or the reorganization that occurs in chromosomes following major genomic events such as whole-genome duplications. In this work, I utilized the rapid development of primary genomic resources in the non-model teleost sablefish (*Anoplopoma fimbria*) to demonstrate that conserved linkage to a model genomic reference can identify the gene most likely responsible for genetic sex-control. I then assembled the first genome for a

non-duplicated member of the teleost lineage Protacanthopterygii, the northern pike (*Esox lucius*), and demonstrated the conservation of synteny between three major lineages of teleosts, the Protacanthopterygii, the Acanthopterygii and the Ostariophysi. I further showed that the genome of northern pike retains an ancestral teleost organization and pre-duplicated genome in comparison to the economically important Salmoniformes. Finally, with continued improvements of the genome to the chromosome level, I demonstrated the degree of conserved linkage maintained between Atlantic salmon and northern pike and explained how conserved linkage through both genomes could be used to improve the genome assembly of the other, even with over 125 million years of separation. As genomic technology continues to advance and new genomic resources become available, the continued refinement of genome re-organization post duplication will be revealed, and this pre-duplication outgroup will continue to push our understanding of the effects of genome duplication, as we transition from genome organization to functional modifications of gene duplicates following duplication.

Table of Contents

Supervisory Committee ii
Abstract iii
Table of Contents v
List of Tables ix
List of Figures x
Acknowledgments xii
Dedication xiii
Chapter 1 : Introduction 1
Chapter 2 : Genomics of Sablefish (*Anoplopoma fimbria*): Expressed genes, mitochondrial phylogeny, linkage map and identification of genetic sex markers 19
Abstract 19
Background 19
Results 19
Conclusions 20
Background 20
Results and Discussion 24
Gene Identification 24
The mitochondrial genome and phylogenetic placement of sablefish 24
Identification of Genetic Markers 27
Production of first-generation linkage map 28
Significantly conserved synteny observed between *A. fimbria* and *G. aculeatus* 30
Additional markers by Genotyping-by-sequencing (GBS) 36
Genetic identification of sex in sablefish 37
Significance of the Genomic Tools 44
Conclusions 45
Methods 45
Sample Collection and extractions 45

cDNA libraries 47
Mitochondrial Genome Sequencing 47
Primer Development and Marker Identification - SNPs 50
Primer Development and Marker Identification – Microsatellites 50
Marker Naming 51
Linkage Analysis 52
Anoplopoma fimbria vs *Gasterosteus aculeatus* (threespine stickleback) 52
Genotyping-by-sequencing (GBS) Library 53
Genome Walking 53
Identification of Sex-specific markers and sequences 54
Acknowledgements 55
Supporting Data 55
Chapter 3 : The genome and linkage map of the northern pike (*Esox lucius*): Conserved synteny revealed between the salmonid sister group and the Neoteleostei 56
Abstract 56
Background 57
Results and Discussion 62
A high-quality genome assembly for *Esox lucius* 62
Transcriptome assembly and Gene Identification 65
A snapshot of tissue-specific gene expression 66
E. lucius homologues suggest a pre-duplication genome 68
Restriction-site associated DNA (RAD) tag based linkage mapping inefficient in pike 72
A first-generation, Microsatellite-based linkage map 73
Synteny analyses suggest Neoteleostei-like genome in northern pike 79
The usefulness of salmonid duplication analysis 85
Conclusions 88
Methods and Materials 89

Animal Care 89
Whole-Genome Sequencing 89
Transposable element library creation and genomic repeat annotation 90
RNA-seq and assembly 91
RAD-tag based linkage mapping 93
Microsatellite Primer Design 93
Linkage Mapping 95
Synteny analyses 96
Acknowledgements 96
Supporting Data 97
Chapter 4 : A Chromosome-level assembly of the northern pike (*Esox lucius*) genome traces the evolution of the salmonids following whole-genome duplication 98
Abstract 98
Background 99
Results and Discussion 102
Esox lucius Genome assembly 102
Synteny in Teleostei 104
Synteny with the Atlantic salmon: 111
Guiding assemblies 112
Conclusions 117
Methods 118
Sequencing and Assembly 118
Assembly into Chromosomes: 120
Genome Alignments 120
Acknowledgements 121
Supporting data 121
Chapter 5 : Discussion, Future Directions and Conclusions 122
Mapping methods 122

Continued progress with genomic tools developed for sablefish 124
Fate of gene duplicates – pike as a proxy for last common ancestor 126
Genome assembly Improvements – *Esox lucius* Chapter 3 to Chapter 4..... 131
Future Directions 133
Conclusions 137
Bibliography 139
Appendix A Publications and presentations during program..... 191
Appendix B Supplementary Files, Chapter 2..... 194
Appendix C Supplementary Files, Chapter 3..... 197
Appendix D Supplementary Files, Chapter 4 200

List of Tables

Table 2.1: Comparative synteny between threespine stickleback and sablefish 32

Table 3.1: *Esox lucius* genome assembly statistics. 63

Table 3.2: Sequences from mapped Atlantic salmon SNPs blasted against the *Esox lucius* genome (≤1e-10)..... 87

Table 4.1: Genome sequencing stats 103

Table 4.2: Conserved synteny between northern pike, threespine stickleback, Atlantic salmon and rainbow trout 106

List of Figures

Figure 1.1: Synteny and linkage 2
Figure 1.2: The basics of generating a linkage map 4
Figure 1.3: Basic steps in Genome assembly 8
Figure 1.4: Uses of conserved synteny and linkage 11
Figure 1.5: Phylogeny of whole-genome duplications in the teleost phylogeny 13
Figure 2.1: Mitogenome phylogeny of orders: Scorpaeniformes, Gasterosteiformes, select Perciformes 26
Figure 2.2: Conserved linkage between sablefish and threespine stickleback chromosomes I-VII 34
Figure 2.3: Conserved linkage between sablefish and threespine stickleback chromosomes VIII-XV 35
Figure 2.4: Conserved linkage between sablefish and threespine stickleback chromosomes XVI-XXI 36
Figure 2.5: Sex-specific sequences in sablefish 41
Figure 3.1: The phylogenetic relationship of teleosts with publicly available genome assemblies, as well as select species with ongoing genome projects (*) 58
Figure 3.2: The percent similarities identified through transcriptome blastn comparisons. 69
Figure 3.3: The northern pike merged linkage map: LG01 to LG10. 75
Figure 3.4: The northern pike merged linkage map: LG11 to LG20. 76
Figure 3.5: The northern pike merged linkage map: LG21 to LG25. 77
Figure 3.6: The synteny between the northern pike and model teleosts. 81
Figure 4.1: Threespine stickleback genomes aligned to northern pike. 107
Figure 4.2: Atlantic salmon genome aligned to northern pike. 108
Figure 4.3: Northern pike-specific sequence through mapping to Atlantic salmon and threespine stickleback 109

Figure 4.4: Threespine stickleback genomes aligned to northern pike – GacuXII vs Eluc12. 110
Figure 4.5: Adding additional scaffolds through comparative mapping – Eluc23 vs Ssal07q 113
Figure 4.6: Identifying mismapped scaffold candidates in Atlantic salmon – Eluc02 vs Ssal26 and Ssal11 116
Figure 4.7: “Pike-ized” regions of residual tetraploidy – Eluc02 vs Ssal26 117
Figure 5.1: Sex-specific PCR assays designed for lingcod and yelloweye rockfish using highly-conserved domains between threespine stickleback and sablefish. 126
Figure 5.2: Heatmap of 2,272 triplets (two salmon homeologues and a pike orthologue), in which one of the Atlantic salmon homeologues has diverged in gene expression regulation (Figure 4b, Lien et al. (Lien et al. 2016)). 130
Figure 5.3 Circos plot of homeologous regions in coho salmon 135

Acknowledgments

I would like to acknowledge all the help received during my time at the University of Victoria. My collaborators in Chapter 2, Amber Messmer, Dan Sanderson and Stuart Jantzen on their early work assisting with microsatellite development and EST sequencing, Kristian von Schalburg for all his help training me on the art of RNA extraction, David Minkley and Jong Leong for their support on various aspects of the Bioinformatic approaches used in this work, Graham Macdonald, Amanda Davidsen, William Parker and Rosetta Mazzola for their assistance in running and analyzing the sheer volume of genotyping produced in this work, my supervisor Dr. Ben Koop's continuous advice, and Dr. Briony Campbell and all the staff who may have assisted in the generation and care of the sablefish used in our work. My collaborators for Chapter 3, David Minkley and Jong Leong for their enormous help with much of the heavy lifting involved in the Bioinformatics, Amber Messmer, Johanna Jantzen and Nathan Bird with microsatellite primer development and genotyping, and Kristian von Schalburg in his continued support with anything RNA, and Craig Lemon and staff at the Hackettstown Fish Hatchery (NJ Fish and Wildlife) and staff at the Whiteshell Fish Hatchery (Whiteshell Provincial Park, Manitoba) and anyone who may have assisted in generation and care of the northern pike used in our work. In Chapter 4, David Minkley and Jong Leong were again instrumental in their assistance with the core bioinformatics. My time at the University of Victoria was further enriched by enormously helpful discussions and support from Dr. Ben Sutherland, Dr. Kris Christensen, Stephanie Puckett and Cody Despins, and the UVic molecular evolution discussion group including Marie Vance, Graeme Keais and Dr. David Marques who sparked a number of helpful conversations and support. To all others I may have forgotten, I apologize and thank you for your help profusely!

Dedication

For my loving wife, my wonderful family, my supportive lab and my patient supervisor.

Chapter 1 :

Introduction

With his work “On the Origin of Species” published in 1859, Charles Darwin popularized the idea of evolution through common descent (Darwin 1859). Common descent states that species evolve in a series of split branches and that all lineages are traced back to a shared common ancestor, rather than to independent origins. This concept may be best visually illustrated by the concept of a “tree of life” or evolutionary tree, whereby all branches eventually meet at the base of the tree, representing the last common ancestor. It has been theorized that all life on earth likely shares a common origin, or last universal common ancestor, with recent work identifying a set of core genes that defines this ancestor (Weiss et al. 2016).

The concept of conserved synteny and linkage can be traced as a natural extension of common descent. In a single species, the genes found on a single chromosome are said to be syntenic, with the origin of the word meaning “same thread” (Passarge et al. 1999) – that is, all genes on the same chromosome are syntenic, no matter how far apart they may be (figure 1.1a). Under this definition, synteny would only be broken by a chromosomal translocation or fission within the same species, such as the translocation occasionally observed in Down syndrome between human chromosomes 14 and 21 (Polani et al. 1960). However, as the field of genetics and genomics expanded, it became possible to compare genomes between species rather than just within species. At the point of speciation, two diverging lineages share a common set of chromosomes; between the two species, these chromosomes will at least initially share the same genes and the same chromosomal structure. This clustering of sequences on the same homologous chromosome between two species is the definition of conserved synteny, also known as shared synteny (Nadeau 1989) (figure 1.1b). While these blocks of syntenic sequences need not have a conserved order,

where they do it can be said that they have conserved linkage (Nadeau 1989) or collinearity of conserved gene order (figure 1.1c). While the terms conserved synteny and conserved linkage are often simply described as "synteny" in the literature, synteny as originally described by J.H. Renwick (Renwick 1971) refers exclusively to the presence of two genes on a single chromosome.

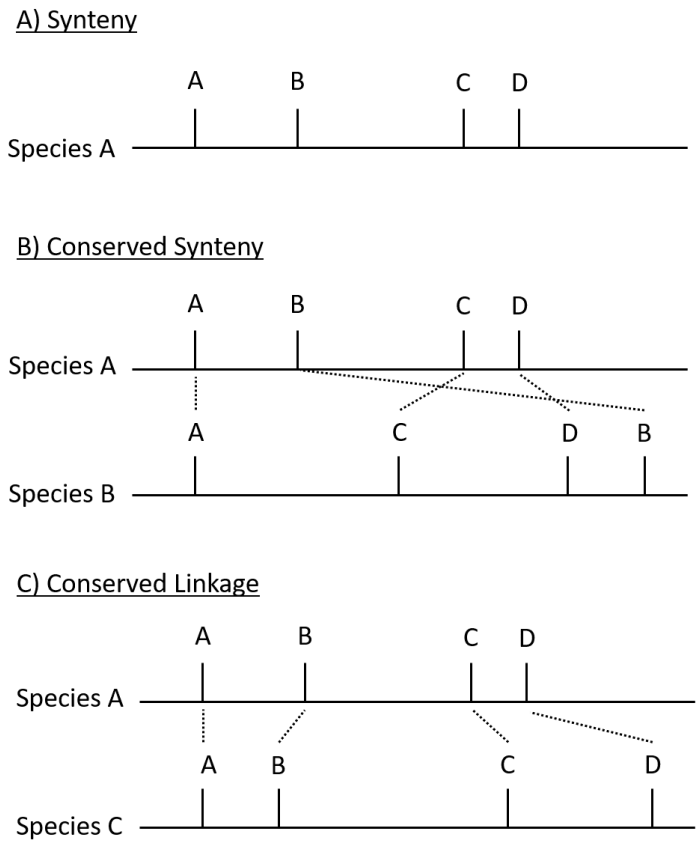


Figure 1.1: Synteny and linkage

Synteny (Panel A) refers specifically to a group of features (e.g. Genes) that are collected on a single chromosome, in a single species. Conserved synteny (Panel B) refers to a group of features on the same chromosome in two distinct species, regardless of order. Conserved linkage (Panel C) refers to a group of features on the same chromosome that maintains the same order in two distinct species. Conserved synteny and conserved linkage are identified through comparative mapping.

To distinguish between such situations, it must be possible to identify the position of at least two points within a chromosome. In the early 1900s, Thomas Hunt

Morgan and his students at Columbia University began to incorporate the ideas of inheritance and evolution into the emerging field of genetics. Working on the fruit fly *Drosophila melanogaster*, Morgan began cross-breeding flies to identify heritable mutations. In a fundamental paper in 1911 (Morgan 1911), following the identification of a series of five eye colour mutations, he was able to tie the inheritance of a particular eye colour allele with the sex of the fly. This demonstrated two seemingly independent traits that could co-segregate, breaking Gregor Mendel's second law of independent assortment. This concept became known as genetic linkage, with Morgan postulating that it was "due to an intimate physical relation between the …factors in question; and the most obvious connections is that the relation is to be found in the chromosomes that carry both the sex factor and those factors that are sex-limited" (Morgan 1911). In a subsequent manuscript in 1913 (Sturtevant 1913), Morgan's undergraduate student Alfred Sturtevant was able to order the genes along the chromosome by looking at the frequency with which linkage was broken in between adjacent traits on the sex chromosome of *D. melanogaster*. Sturtevant translated this frequency into a distance measurement, thereby creating the first linkage map. This distance measurement, the frequency of broken linkage between a pair of genes per one hundred offspring examined is called a centiMorgan (cM). Following up this work in "A Critique of the Theory of Evolution" (Morgan 1916), Morgan describes the theory of recombination, whereby paired chromosomes "cross-over" during meiosis, breaking the linkage between alleles. While these early experiments were based on observable phenotypic alleles (eye colour, wing morphology, etc.), the fundamental technique of generating a linkage map remained consistent as the field transitioned from classical to molecular genetics. While the number of markers and the statistical difficulty increased over time, the underlying method remains remarkably unchanged (figure 1.2).

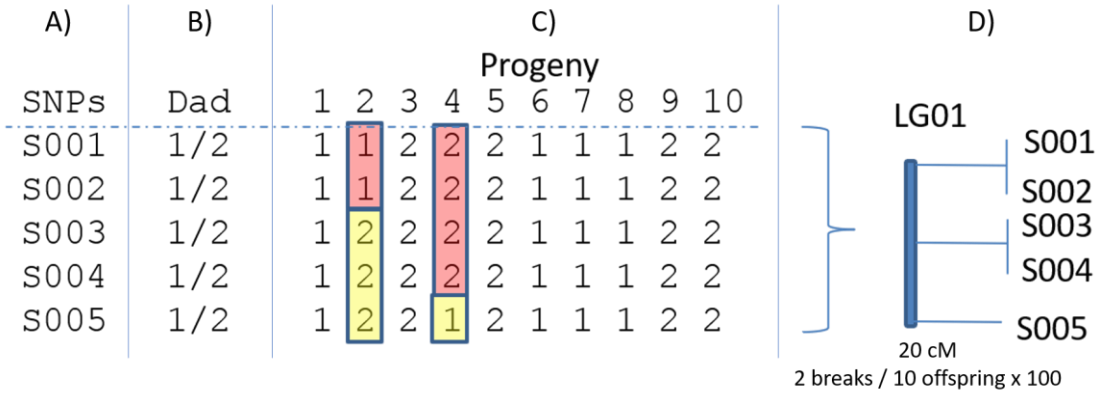


Figure 1.2: The basics of generating a linkage map

A simple example of linkage mapping. A) a series of markers, such as SNPs, are identified in a species. B) Markers are identified that are polymorphic in a parent, represented here by allele 1 and allele 2. C) The markers are genotyped in the offspring of the parent, determining which of the two alleles each received. The markers are sorted to maximize linkage between alleles in the offspring. Instances where the linkage is disrupted by recombination in the offspring are identified (red and yellow boxes). D) Markers are arranged on a map, with the length of cM based on recombination frequency.

As phenotypic attributes have given way to genotypic markers in the generation of modern genetic maps, the importance of identifying and genotyping genetic polymorphism has become particularly important. Large-scale success came with the development of the polymerase-chain-reaction (PCR) that enabled the amplification of defined segments of DNA using short, highly specific nucleotide primers (Mullis et al. 1986). Microsatellite markers, made up of short, tandem 2-5 base pair repeats found throughout the eukaryotic genome (Tautz and Renz 1984), could be amplified by PCR (Litt and Luty 1989, Tautz 1989, Weber and May 1989) and genotyped by gel or capillary electrophoresis (Butler et al. 1994), providing a cheap, abundant and highly polymorphic source of genetic markers. In particular, the ability to use some microsatellites across closely related species (Schlotterer et al. 1991) allowed for the rapid development of linkage maps with shared markers for direct comparison. A rapid development of linkage maps using shared markers can be found in salmonid species (e.g. (Sakamoto et

al. 2000, Gilbey et al. 2004, Woram et al. 2004)), which allowed for early comparisons of conserved linkage within the Salmoniformes (Danzmann et al. 2005).

With the evolution of sequencing technologies, it became possible to harness the power of single nucleotide polymorphisms (SNPs), the most abundant markers in the genome. As SNPs are only bi- or occasionally tri-allelic markers and tend to be less polymorphic than microsatellites, the number of SNPs required to match the power of microsatellites may be more than ten-fold higher (Kalinowski 2002), and it is rare to find particular SNPs shared across distinct species (Leffler et al. 2013). SNP analysis had been confined to probe-binding assays, such as TaqMan (Holland et al. 1991), or direct analysis in capillary sequencing, which could be cost-prohibitive relative to microsatellites. While the development of SNP chips (Wang et al. 1998) allowed for more efficient analysis in model organisms, the true power of SNPs has only recently begun to emerge with advances in technology, primarily driven by Illumina-based short-read sequencing (Bentley et al. 2008). Reduced-representation methods such as RAD mapping (Baird et al. 2008, Amores et al. 2011, Hohenlohe et al. 2011) and Genotyping-By-Sequencing (Elshire et al. 2011), in which sequencing is focused on regions surrounding a restriction-enzyme cut site, allowed sequences across many individuals to be stacked (Catchen et al. 2011), with SNPs in stacks being simultaneously identified and genotyped. These techniques allowed a relatively consistent sampling of genome-wide polymorphism without prior knowledge of the genome.

Linkage mapping gives a framework organization to the genome, but it cannot, under most circumstances, fully resolve gene or sequence order in a genome because the fundamental measurement is the disruption of linkage by recombination. In general, there is an average of 1 recombination event per 100 offspring in 1 million base pairs (Mbp) in humans (Chakravarti 1991). To correctly order 20 genes within this 1 Mbp window would require analysing 2,000 full-siblings. Difficulties in calculating recombination frequencies can be further compounded by inconsistent recombination frequencies across the genome (Lichten and Goldman 1995) or between the sexes

(Lenormand and Dutheil 2005). While mixing multiple families or increasing marker density can improve mapping efficiency, the difficulty in identifying informative genetic markers shared between different families can make merging the data enormously difficult. Fortunately, the demonstration of conserved linkage to a model reference genome allows the approximation of gene order within the non-model organism, allowing the relatively significant cost of generating new genomes in a single species to be distributed among a wider number of species.

The release of full-genome assemblies has experienced near exponential growth over the past 50 years. While the smallest viral genomes could be sequenced as early as the mid-1970s (Fiers et al. 1976, Sanger et al. 1977), it was not until 1995 that the first bacterial genome could be fully sequenced (Fleischmann et al. 1995). Eukaryotic genome assemblies followed with increasingly complex organisms: yeast (Goffeau et al. 1996), nematode (The C. elegans sequencing consortium 1998), *Arabidopsis* (The Arabidopsis Genome Initiative 2000) and *Drosophila melanogaster* (Adams et al. 2000) all were released before the turn of the century. With vertebrate genomes being released through the beginning of the 2000s – human (Lander et al. 2001, Venter et al. 2001), mouse (Waterston et al. 2002), and the first fish, fugu (Aparicio et al. 2002) – biology truly entered the genomic age. While the complexity of generating these early assemblies was enormous, improvements to sequencing technologies, expansion of computing resources, and decreasing costs meant that the generation of new assemblies continued to expand. When I began my degree, there were just 28 eukaryotic genome assemblies available; as of October 7, 2017, there were 418 eukaryotic genome assemblies (including 47 teleost fish) in the publicly accessible RefSeq database of the National Center for Biotechnology Information (NCBI), with many more to be found in various repositories around the Internet. Such growth has been driven primarily by reduced sequencing costs and innovative technologies, which have brought many genome assembly projects from requiring large, well-funded consortiums to the level of the individual lab.

While the underlying technologies have changed, the fundamental process of generating a new sequence assembly has remained remarkably consistent. Chromosomes or other large fragments of DNA are extracted (figure 1.3.1) and broken randomly into smaller fragments (figure 1.3.2). Sequencing libraries are then generated (figure 1.3.3), depending on the sequencing technology used. In older genome assemblies, such libraries may have been generated and sequenced using the first widely adopted sequencing technology, Sanger sequencing (Sanger et al. 1977), often via clone-based libraries such as Bacterial Artificial Chromosomes (BACs) (e.g. (Kelley et al. 1999)). In 2017, such sequencing would be generated using single-molecule methods such as those commercialized by Pacific Biosciences (Eid et al. 2009) or Oxford Nanopore (Olasagasti et al. 2010), or by using a short-read platform such as Illumina (Bentley et al. 2008). An Illumina-based approach using paired-end and medium-insert mate-pair is presented in figure 1.3.3. An assembly into overlapping fragments would then be performed, generating contiguous assembled pieces of the genome known as contigs or unitigs (figure 1.3.4). In instances where libraries of physically-linked fragments were end-sequenced but the sequences did not overlap, the physical linkage can be used to order contigs into scaffolds while the fragment size can be used to determine the size of an approximate gap (figure 1.3.5). Once scaffolds are built, additional sequencing using single-molecule long-read technologies or directed methods such as chromosome-walking (Cole et al. 2008) can be used to fill in gaps (figure 1.3.6). The order and separation of steps represented in figure 1.3.4-1.3.6 are highly dependent on the sequencing technologies used as well as the choice of the assembly program. While some assembly programs will utilize multiple read types in generating a final assembly (e.g. (Miller et al. 2017)), others deal only with single data-types and are used in a step-wise approach (e.g. (Gnerre et al. 2011, English et al. 2012)). Ultimately, the choice of sequencing and assembly strategy will depend highly on the budget, the complexity of the genome, and the availability of closely related genomic resources for transitioning from assembled fragments to chromosomes.

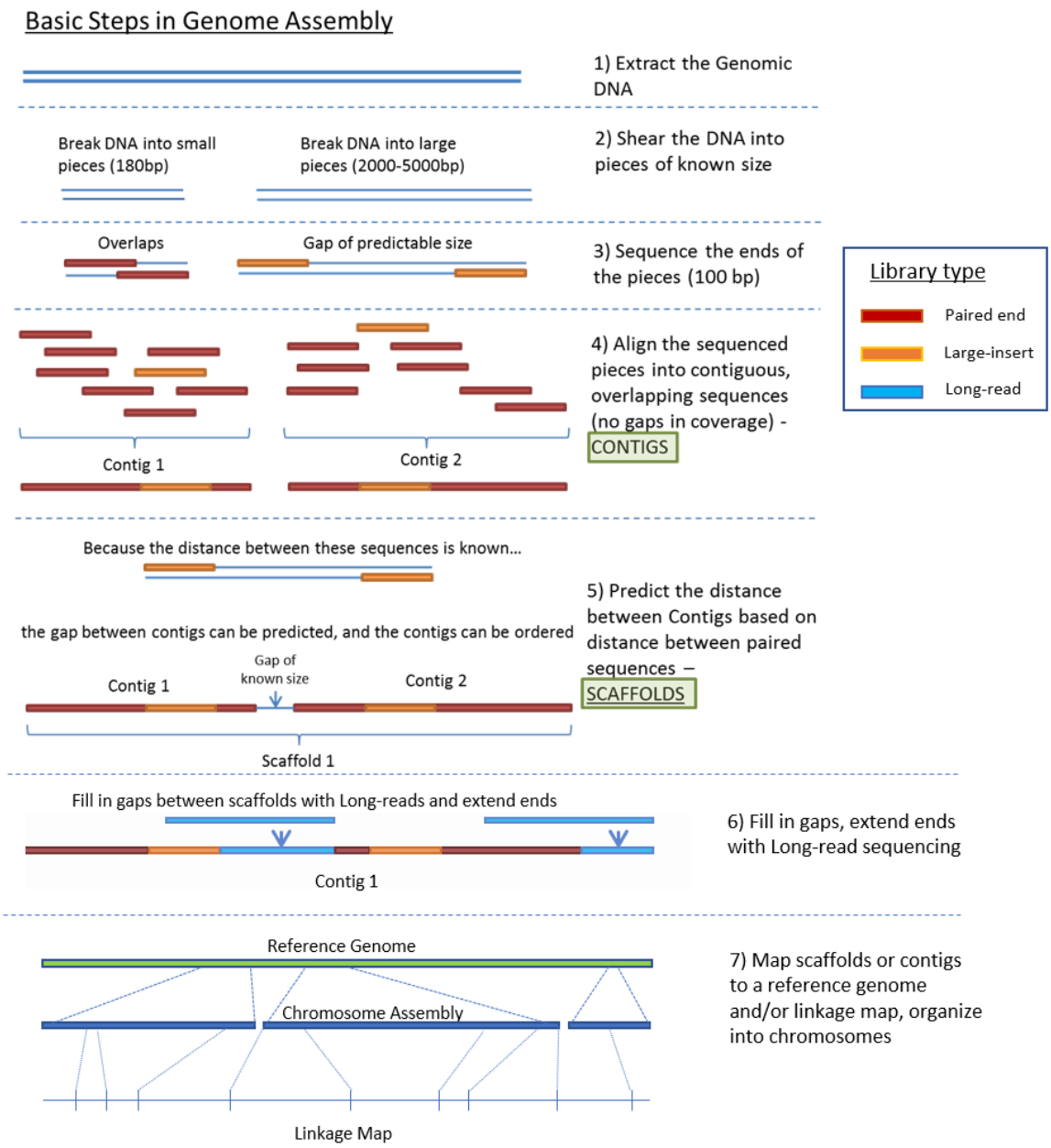


Figure 1.3: Basic steps in Genome assembly

Steps are summarized to the right

Most assembly approaches are not able to successfully bridge all gaps. Whether due to technological limitations of the sequencing method, the sheer size of the genome, or regions of reduced complexity or high-similarity (Schatz et al. 2012), the vast

majority of genome projects yield pieces of the genome that fall short of full-chromosomal length. While such fragmented approaches generally allow nearly full representations of gene content, structural analyses of the genome rely on the ability to organize the fragments into a linear order representing the chromosome. Thus, even as comparative mapping has moved from phenotype to genotype and now to full-genomic approaches, the root of all such analyses has remained a map (Fierst 2015). As shown in figure 1.3.7, a linkage map can be used to guide scaffolds into an ordered representation of the chromosome. Where two or more markers on a linkage map can be mapped to the same scaffold, these markers can then be used to orient the scaffold into a chromosome. In instances where a reference genome is available, common markers between the two can be used to generate a chromosomal representation (Shen et al. 2016). The use of common markers is generally done between species within a genus or perhaps within the family classification, although early efforts between human and mouse demonstrated the degree to which reference-guided efforts could exploited if conserved linkage could be demonstrated (Thomas et al. 2000). While we are rapidly approaching the ability to generate complete chromosome-level assemblies (Peichel et al. 2016), it is likely that in larger, more complex, and highly repetitive genomes, guiding chromosome construction by way of mapping will remain an important method.

Genomic representations, either through maps or through chromosome anchored assemblies, can be used to demonstrate conserved synteny and conserved linkage, by mapping a representation of the genome of one organism to that of another. This task can be done by comparing the placement of common markers between two linkage maps, such as was accomplished in early comparative mapping analyses with human and zebrafish (Postlethwait et al. 1998, Gates et al. 1999, Barbazuk et al. 2000). Conserved synteny can be revealed by mapping the markers on a linkage map from one species to the genome of another, such as was performed with the first teleost genome (Kai et al. 2005). Conserved synteny can be demonstrated by aligning entire chromosomes between two organisms, such as was done with the first two vertebrate

genomes, human and pufferfish (Aparicio et al. 2002). While comparing two linkage maps may only reveal limited information based on the necessity of common markers, the availability of at least one annotated genome enables numerous possibilities. If the conserved linkage is high, linkage can allow for the estimation of gene order as when two markers are collinear, the sequence between the two markers is highly likely to be co-linear as well. On a structural level, conserved linkage allows for the identification of chromosomal inversions and translocations between species. As such, the immense effort required to produce genomic resources in a model organism can be harnessed to inform efforts in species for which resources have not yet been developed.

With a linkage map or a genome, how might comparative mapping be utilized? Several examples can be found in figure 1.4. Comparative mapping can be used to identify gene gain or loss (Lai et al. 2004) (figure 1.4a). Comparative mapping can be used to generate a list of gene candidates in a species without a genome (Myosho et al. 2012, Takehana et al. 2014) (figure 1.4b). Comparative mapping can be utilized to inform a phylogenetic placement by tracing inversions, fusions, or translocations throughout evolutionary lineages (Prada et al. 2011, Sutherland et al. 2016) (figure 1.4c). Comparative mapping through linkage maps or genomes can provide evidence for genome duplication and subsequent re-organization of the chromosomes (Ahn and Tanksley 1993) (figure 1.4d). Thus, comparative mapping can be used to query a number of evolutionary and biological patterns.

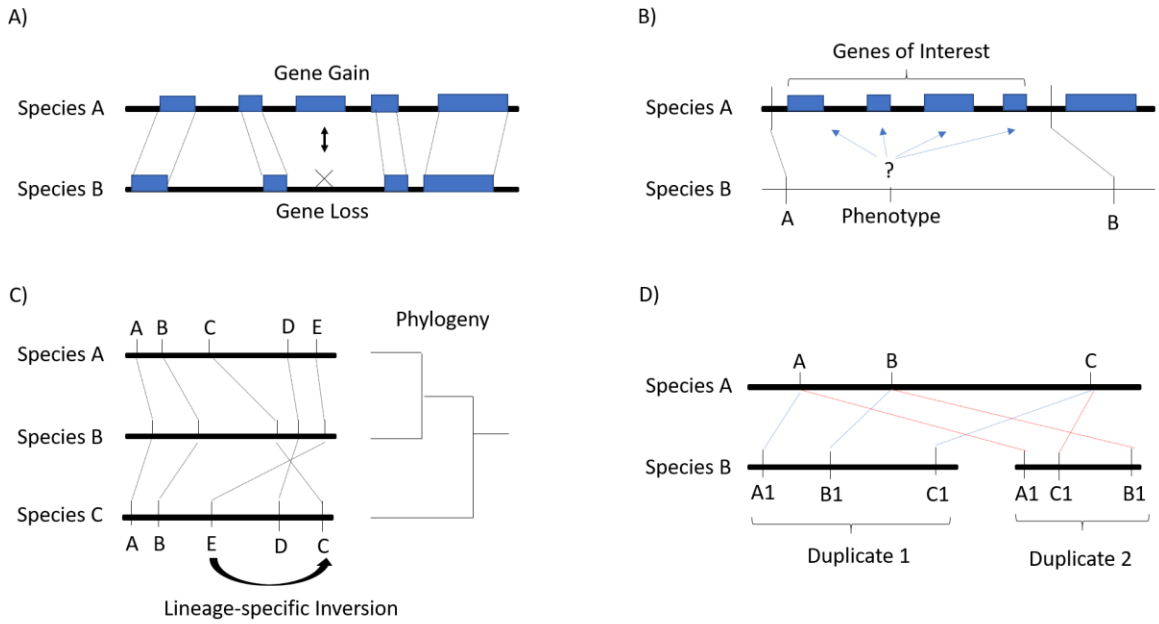


Figure 1.4: Uses of conserved synteny and linkage

Four potential uses of conserved synteny are presented. In Panel A, the gain or loss of a gene in one species can be determined by comparative mapping to another. In Panel B, a species for which a genome is available can inform a list of gene candidates in a species for which only a linkage map is available. In Panel C, the structural organization can be used as a component of the phylogenetic analysis. In Panel D, the re-organization post-genome duplication can be inferred from comparison to another species which has not experienced a duplication.

Comparative mapping and synteny analyses need not necessarily be between species to be informative. Although first suggested by Stanley Stephens in 1951 (Stephens 1951), it was Susumu Ohno, informed by work on chromosome numbers (karyotypes) on a number of animal lineages through the 1960s (Ohno et al. 1967), that theorized in his seminal book “Evolution by Gene Duplication” (Ohno 1970) that whole-genome duplication (WGD) may play a major role in evolution. Generating a “redundant” copy of the genome allowed the raw material to be rapidly moulded into generating new functions. While the evolutionary fate of gene duplicates remains an active area of research (e.g. (Braasch et al. 2016, Lien et al. 2016)), the genomic evidence for duplications themselves is well established. Molecular evidence for the two

rounds of genome duplication in animals proposed by Ohno (1R and 2R; R = Round) came to prominence in the early 1990s with analysis of duplicated gene clusters (Lundin 1993) with the hox and homeobox genes, in particular, reinforcing the idea (Holland et al. 1994). However, this evidence was not considered conclusive at the time as so many duplicates had been lost that tandem duplication and chromosomal translocation remained a possibility to explain these patterns. In 1997 following the publication of the yeast genome, Kenneth Wolfe and Denis Shields were able to provide evidence of an ancient genome duplication event as identified through blocks of conserved synteny between homeologous pieces of *Saccharomyces cerevisiae* chromosomes (Wolfe and Shields 1997). Subsequently, a number of ancestral duplication events have been identified in a similar manner in plants (Vision et al. 2000) and arthropods (Kenny et al. 2016, Schwager et al. 2017). Two rounds of duplication in the ancestral vertebrates are also now relatively well accepted (Panopoulou et al. 2003, Dehal and Boore 2005, Singh et al. 2015) (figure 1.5), particularly following the results from large comparative mapping studies with full genome sequences.

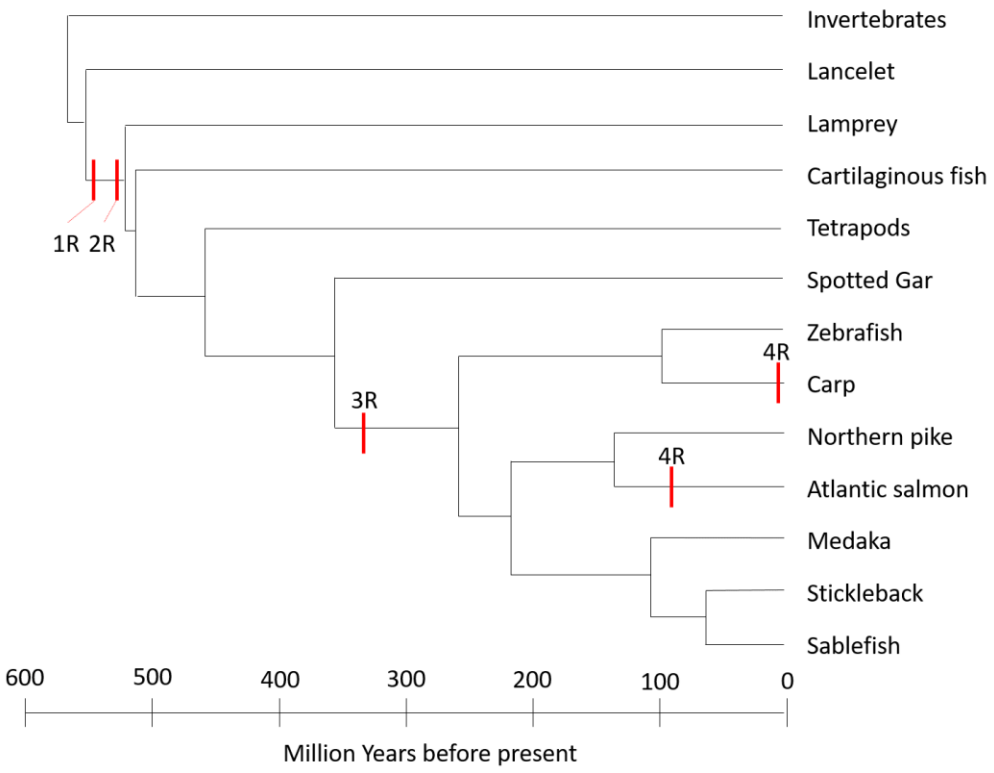


Figure 1.5: Phylogeny of whole-genome duplications in the teleost phylogeny

The phylogenetic timing of known or supported whole-genome duplications (WGD) that have occurred in the ancestry of all teleosts (1R, 2R, 3R) and lineage-specific duplications in carp and salmon (4R). R = Rounds of WGD.

Of the approximately 50,000 extant vertebrate species, about half are in the diverse Infraclass of fish known as the Teleostei, which arose approximately 250-350 mya (Near et al. 2012). One of the distinguishing features of this group is that their common ancestor experienced a whole-genome duplication event, known as the 3R duplication (Taylor et al. 2001, 2003, Hillier et al. 2004, Jaillon et al. 2004, Vandepoele et al. 2004). As the name would suggest, it is a third round of genome duplication following the separation from the remaining basal vertebrates that only share the first two rounds (summarized in figure 1.5). Following this 3R duplication, large-scale gene loss, genome rearrangement, and an increase in transposable element activity are thought to have occurred (Neafsey and Palumbi 2003, Inoue et al. 2015, Gao et al. 2016a), leading to

genomes in teleost fish that vary wildly in size (Hinegardner and Rosen 1972, Pie et al. 2006, Canapa et al. 2015). It is therefore a bit puzzling perhaps that chromosome number in the teleosts remain remarkably consistent. In their 2006 work, Judith Mank and John Avise presented a phylogenetic representation of chromosome number that shows well over half of teleost fish have a karyotype of 2n=48 or 2n=50 (Mank and Avise 2006) and chromosome number is largely conserved across the two major teleost lineages. The pre-duplication karyotype is thought likely to have been 2n=24 or 2n=26 (Putnam et al. 2008, Kai et al. 2014), suggesting that relatively little inter-chromosome rearrangement has taken place since the duplication event. Stable chromosome numbers suggested conserved synteny may remain high for much of the teleost lineage, implying the wide-spread use of comparative mapping in the bony fish may be particularly useful for non-model species to inform general structural organization. Demonstration of large-scale conserved synteny within the teleosts could allow rapid development of genetic resources, particularly with the introduction of *de novo* mapping techniques such as RAD-seq (Amores et al. 2011) and Genotyping-by-sequencing (Elshire et al. 2011), which require no prior genomic information.

It is unknown whether residual effects of WGD are a factor in the rate of intra- and inter-chromosomal rearrangement following a return to a diploid state (see (Sémon and Wolfe 2007)); such conservation of conserved synteny along a chromosome is not seen to such an extent in mammalian systems (e.g. Rodents (Zhao et al. 2004)) but can be seen in other vertebrates (e.g. amphibians (Sun et al. 2015)). Exceptions to chromosome conservation do exist, although many of the exceptions occur in lineages such as carp and the Salmoniformes, which are thought to have undergone a fourth round of duplication. With WGD estimates at 8 million years for carp (Xu et al. 2014), 86 million years for salmonids (Macqueen and Johnston 2014), and 250-350 million years for the 3R duplication (Christoffels et al. 2004, Jaillon et al. 2004, Vandepoele et al. 2004), teleosts offer the opportunity to look at least three different time-scales in the evolution of genomes post-WGD. Demonstrating the syntenic stability of the Teleostei

should allow for the identification of an outgroup species to the lineage-specific whole-genome duplication events. This would allow for the reconstruction of genomic rearrangements that occur following the more recent duplications in a similar way to which the Spotted Gar is to be used to inform the 3R duplication event (Braasch et al. 2016).

My goal in this work was to demonstrate the degree to which conserved synteny and conserved gene order has been maintained in the Teleostei and to demonstrate how rapidly evolving sequencing and genotyping technologies can be co-opted along with an understanding of conserved synteny to rapidly bring genomic understanding of little-studied organisms to model organism level quickly and economically. In Chapter 2, I sought to demonstrate the benefits of comparative mapping and the utility of conserved linkage to rapidly inform questions of biological significance in species for which very little genetic data exists. My species of interest was the sablefish (*Anoplopoma fimbria*), a high-value, long-lived finfish new to aquaculture in the Pacific region. The phylogenetic placement of the sablefish has long been in question (Imamura and Yabe 2002, Smith and Wheeler 2004, Shinohara and Imamura 2007), so I began by direct sequencing of the mitochondrial genome to resolve the phylogenetic placement. An Expressed-Sequence Tag (EST) library was developed and sequenced to identify potential microsatellite markers directly linked to genes. With help from Briony Campbell at Sablefish Canada Ltd. (now Golden Eagle Sable Fish), three families were generated, and microsatellite markers were used to develop a first-generation linkage map, the first linkage map for any fish in the order Scorpaeniformes. By designing the markers to conserved, transcribed sequences, I found enough matches through comparative mapping to demonstrate conserved linkage with the threespine stickleback, a model organism for which a high-quality genome assembly was available. Mapping the sex of the offspring as a phenotype on the linkage map, I linked the sex-determination gene to a small portion of sablefish LG14 with which synteny to the stickleback allowed the generation of a list of gene candidates. Finally, using a newly

developed Genotyping-by-Sequencing protocol for large-scale *de novo* SNP discovery (Elshire et al. 2011), markers were identified that both perfectly linked to sex and mapped to the region of the stickleback genome that was identified through comparative mapping. I identified a putative sex-determination gene and developed a simple PCR Assay to genetically sex sablefish. This work was published in *BMC Genomics* in 2013 (Rondeau et al. 2013) and was a collaborative effort with many of my fellow Koop lab members. This result not only demonstrated the utility of comparative mapping to leverage the resources developed for model organisms in non-model species, but also developed a manageable assay with which to genetically sex individual sablefish at any stage of development.

In Chapter 3, I extend the use of comparative mapping to more distant evolutionary relationships by focusing on the Protacanthopterygii. This group of fish includes the economically and culturally important order Salmoniformes. However, utilizing the salmonids in evolutionary comparisons is complicated by the whole-genome duplication event in the ancestor from which all members of the order descend. The most closely related order to Salmoniformes, the Esociformes, did not appear to experience this duplication event and thus was hypothesized to be more suitable for evolutionary comparisons of the Protacanthopterygii to the remaining teleosts. Of the 13 species that make up the Esociformes, the northern pike (*Esox lucius*) is the most studied and was the only species for which some precursory genetic resources were available (Leong et al. 2010).

I sought to leverage these earlier studies to develop genomic tools for the northern pike to represent a non-duplicated genome of the Protacanthopterygii. Beginning with microsatellites, tied to the transcribed sequences, and two families, we developed the first linkage map for the Esociformes. A preliminary genome assembly for pike was developed using Illumina Paired-End and Mate-Pair libraries. Roughly half of the resulting scaffolds could be anchored to chromosomes by the linkage map. A series of 13 tissues were sequenced, using Illumina RNA-seq technology, to produce a gene

expression atlas for the species and for use in annotation of the genome. Through comparative mapping of annotated genes anchored to chromosomes, I demonstrated that conserved synteny was extremely high between the northern pike and the model organisms' stickleback, medaka, and zebrafish, even after 210-230 million years of evolution. Finally, I demonstrated, through comparative mapping, that the pike represents a non-duplicated outgroup to the Atlantic salmon. This work was again a collaboration with many of the members of the Koop lab, particularly Amber Messmer, Jong Leong, and David Minkley, and was published in *PLOS One* in 2014 (Rondeau et al. 2014). With this work, I demonstrated not only that conserved synteny is maintained across much of the teleost lineage, but also that with very little inter-chromosomal rearrangement observed, the northern pike genome represents a potential outgroup for analyses of the rearrangements that occurred following the salmon-specific genome duplication.

In the rapidly developing field of sequencing technologies, new techniques now allow for genome refinement at a speed and cost that were previously unimaginable. Therefore, in Chapter 4 I discuss my attempt to improve the genome assembly of the northern pike and to use the newly published Atlantic salmon genome (Lien et al. 2016) to explore the structural rearrangements that have occurred between the two lineages. Re-scaffolding of the pike genome with large-insert mate-pair reads allowed for a four-fold improvement in the scaffold N50, a metric that defines the size of the scaffold from largest to smallest at which half the genome is represented. Utilizing the long-reads produced by a Pacific Biosciences RSII sequencer, the gaps in scaffolds between contigs were filled in, increasing the contig N50 nearly ten-fold, exceeding that of many model organisms. Finally, with help from Dovetail Genomics and an in-vitro proximity-ligation and assembly protocol known as Chicago/Hi-Rise, the scaffold N50 was further improved to 7.95 Mb, with a contig N50 of 125 kb. The northern pike genome at time of release had the 4th highest Contig and Scaffold N50s for any fish genome in NCBI's RefSeq assembly database. With 88% of the assembly mapped to chromosomes by the

genetic linkage map from Chapter 3, I demonstrated through comparative mapping the degree to which the Atlantic salmon genome has re-arranged post duplication. I further demonstrated a potential solution to linkage mapping in regions of residual tetraploidy by using northern pike as a reference guide. This work was a collaborative effort with Jong Leong and David Minkley. This chapter demonstrated the degree to which emerging sequencing technologies can rapidly and economically expand and refine genome sequences, brought the completeness of the northern pike to model organism level, and demonstrated a way forward to resolve regions of difficulty in current and future salmon genome assemblies.

This dissertation illustrates the utility of conserved synteny in species for which few genomic resources are available, which is the case for many non-mammalian vertebrates, including fishes. In sablefish, I specifically demonstrated that comparative mapping could be used to identify phenotype-linked genes without the need to develop an annotated genome. In northern pike, I reinforced the idea that synteny is generally conserved across many teleosts, thereby concluding that comparative mapping approaches should be practical in many non-model fish to leverage the availability of currently available genomic resources. I further developed the northern pike genome as a new model organism to represent the non-duplicated species within the Protacanthopterygii. Finally, I demonstrated the utility of this newly developed genome to inform major and minor structural rearrangements that occur in salmon lineage post genome-duplication.

Chapter 2 :

Genomics of Sablefish (*Anoplopoma fimbria*): Expressed genes, mitochondrial phylogeny, linkage map and identification of genetic sex markers

Abstract

Background

The sablefish (order: Scorpaeniformes) is an economically important species in commercial fisheries of the North Pacific and an emerging species in aquaculture. Aside from a handful of sequences in NCBI and a few published microsatellite markers, little is known about the genetics of this species. The development of genetic tools, including polymorphic markers and a linkage map will allow for the development of genetically diverse broodstock and mapping of phenotypes of interest. The sexual dimorphism between females and males makes a genetic test for early identification of sex desirable.

Results

A full mitochondrial genome is sequenced, and the resulting phylogenetic analysis verifies the placement of the sablefish within the Scorpaeniformes. Nearly 35,000 assembled transcript sequences are used to identify genes and obtain polymorphic SNP and microsatellite markers. 360 transcribed polymorphic loci from two sablefish families produced a map of 24 linkage groups. The sex phenotype maps to sablefish linkage group 14 of the male map. We show significant conserved synteny and conservation of gene-order between the threespine stickleback *Gasterosteus aculeatus*

and sablefish. An additional 1843 polymorphic SNP markers are identified through next-generation sequencing techniques. Sex-specific markers and sequence insertions are identified immediately upstream of the gene gonadal-soma derived factor (*gsdf*), the master-sex determinant locus in the medaka species *Oryzias luzonensis*.

Conclusions

The first genomic resources for sablefish provide a foundation for further studies. The genetic map represents, as far as we can determine, the first linkage map for a member of the Scorpaeniformes. The level of conserved linkage seen through comparative mapping will allow the use of the stickleback genome in future genetic studies on sablefish and other related fish, particularly as a guide to whole-genome assembly. The identification of sex-specific insertions immediately upstream of a known master-sex determinant implicates *gsdf* as an excellent candidate for the master sex determinant for sablefish.

Background

The order Scorpaeniformes is a diverse group of species that includes the rock cod, greenling, sculpins, and sablefish among others. Grouped originally by the presence of the suborbital stay, a posterior extension of the third circumorbital bone (Smith and Wheeler 2004), the order is now considered to be paraphyletic with members of the orders Perciformes and Gasterosteiformes (Shinohara and Imamura 2007). While several species among the currently defined order are considered commercially important, the most economically valuable in North America is the sablefish, *Anoplopoma fimbria*.

The sablefish, also known as Alaskan black cod, is a long-lived demersal species located mainly between 200-1500m along the continental shelf of the north Pacific ocean (Kimura et al. 1998). Found from eastern Japan through Alaskan waters and down

to Baja California in Mexico (Allen and Smith 1988), the sablefish is most important to American and Canadian commercial fisheries, with sablefish landings exceeding $200 million value on a combined harvest of 22,000 metric tonnes in 2011 (NOAA Fisheries: Office of Science & Technology 2012, Fisheries and Oceans Canada 2013). Sablefish is also in the preliminary stages of commercial aquaculture, with farms in British Columbia producing 500 tonnes in 2009 (Canadian Aquaculture Industry Alliance 2009). While commercially important, little genetic information is available, with genetic resources limited to a few sequences and 29 microsatellite primer pairs (Consortium Molecular Ecology Resources Primer Development et al. 2011, McCraney et al. 2012). Indeed little is known about the 1,477 species within the economically important order Scorpaeniformes (Nelson 2006).

For the successful development of sablefish as a sustainable aquaculture species and for the protection of wild populations, modern molecular tools will be of immense value. The identification and exploitation of genetic markers can be used in the characterization and identification of strains, parental identification, and analysis of diversity in the broodstock or in the construction of a linkage map (Liu and Cordes 2004). Linkage maps have been developed for numerous fish species including Atlantic salmon (Moen et al. 2008, Lien et al. 2011), channel catfish (Kucuktas et al. 2009), common carp (Zheng et al. 2011), grass carp (Xia et al. 2010), Atlantic halibut (Reid et al. 2007) and gilthead sea bream (Bargelloni et al. 2007). Traditionally maps were developed with markers such as allozymes, AFLP and RAPD markers, although microsatellites and SNPs are the current markers of choice. Microsatellites are relatively abundant, highly polymorphic and easy to genotype, and SNPs, while less informative due to a limit of two (or very occasionally three) alleles per locus, are easily identifiable and are the marker of greatest abundance in the genome. With advances in next-generation sequencing (NGS) protocols such as RAD-mapping (Amores et al. 2011) and Genotyping-by-Sequencing (Elshire et al. 2011), it is possible to generate a dense SNP map of primarily anonymous markers (type II markers) with relatively little prior DNA

sequence information. Having markers linked to genes (type I markers) rather than anonymous sequences, however, allows for the putative placement of genes on a linkage map, which can make them of greater use than type II markers in linkage mapping for aquaculture species (Liu and Cordes 2004). Type I markers, both microsatellites and SNPs, can easily be identified in libraries of expressed sequences, either through traditional EST library sequencing or through NGS transcript sequencing methods such as RNA-Seq. While linkage maps can have many uses, one of the most useful is in mapping phenotypes to a map to identify the genes controlling phenotype and look for associations or linkage between traits. For single-locus phenotypes, the phenotype can be scored and analyzed in the way a genetic marker would; multi-locus phenotypes require more powerful quantitative trait loci (QTL) analyses to predict regions of interest statistically.

Of the phenotypes often identified through mapping, sex is one of the most important and most common for many species. Sexual dimorphism is of significance when one sex can often grow larger or faster, both important traits to understand in fisheries and aquaculture. While mapping of the sex phenotype to a chromosome is relatively easy given a single-locus sex determination system, mapping the sex phenotype to a gene has been more difficult, and the master sex-determinant has only been identified in a few fish species. The gene *Dmy* determines sex in medakas *Oryzias latipes* (Matsuda et al. 2002) and *O. curvinotus* (Matsuda et al. 2003), *amhy* in the Patagonian pejerrey *Odontesthes hatcheri* (Hattori et al. 2012), and *sdY* determines sex in most salmonids (Yano et al. 2012, 2013). These master sex-determining genes (MSD) are the result of divergent duplicated copies of autosomal genes, present only on the Y-chromosome. $Gsdf^{Y}$ in the medaka *Oryzias luzonensis,* on the other hand, is up-regulated in males during sexual differentiation due to changes in the upstream promoter region (Takehana et al. 2014), while the male-specific $Amhr2^{Y}$ in 3 species of *Takifugu* appears to be the result of a single coding change (Kamiya et al. 2012). While all these genes have been previously described as playing a role in the sexual

determination of teleosts (aside from *sdY*), none so far have been described as the master sex determinants in distantly related groups of fish.

Comparative mapping (Lee and Kocher 2007) can be used to facilitate comparison of newly developed maps to pre-existing genomic resources to augment available genetic information. Comparison between non-model fish species and those with a fully-sequenced reference genome has identified significantly conserved synteny in numerous species (Sarropoulou et al. 2008, Kucuktas et al. 2009, Xia et al. 2010, Liu et al. 2012, Palti et al. 2012). With preliminary or complete whole-genome assemblies available for ten bony fish species (October 28, 2012), and the increasing ease with which genetic information and complete genomic sequences can be obtained (Catchen et al. 2011), comparative mapping will only become easier. The degree to which the identification of conserved synteny will be useful depends on how closely related the two species are; thus, for species of relatively close ancestry and significant macro-synteny, comparative mapping can be used to predict gene locations and order, and provide a list of potential candidates responsible for a phenotype or underlying a QTL.

In this work, we describe the sequencing of the mitochondrial genome and the resultant verification of the phylogenetic placement of the sablefish within the Scorpaeniformes. We present a library of assembled transcript sequences to identify genes and use them to develop type I SNP and microsatellite markers. The markers were scored across two families and used to produce the first-generation sablefish linkage map and locate the sex phenotype onto the male map. We show conserved synteny between the threespine stickleback *Gasterosteus aculeatus* and sablefish and use a comparative mapping approach to predict gene locations in the sablefish. An additional collection of markers, mainly type II SNPs are identified through genotyping-by-sequencing and used to identify sex-specific markers and sequence insertions immediately upstream of a known master-sex determinant.

Results and Discussion

Gene Identification

To obtain a comprehensive picture of the gene sequences of sablefish, we examined transcriptomes by EST sequencing and RNA-seq. In total, 19,968 cDNA clones were sequenced in both forward and reverse directions. Following trimming and removal of any contaminants, 34,080 EST sequences were obtained and deposited in the GenBank EST database under accessions GenBank: GO615858-GO649937. EST sequences were assembled into 12,060 unique contigs using PHRAP assembly. 1,249 full length gene sequences were submitted as GenBank: ACQ57837– ACQ59081, GenBank: C3KHG1, GenBank: C3KJF2, GenBank: C3KHF2 and GenBank: C3KJE6. A 3'-UTR analysis of all contigs was used for primer design in the search for polymorphic markers. RNA-seq yielded 96,733,584 reads, which were assembled into 92,888 UniGene sequences. Combined with the EST sequencing data into NCBI TSA BioProject 71237, the assembled data produced 34,728 contigs > 400bp; these can be found in accession numbers GenBank: JO657891-JO692618.

The uploaded EST and TSA datasets represent all but 84 of the 70,138 sequences for sablefish in the NCBI nucleotide and EST databases (retrieved October 7, 2012). In addition, sablefish is currently the Scorpaeniformes species with the largest percentage of sequences in both the EST (46.9% of total) and nucleotide databases (72.7%). This work represents a significant increase in the available expressed sequence data for the sablefish, and for the order Scorpaeniformes in general.

The mitochondrial genome and phylogenetic placement of sablefish

To help resolve the general phylogenetic placement of sablefish we sequenced the mitochondrial genome and used phylogenetic tree analysis to compare it to other Percomorpha and Scorpaeniformes. The sablefish mitochondrial genome, seen in Additional file B1, consists of 16,507 base pairs, encompassing 13 genes, 2 ribosomal

RNAs, and 22 tRNAs, with all features following the order of the typical vertebrate mitochondrial genome (reviewed in (Pereira 2000)). The sequence was submitted to NCBI as GenBank: JX070112.

The mitochondrial genome was analyzed using the methods and species analyzed in Kawahara et al. (Kawahara et al. 2008). The larger phylogenetic analysis of the mitogenome sequences produced the same major phylogenetic groupings as previously described, so we focused on the grouping of Scorpaeniformes, Zoarcoidei, and Gasteroidei. As seen in figure 2.1, we recovered a very similar set of phylogenetic placements for the species previously analyzed (Subgroup G in (Kawahara et al. 2008)). Additionally, we recovered a similar phylogenetic ordering of the Gasterosteiformes mitogenomes (Kawahara et al. 2009). The sablefish mitogenome was placed in an ancestral branch prior to the split of the Scorpaeniformes family Cottidae from the most-closely related Gasterosteiformes and Perciformes families, although the bootstrap value of this placement is not strong.

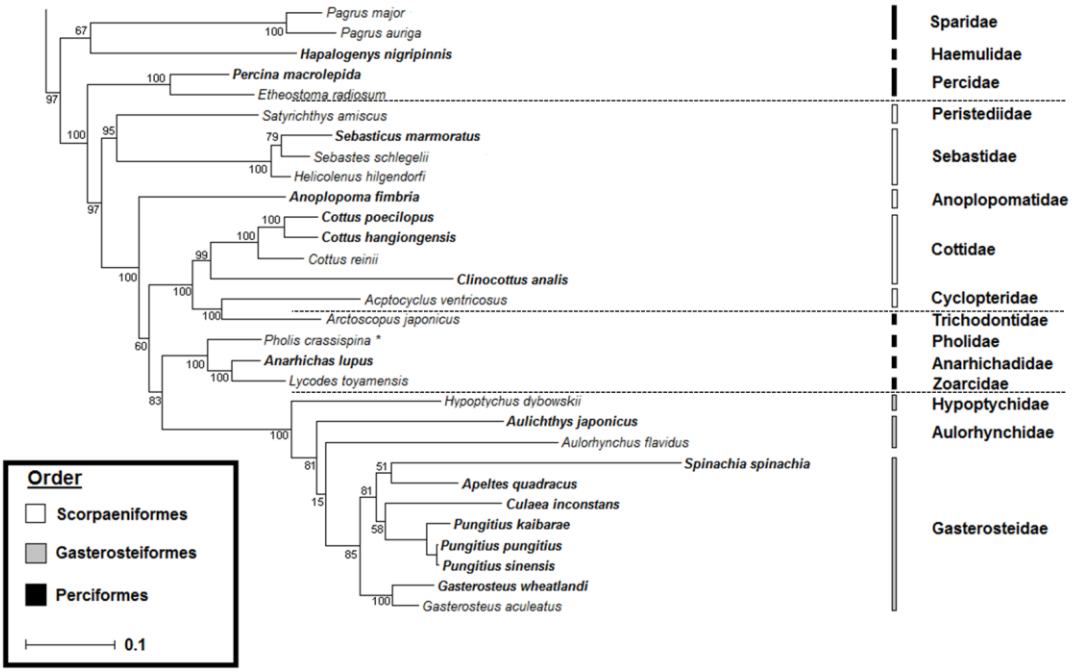


Figure 2.1: Mitogenome phylogeny of orders: Scorpaeniformes, Gasterosteiformes, select Perciformes

Species represented in bold text are not part of the previous analysis by Kawahara et al. (Kawahara et al. 2008). Families currently assigned to the order Scorpaeniformes are represented by white bars down the right-hand side, while Perciformes and Gasterosteiformes are in black and grey respectively. Boot-strap values are indicated for all branch points. The asterisk (*) indicates a change of genus name from previous analyses.

The Anoplopomatidae is a family comprising two known species, the sablefish and the skilfish (*Erilepis zonifer*) placed in the order Scorpaeniformes. Molecular ((Smith and Wheeler 2004); figure 2.1) and morphological (Imamura and Yabe 2002) analyses both suggest a closer relationship between sablefish and the Cottidae than to the scorpanoids, but the exact placement remains unclear. While our analysis agrees with that of Smith and Wheeler (Smith and Wheeler 2004) in suggesting a relatively ancestral branch for the divergence of the sablefishes, our results conflict on whether divergence occurred before or after the divergence of the Cottidae from the Gasterosteodei and the Zoarcoidei; both analyses produce weak branch support in favour of the respective

positions. Given the difficulty in accurately placing deep phylogenetic branches, a definitive molecular placement of the Anoplopomatidae will await further data.

While the exact placement of the Anoplopomatidae remains to be determined, sablefish represent an older unique branch that branched at or around the timing of the split of Cottoidei and Gasterosteodei. This is of interest as one member of the Gasterosteodei, the threespine stickleback *Gasterosteus aculeatus* has a fully sequenced draft genome available for comparison. The stickleback is without question the most closely related species for which a well-annotated genome sequence is available. Previous estimates based on mitochondrial genomes place the time of divergence of the more ancestral suborder Scorpaenoidei from the Cottoidei and Gasterosteodei at approximately 150 mya (Setiamarga et al. 2009), meaning the divergence of sablefish from stickleback lineages could be less than 150 mya. More recent estimates using nuclear data suggest a much more recent time of divergence in the range of 50-70 mya (Near et al. 2012). The next most closely related species with fully sequenced genomes, those of the pufferfishes *Takifugu rubripes* and *Tetraodon nigroviridis* are separated from sablefish by 80-180 mya (Setiamarga et al. 2009, Near et al. 2012). As all the phylogenetic data points to the stickleback as the closest fully-sequenced genome, *G. aculeatus* was used in our analysis of conserved synteny.

Identification of Genetic Markers

A set of polymorphic microsatellite and SNP DNA markers from our EST assemblies were identified for use in genetic mapping and population genetic studies. As the EST library construction and subsequent primer design were performed prior to obtaining sablefish family tissues, the sequences do not represent either of the two families used in linkage map construction. This did not affect the use of microsatellites as the identification and amplification of polymorphic repeats proved relatively easy, with over half the 460 primer pairs designed successfully amplifying the desired product and yielding polymorphic repeats in one or both sablefish families. In contrast, the

identification of SNPs from the assembled EST dataset proved much less useful when targeting expected SNPs in our mapping families. However, in analyzing the adjacent sequence, mainly the 3'-UTR, it was found that sablefish were indeed highly polymorphic. This led to the strategy of identifying SNPs through the direct sequencing of 250 unrelated 3'-UTR gene regions of the sablefish family parents. Additional primers were designed to genes of interest based on conserved sequence identified in related species. Primer pair sequences, repeat motifs and other relevant information for the 233 newly identified microsatellites, as well as the 13 previously published (Consortium Molecular Ecology Resources Primer Development et al. 2011), can be found in Additional file B2. For the SNPs, primer and annotation information can be found in Additional file B3, with individual SNPs identified in Additional file B4.

A preliminary set of markers was used to analyze a second, half-sibling family to determine the number of progeny from each of the three fathers. Most of the progeny were descended from a single father (96 offspring) with only minor contributions from the other two fathers (13 and 6). It was therefore decided to analyze Family 2 as a single full-sib family, excluding the handful of individuals descended from the other two fathers. Following genetic marker identification, we began locating these markers in two sablefish families.

Production of first-generation linkage map

Equipped with our gene-associated markers and two sablefish families, we examined all the markers in all the individuals from the two families to create a first-generation linkage map. In all, 246 microsatellites and 509 SNPs were scored for all the individuals from the two families. When it came to assembling the linkage map, the strategy of identifying multiple SNPs in a single sequence proved very useful. While most of the microsatellite markers were highly polymorphic and useful for mapping in both families, it was very rare that individual SNPs were shared between families. It was common, however, to have two or more markers from a single EST assembly-derived

sequence polymorphic in separate parents and families. As these markers are physically linked, separated by 600bp or less in a single contiguous sequence, we analyzed the SNPs as representing the locus rather than a single marker and placed the locus on the linkage map rather than the individual nucleotide marker. This strategy was also used when multiple microsatellite primers designed from the same contig were analyzed. This allowed merging of the individual parental maps and sex-specific maps at many more sites than with the microsatellites alone, leading to a more accurate representation of marker order and distances. The 509 scored SNPs, therefore, were mapped as 133 SNP loci; between 1 and 13 SNPs contributed to the mapping of an individual locus.

We produced maps for each of the mapping parents, with subsequent integration into sex-specific maps and finally a merged map. The final merged map, seen in Additional file B5, consists of all 133 SNP loci and 227 microsatellite loci (234 individual microsatellites) mapped across 24 linkage groups. Two additional microsatellite markers, AfiMI0079UVic and AfiMI0131UVic, remain unlinked and 10 primer pairs, while initially observed to be polymorphic in unrelated sablefish, were subsequently found to be uninformative in both mapping families (AfiMI0005UVic, AfiMI0015UVic, AfiMI0104UVic, AfiMI0153UVic, AfiMI0165UVic, AfiMI0196UVic, AfiMI0240UVic, AfiMI0304UVic, AfiMI0411UVic, AfiMI0416UVic).

The 24 linkage groups span 1332.8 cM in the merged map, with individual linkage groups ranging from 20.9 cM (LG22) to 80.3 cM (LG15). A significant difference in recombination was observed between the sexes as the male map had a length of 860.4 cM, while the female map had a length of 1610 cM; this gives a female:male recombination rate of 1.87:1, although the recombination rate in individual linkage groups varied greatly. This is not unusual, as higher recombination rates have been reported in the females a number of fish species, including Atlantic salmon (Lien et al. 2011), catfish (Kucuktas et al. 2009), rainbow trout (Rexroad et al. 2008) and zebrafish (Singer et al. 2002). In contrast, the recombination rates between the individual fathers

or between the individual mothers of our mapping families were both approximately 1:1. The overall length and female/male recombination ratios are likely to be underestimated, however, as there are still a few gaps in the female-specific linkage map.

With only 2 markers remaining unlinked after mapping, we were confident that all the chromosomes were covered by the map. This was subsequently confirmed after the publication of the sablefish karyotype (Phillips et al. 2013). A diploid chromosome number of 2n = 48 was identified, the same number as was predicted here through linkage mapping. Of note, this karyotype represents the most-commonly identified karyotype in teleost fish and is thought to represent that of the common teleost ancestor after the last whole-genome duplication (Jaillon et al. 2004).

Significantly conserved synteny observed between *A. fimbria* and *G. aculeatus*

A comparison to the most closely related species for which a whole genome sequence is available was undertaken to look for syntenic chromosomes and conservation of gene order. All the markers were derived from EST transcripts. These EST contigs, which include markers that mapped to one of the 24 linkage groups, or the 2 singletons, were mapped against the threespine stickleback genome. Most (278/360) of these contig comparisons produced a significant “hit” to the stickleback genome; furthermore, as can be seen in Table 2.1, the level of conserved synteny between the two species is quite high, with many markers from one sablefish linkage group producing significant “hits” to the corresponding stickleback chromosome. In most cases, the relationship between stickleback and sablefish is 1:1 – that is, one stickleback chromosome corresponds to a single sablefish linkage group. In three cases, there were two linkage groups for one stickleback chromosome. Analyses of end-markers, however, show no hints of linkage between any of the three pairs of linkage groups. This is not surprising given the haploid chromosome number in the threespine stickleback is 21 while the haploid karyotype for sablefish is 24. We used these relationships to assign

numbers to each of our linkage groups to make the comparison to stickleback easier; thus, LG02 corresponds to stickleback ChrII, and LG11 corresponds to stickleback ChrXI. In the case of the three linkage group pairs mapped to single stickleback chromosomes, the largest linkage group was assigned the stickleback chromosome number, while the smaller received the next number available above 21; resulting pairs were LG01 and LG22, LG04 and LG23 and finally LG07 and LG24.

		Sablefish LGs																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Stickleback Chromosomes	I	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	1	-
	II	-	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	III	-	-	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	IV	-	-	-	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	-
	V	-	-	-	-	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-
	VI	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	VII	-	-	-	-	-	-	13	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	3
	VIII	-	-	-	-	-	-	-	13	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-
	IX	-	-	-	-	-	-	-	-	14	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-
	X	-	-	-	-	-	-	-	-	-	12	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	XI	-	-	-	-	-	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-	-	-
	XII	-	-	-	-	-	-	-	-	-	-	-	19	-	-	-	-	-	-	-	-	-	-	-	-
	XIII	-	-	-	-	-	-	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-
	XIV	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-	-	-	-	-	1	-	-	-	-
	XV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	-	-	-	-	-	-	-
	XVI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-	-	-	-	-
	XVII	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11	-	-	-	-	-	-	1
	XVIII	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-	-	-
	XIX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7	-	-	-	-	-
	XX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-
	XXI	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	8	-	-	-
	Un	-	-	-	-	-	1	-	1	-	2	1	2	3	1	1	-	7	-	-	-	2	-	1	2
No Significant Hits		0	1	2	0	2	2	0	6	4	1	3	5	1	1	2	3	2	4	7	2	2	2	2	2
Total Markers		6	11	18	14	16	11	13	20	18	15	26	26	12	18	12	17	21	18	14	20	12	5	9	8
Orthologous to Gacu		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	I	IV	VII

Table 2.1: Comparative synteny between threespine stickleback and sablefish

The number of sablefish loci with significant BLAT hits to the stickleback genome and the predicted orthologous chromosomes for each linkage group.

Linkage groups showed anywhere from 3 (LG24) to 17 (LG04) hits to a single stickleback chromosome. Only 8 of the 278 loci that were located to the 21 chromosomes in the stickleback produced significant hits to chromosomes other than that predicted for the linkage group. Six of these eight markers were located 3 markers or less from the end of the linkage group. An additional 24 loci, interspersed throughout the linkage groups, mapped to the stickleback ChrUn, a “chromosome” composed of all sequence data unassigned to one of the 21 chromosomes. While most linkage groups had 0-2 loci that hit this “Un” chromosome, 7 hits were from loci assigned to linkage group 1. Only one of the two unlinked microsatellite markers produced a significant hit, mapping to ChrI in the threespine stickleback; this suggests that the chromosome represented by either LG01 or LG22, the two linkage groups that are predominately associated with stickleback ChrI, is likely larger than predicted by the current sablefish linkage map.

Gene order also appears relatively well conserved between sablefish and the threespine stickleback, as shown in figures 2.2-2.4. In most cases, 1-3 inversions can be used to explain the difference in marker ordering between the two species, with markers located between inversion points consistently ordered. This is important and will be revisited later, as it allows for prediction of candidate genes of interest in the sablefish based on the annotated whole-genome sequence in stickleback. Development of the sablefish linkage map and the identification of conserved linkage may present an opportunity to re-evaluate the stickleback genome assembly and successfully locate some of the loci previously pooled into ChrUn.

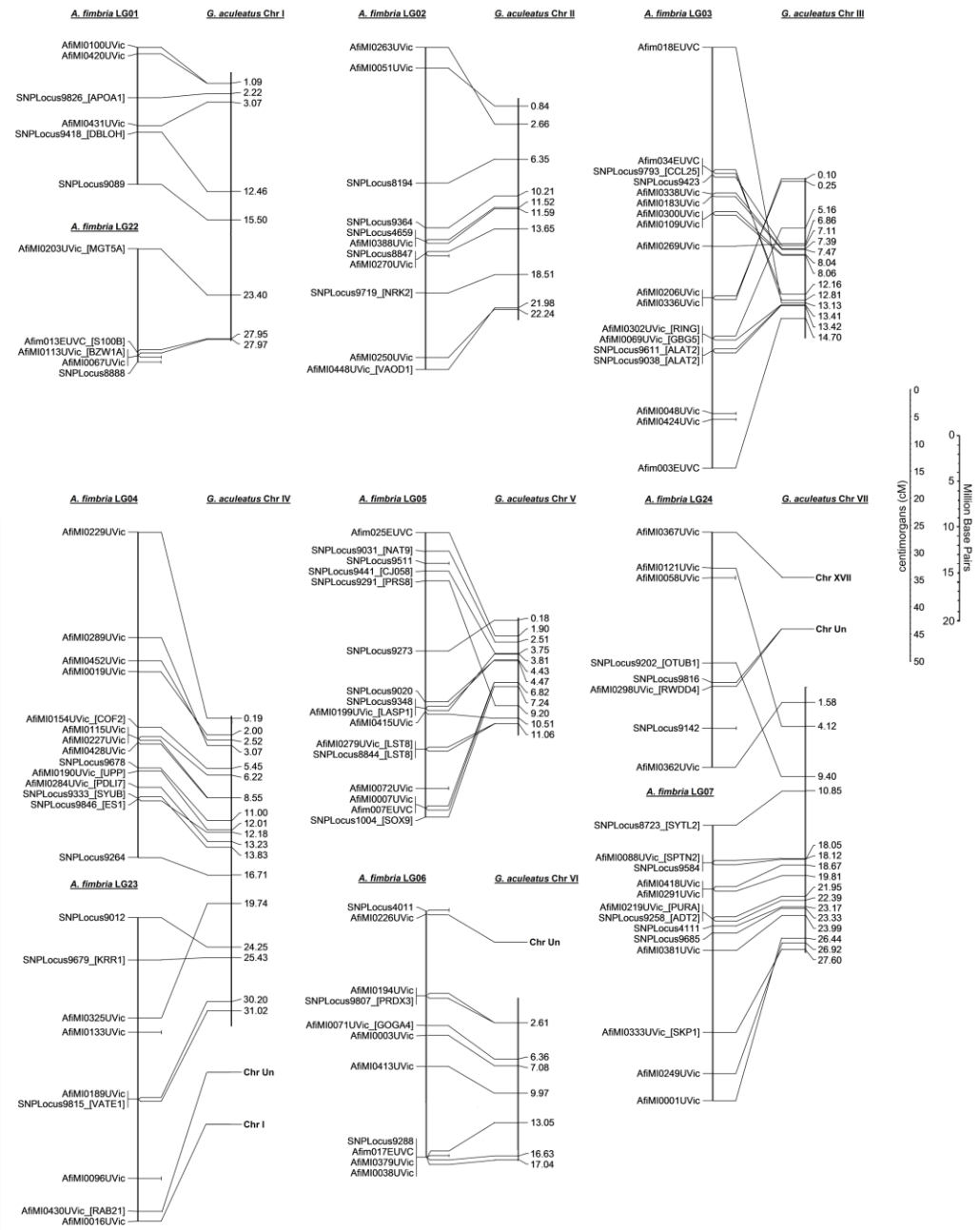


Figure 2.2: Conserved linkage between sablefish and threespine stickleback chromosomes I-VII

A line is drawn to compare the position of a marker on a linkage group to the strongest BLAT hit (>100 BLAT score) for the contig sequence used in primer design.

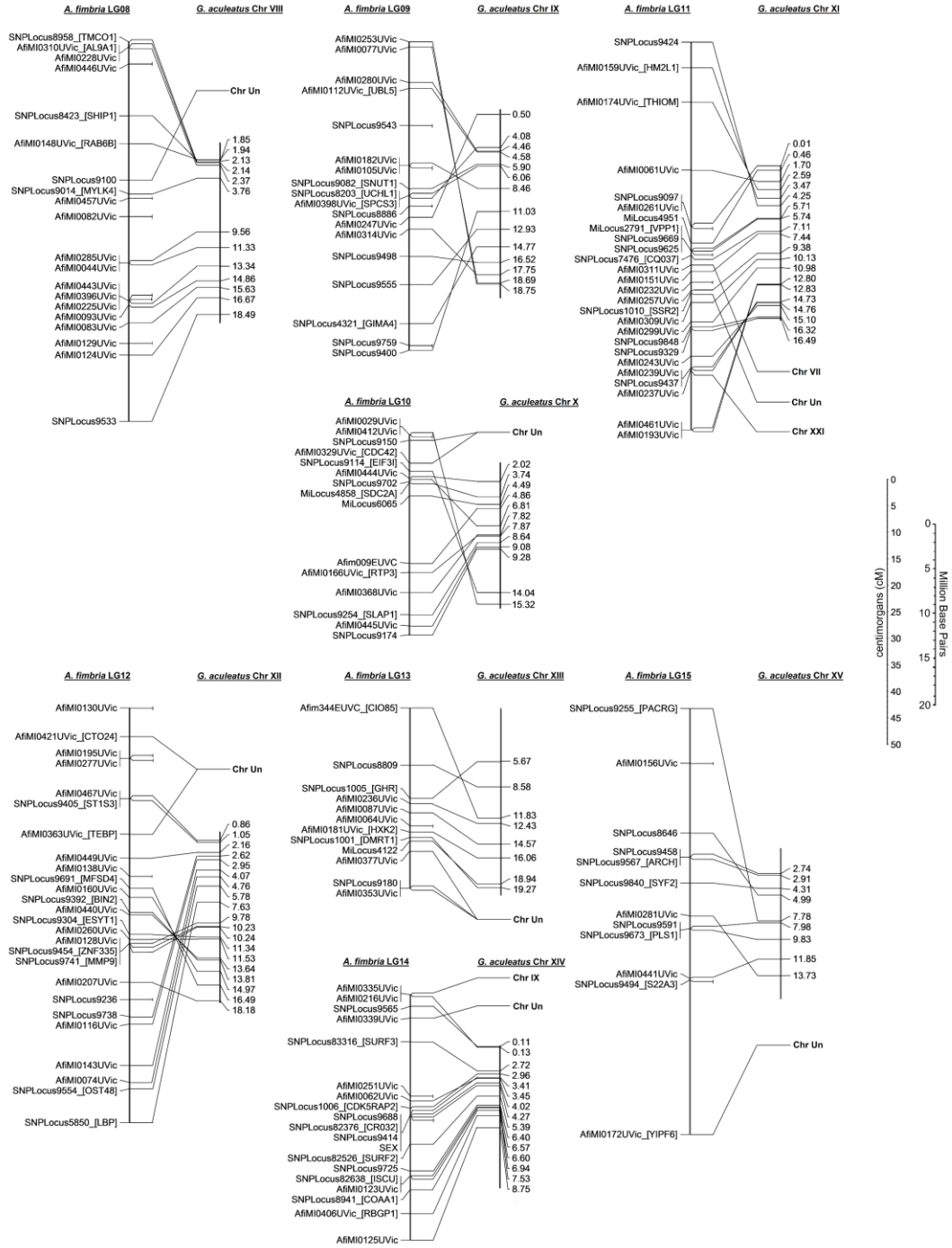


Figure 2.3: Conserved linkage between sablefish and threespine stickleback chromosomes VIII-XV

A line is drawn to compare the position of a marker on a linkage group to the strongest BLAT hit (>100 BLAT score) for the contig sequence used in primer design.

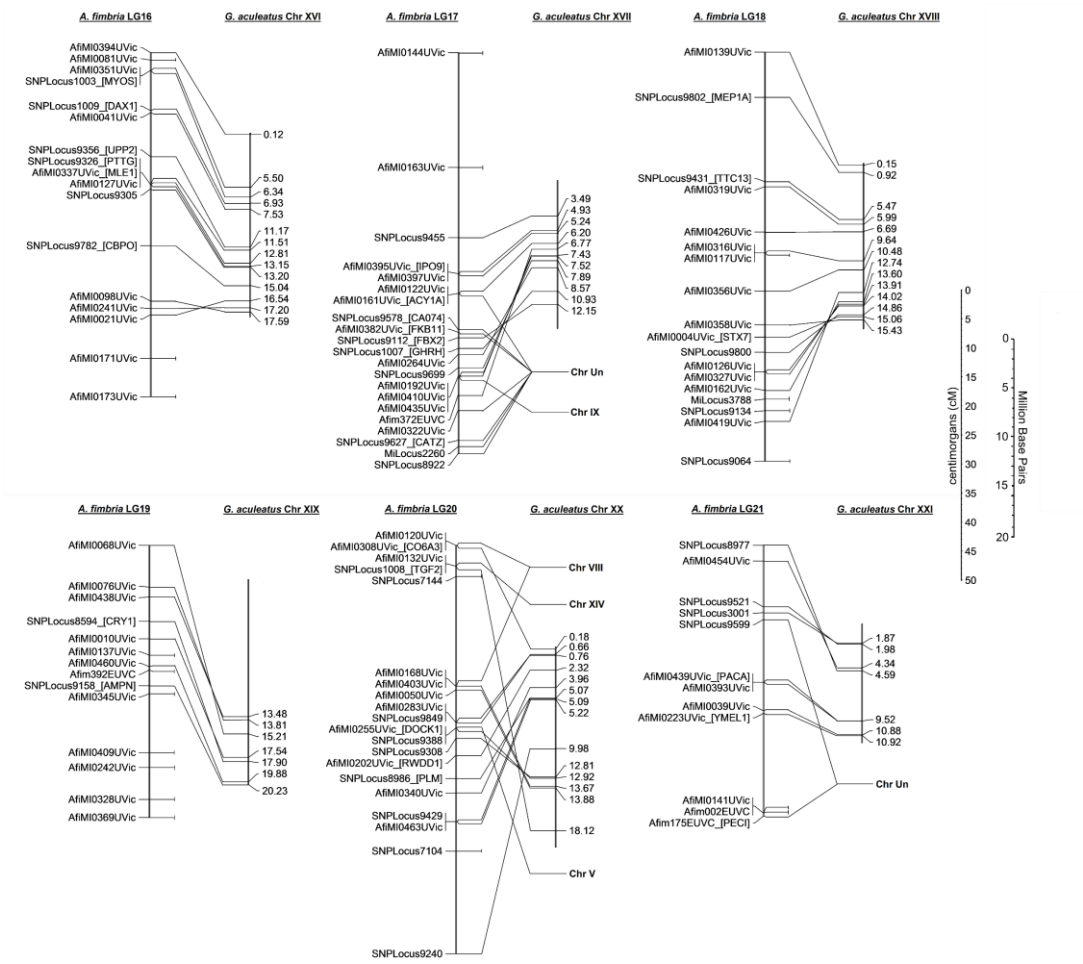


Figure 2.4: Conserved linkage between sablefish and threespine stickleback chromosomes XVI-XXI

A line is drawn to compare the position of a marker on a linkage group to the strongest BLAT hit (>100 BLAT score) for the contig sequence used in primer design.

Additional markers by Genotyping-by-sequencing (GBS)

The construction of a GBS library was performed to create a second-generation linkage map to supplement our microsatellite/SNP sablefish map and to identify potential sex-specific sequences. Two Illumina lanes of GBS sequencing were produced to test this technique for sablefish linkage mapping and to determine whether the *Ape*KI restriction enzyme used was appropriate for sablefish; most of the individuals were

from our second mapping family, while a few additional, unrelated and sexed individuals were used to aid in sex identification (see the following section). We obtained ~490 million barcoded reads, with between 500,000 and 1,300,000 unique reads per individual. We determined there was an average of 8-10 fold read coverage per unique genomic sequence, well below the desired 60x for a *de novo* sequencing study (Davey et al. 2011). While overall coverage was less than desired, we were still able to utilize the data to identify additional SNP markers. With relatively conservative parameters, a total of 1843 polymorphic loci were identified through next-generation sequencing; see Additional file B6 for details. These SNPs, however, were not used to further expand the linkage map.

Genetic identification of sex in sablefish

In sablefish, as in other fish species, one of the most economically and biologically important and perplexing phenotypes is sex. Sexual dimorphism in fish is often observed, with one sex growing both faster and larger than the other, commonly the female in sablefish. Rapid and substantial growth is a desirable trait, particularly in aquaculture and ocean ranching. Thus, selection for females early in rearing, or the production of monosex female offspring is of great interest. While external identification of sex is difficult and inaccurate in immature sablefish, internal analysis of the gonads can usually determine the sex of the fish but sacrificing the fish is required. At 16 months, the sex of each of our fish was easily identified by the size of the gonads themselves, with the ovaries well over twice the size of the testis in males and females of equal size and weight. Assuming a single locus phenotype, we attempted to place sex on each of the sex-specific maps. While unsuccessful in linking sex to the female map, we were successfully able to map the trait to linkage group 14 of the male map, suggesting an XX-XY sex determination system. As previously described in the comparative mapping section, this corresponds to chromosome XIV of the stickleback genome. Adjacent markers further narrowed this region to between 3.0 and 5.4Mb on

the stickleback genome. This region does not correspond to the Y-specific region of the threespine stickleback linked to ChrXIX (Peichel et al. 2004) nor to the Y-chromosome associated with LG12 (ChrXII) of the ninespine stickleback *Pungitius pungitius* (Ross et al. 2009). During review, a report was published on the expression of five genes of interest in sex, including three shown to be significantly elevated in juvenile testes (Smith et al. 2013). Two of these three are located on the linkage map, *dmrt1* mapping to LG13 and *sox9a* to LG05, both predicted by comparative mapping to the threespine stickleback genome. Neither the remaining testis-elevated gene, *amh*, nor the two genes found with elevated expression in the ovaries, *foxl2* and *cyp19a1a* were predicted to be found on LG14; *amh* is predicted to be located on LG08, *foxl2* on LG01 and *cyp19a1a* on LG02.

Examination of the sablefish GBS library was undertaken to determine whether we could identify any sex-specific sequences. The sequences were processed using the program Jellyfish (Marçais and Kingsford 2011) into short fragments (kmers) with length 31, and each kmer was counted in each individual, and placed into a matrix. We searched for unique kmers found only in male or female fish, with a minor allowance for sequencing errors (max two individuals). After assembling sex-specific kmers into overlapping mini-contigs, we could identify 11 mini-contigs of interest, which were then mapped back to the original 100bp paired-end reads. These reads were BLATed against the stickleback genome, and two of these reads produced hits >50 (default BLAT score), both of which BLATed against the region where we expected to find sex in ChrXIV of stickleback. These sites have been designated Tag2 and Tag10 based on the order that they were identified. Genome walking was performed to design primers to score the polymorphic markers in our sex-specific tags, and this process recovered well over 1000bp surrounding each tag for primer design. Following amplification with the new primers on 53 unrelated, definitively sexed sablefish, the alleles identified by Tag 2 and Tag10 were present in every male and absent in all females.

Both the tag sequences BLAT to the region between the genes PPEF2 and AFF1 on the stickleback genome. In this region, and in particular between these two tags, two

coding regions were predicted to be found based on stickleback ESTs. Based on tBLASTx, the first coding region (ESTs GenBank: DW615685, DN718296, DW662322, DN733719, DT981615) was an *aff1* homolog. The second gene region (ESTs GenBank: CD507187, DW624794) was most likely gonadal soma-derived factor (*gsdf*). Primers were designed to amplify the region between our tags and a short region of *gsdf* available in our sablefish RNA-seq assembly. Amplification of the long PCR products that stretched between Tag10 and gsdf revealed a male-specific fragment that was ~500bp larger than the PCR products obtained from female samples. No differences between male and female samples were observed between the coding region of *gsdf* and Tag 2 (data not shown).

Primers were designed to span the region between Tag 10 and gsdf in order to identify the male-specific sequences. Sequencing the resulting amplified products yielded a number of interesting features in the sequence upstream and within the *gsdf* gene (see Figure 2.5a). Two regions of ~180bp and ~140bp are each present in duplicate (95% identity) in the sequence 5' of the gene, while a number of smaller repeats are found throughout the sequence. A number of SNPs in the upstream and intronic portions of gsdf appeared to be linked to sex, and there were two exonic polymorphisms, one of which caused a sex-specific coding change from a phenylalanine in the X-chromosomal copy to a leucine in the Y. Most surprising though was the presence not only of a Y-specific sequence of 935bp but an X-specific sequence of 412bp that produced the previously observed 500bp difference in amplified size. Primers were designed to confirm X and Y specificity of the sequences. A PCR was designed to amplify across the X-specific sequence, with primers designed just outside of the specific sequence to produce amplified segments from both chromosomes; as seen in Figure 2.5b, amplification in both sexes show the larger band corresponding to the X-chromosome, while only amplification in the males produced the small band from the Y-chromosome demonstrating the lack of X-specific insert sequence. Amplification of the Y- specific sequence was more difficult due to the insertion falling in the middle of a

repeated sequence; a nested PCR was therefore performed targeting only the male-specific sequence. As shown in Figure 2.5c, the nested primers produced bands only in the male, not in the females. In the 53 unrelated, definitively sexed sablefish analyzed, all tested females carried two copies of the X-specific insertion, while males carried one chromosome with the X-specific insertion and one with the Y-specific, adding further evidence to the male being the heterogametic sex. Both sequences are flanked by inverted repeats of 12 (Y-Chr) to 16 (X-Chr) base pairs, suggesting the original insertion may have involved a transposase, although no transposase could be identified in the inserted sequences. Representative sequences for the X and Y insertions were uploaded to the NCBI nucleotide database as GenBank: KC623942 and GenBank: KC623943 respectively; masked sequences with all described elements can be found in Additional file B7.

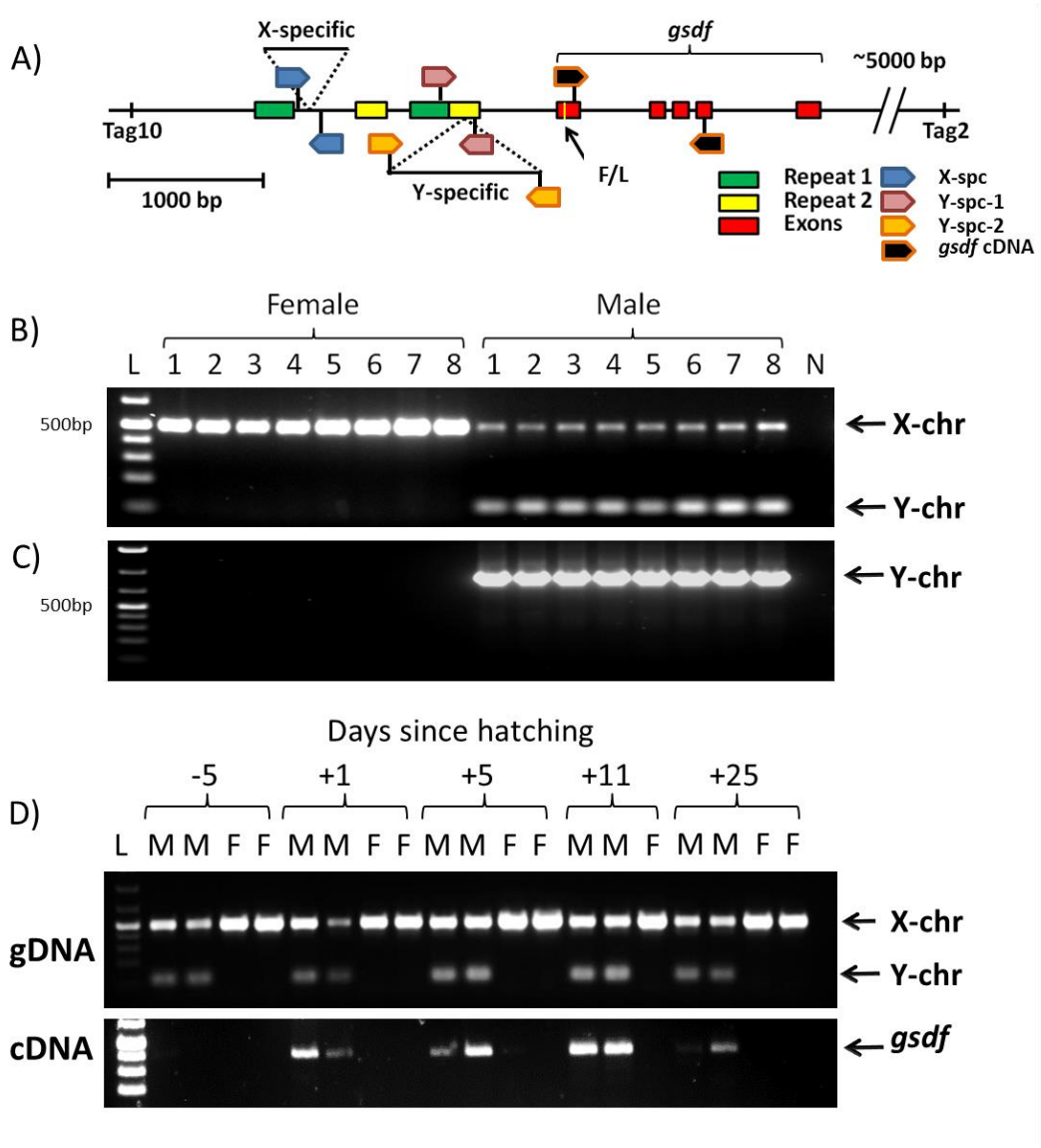


Figure 2.5: Sex-specific sequences in sablefish

A) The major features of the sablefish sex region, including the relative location of the X and Y-specific insertions to the start of the *gsdf* coding region and the location of the subsequent PCR reaction primers; coding change in first exon from Phenylalanine to Leucine marked by F/L B) PCR reaction showing the amplification of a single, larger fragment in 8 unrelated female fish, and two fragments in the male with the Y-fragment 412bp smaller in size C) Nested PCR demonstrating the presence of the Y-specific insertion solely in the males. D) time-series of pre-hatch and post-hatch sablefish showing genetic sex-identification based on the X-specific insertion, and *gsdf* expression in these individuals. For B), C) and D) "L" represent the 1kb generuler plus ladder (ThermoScientific), with the brightest two bands at 500 and 1500bp, "N" represents the negative control.

Portions of the X-specific sequence were identified elsewhere in the sablefish TSA library, but with no strong hits to the adjacent *gsdf* promoter sequences, it is likely that the 412 bp sequence is a repeat sequence or transposable element. Significant BLAST hits for this element to non-sablefish sequences were to EST and WGS sequences of Atlantic cod, *Gadus morhua*, (91-100% max identity). The Y-specific insertion also only produced significant BLAST hits to *G. morhua* sequences, although the Y-specific insertion was not observed in the expressed sablefish sequences. Even though sablefish is commonly called Alaskan black cod, it is not a true cod and is only distantly related to the Gadiformes (last common ancestor ~130-200mya (Setiamarga et al. 2009, Near et al. 2012)). While the mechanism for the observed sequence similarity between this element in the two species remains unknown, others have proposed that parasites such as the trematodes *Schistosoma japonicus* and *Diplostomum* spp. (de Boer et al. 2007), or the sea lamprey *Petromyzon marinus* (Kuraku et al. 2009) may facilitate lateral transposition of transposable elements and other sequences between unrelated host fish species; Atlantic cod and sablefish are geographically separated, but the range of the Pacific cod *Gadus macrocephalus* and sablefish overlap significantly allowing for the possibility of some form of lateral transfer between species.

The presence of the major sex-specific insertions in sablefish immediately upstream of the *gsdf* gene is quite intriguing. As a member of the transforming growth factor-β superfamily, the *gsdf* gene was first described in the proliferation of primordial germ cell and spermatogonial proliferation in rainbow trout (Sawatari et al. 2007). It has since been shown that in the evolutionarily conserved cluster of genes that includes *gsdf*, it is the only gene preferentially expressed in teleostean testis (Gautier et al. 2011a). Further, it has been shown that expression of *gsdf* in the zebrafish occurs prior to gonad differentiation in the early development of zebrafish, and is expressed solely in the gonads (Gautier et al. 2011b).The gene was described as a master-sex determinant in a medaka species, *Oryzias luzonensis* (Myosho et al. 2012). While we do not present

enough evidence to describe *gsdf* as the master-sex determinant in sablefish, the location of the sex-specific insertions (or deletions) in the promoter region immediately upstream of a gene that has been described as a master-sex determinant in another fish species, and coupled with its placement in a cluster of genes otherwise expressed preferentially in ovaries suggests that *gsdf* is the strongest candidate for the master-sex determinant in sablefish. Furthermore, if future research supports *gsdf* as the master-sex determinant in sablefish, it will represent the first description of the independent evolution of the same gene to the role of master-sex determinant in different orders of fish. Independent evolution of DMRT as a sex-determinant has been described in African clawed frogs (Yoshimoto et al. 2010) and chickens (Smith et al. 2009), but in fish, this gene has been described as the main sex-determinant only in related medaka species.

From *Oryzias luzonensis*, it is estimated that sex-specific expression differences in *gsdf* are first observed around hatching (10 days post fertilization [d.p.f.]) , and are no longer detected at 10 days after hatching (Myosho et al. 2012). If the same timing was observed in sablefish, significant male over-expression should be observed between around 15.5 d.p.f. (4°C incubation) (Alderdice et al. 1988) and will no longer be significant 15 days post-hatching (d.p.h.) As shown in figure 2.5d, *gsdf* is overexpressed in males around hatching, although unlike our predictions, it remains overexpressed relative to the females up to 25 d.p.h.; products are not visible following 35 PCR cycles in either sex five days prior to hatching. In 15-month sablefish, no difference is observed in the expression of *gsdf* in ovaries and testis (data not shown). It is possible that differential expression may contribute to the development of each sex, although it remains to be determined what role the amino acid substitution from phenylalanine to leucine may have on the actions of the protein.

Significance of the Genomic Tools

Genomic tools can be applied to many questions relating to fisheries, aquaculture, physiology, ecology, evolution, disease, reproduction, growth, and general biology of sablefish. The resources developed here will be fundamental in the management and analysis of wild and domestic commercial stocks. At a landed value of US $4.46/lb, the sablefish is the second most valuable finfish (after halibut, US $4.97/lb) and represent the fourth greatest total value (after walleye pollock, sockeye salmon, and Pacific halibut, ~equal to Pacific cod) to 2011 Canadian and American commercial fisheries. In aquaculture, values increase up to US $9.00/lb (Dr. Stephen Cross, Personal Communication). The large set of microsatellite and SNP markers, as well as the sequences provided by the complete mitochondrial genome, will allow for a robust molecular analysis of sablefish population structure throughout the natural range to determine the most appropriate population management. Additionally, the ease of genetically sexing both immature and mature sablefish will allow for the easy addition of sex as an additional parameter in analyses of population composition and movement. The availability of expressed sequences should allow a genetic base from which to design studies to gain insight into the physiology and biology of this deep-sea organism.

Many of these markers can likely be used in the analysis of closely related species such as skilfish (*Erilepis zonifer*) (McCraney et al. 2012) and perhaps other Scorpaeniformes, Gasterosteiformes and Perciformes species. The degree of conserved synteny between the threespine stickleback and the sablefish suggests that these tools may be very useful for many species including other economically important species such as lingcod and greenling (Family: Hexagrammidae) and or the more distantly related rockfish (Family: Sebastidae).

The use of these tools will also be helpful in the development of species for aquaculture as well as for conservation and management of wild stocks. A linkage map and the demonstrated ability to use the threespine stickleback genome to determine likely gene location will facilitate molecular analysis and development of markers for

selective breeding. The expressed transcript library will provide access to genes of interest, and the development of sex markers will allow for the study of sex-specific phenotypes. If *gsdf* is confirmed as the master-sex determinant in sablefish, the targeted production of monosex or sterile stocks for use in aquaculture may be possible.

Conclusions

With this work, we present the largest collection of genetic data available for sablefish. A mitochondrial genome sequence was produced and used to place sablefish among Scorpaeniformes and Gasterosteiformes fishes. We have sequenced and assembled transcript library of nearly 35,000 sequences for the sablefish, and identified microsatellite and SNP markers for use in broodstock selection and population management. Using our polymorphic markers, we have produced a linkage map consisting of 24 linkage groups, which agrees with the expected number of chromosomes previously identified through karyotyping. Comparative mapping has been used with *G. aculeatus* to show significant conservation of gene order, allowing for the stickleback genome to be used to predict gene location in sablefish. A genotyping-by-sequencing library was used to identify additional SNPs, adding 1843 markers, and to identify sex-specific markers. Finally, the sex-specific markers led directly to the identification of sex-specific sequences in the sablefish, located in the upstream promoter region of the known sex pathway gene, *gonadal soma-derived factor*, the master sex-determinant in the medaka, *Oryzias luzonensis*.

Methods

Sample Collection and extractions

All sablefish used in this work were provided by Sablefish Canada Ltd. The first family is the result of a paired hatchery mating with a single father, a mother, and 83

progeny. Tissue collection from each parent was through fin clip, while the samples from the progeny are through whole body DNA extractions on samples collected just after hatching, all stored in 95% ethanol until use. The second family, also a result of hatchery mating, produced a half-sibling family from three males and a single female. For each of the parents and the 115 progeny, fin clips (stored in 95% ethanol) were used as a source of DNA. The sex of 100 fish in the second family was recorded by the appearance and size of the gonads at 15 months of age, and liver, spleen, kidney, gonad and head kidney tissues were collected, initially frozen on dry ice, and then at -80°C. DNA was isolated using a Chelex extraction protocol following the protocol of (Walsh et al. 1991). A phenol DNA extraction was performed on the liver tissue of one of the progeny from family 2 following a standard protocol (Sambrook et al. 2000); this extraction was used for mitochondrial genome sequencing. Liver tissues from family 2 progeny and fin tissue from the parents and 14 immature, unrelated sablefish (8 males, 6 females) were extracted for use in Genotyping-by-Sequencing using the standard protocol in the DNeasy Blood & Tissue Kit (Qiagen). For sex-specific sequencing, fin clips from 47 mature, definitively sexed broodstock were extracted by Chelex extraction and used along with the 6 family parents.

Total RNA was extracted from each tissue in TRIzol reagent (Invitrogen) by mixer-mill homogenization (Retsch) and spin-column purified using RNeasy Mini kits (Qiagen). Each RNA sample was then quantified and quality-checked by spectrophotometer (NanoDrop Technologies) and agarose gel, respectively. For cDNA libraries, the brain, gill and kidney were taken from sablefish unrelated to the mapping families and processed. For RNA-seq, RNA was extracted from the kidney, liver and gonadal tissues from one male and one female sablefish from the second mapping family.

For time-series analysis, sablefish were collected 5 days before and 1, 5, 11 and 25 days after hatching, and placed in RNAlater (Invitrogen) until use. Whole fish/developing eggs were digested with Proteinase K (Qiagen); following digestion, 25% of the sample was used with the DNeasy Blood & Tissue Kit (Qiagen) while the

remaining sample was purified using the RNeasy mini kit (Qiagen). RNA samples were converted to cDNA for PCR using M-MuLV Reverse Transcriptase (NEB).

cDNA libraries

Anoplopoma fimbria EST libraries were constructed using the methods of Koop et al. (Koop et al. 2008). In short, the library was constructed directionally in pAL17.3 (Evrogen Co.). Libraries were plated, and colonies picked using a Qpix2 array picker. Following overnight growth of glycerol stocks arrayed in 384-well format (Rise et al. 2004), plasmid DNA was extracted using a standard lysis and neutralization procedure followed by an alcohol precipitation. Sequencing was performed on an ABI 3730 sequencer using BigDye™ Terminator V3.1 (ABI) cycle sequencing kit and either the primer M13 forward (5'-GTAAAACGACGGCCAGT-3') for 5' end or SP6WAN (5'-ATTTAGGTGACACTATAG-3') for 3'end sequencing. Base calling from traces and quality scores assigned using Phred (Ewing and Green 1998, 1998), and Phrap used to assemble the sequences into contigs (200 minscore, 0.99 stringency; http://www.phrap.org/).

One lane of RNA-seq 100bp PET Illumina sequencing was run on a HiSeq2000 sequencer, and assembled using SOAPdenovo (k=75) (Li et al. 2010b); sequencing was performed at the Beijing Genomics Institute.

RNA-seq and EST library data were merged into a Transcriptome Shotgun Assembly. Contigs larger than 200bp from both datasets were assembled using Phrap (200 minscore, 0.96 stringency), with resulting contigs 400bp or larger retained. Contigs were annotated using SwissProt and Gene Ontology's annotated protein databases using a threshold of 1e-5.

Mitochondrial Genome Sequencing

A slightly modified version of Miya and Nishida's protocol (Miya and Nishida 1999) was used to amplify the mitochondrial genome in two long, overlapping

sequences with smaller, secondary PCRs for sequencing. Primers were designed to the 16S ribosomal RNA and cytochrome B mitochondrial sequences in NCBI (accessions GU018112 and FJ264496 respectively). Afim-Cb-H (5'-GATATGAGCCGTAGTAAAGACCTCGGCCGA-3') and Afim-16S-L (5'-TCGACAAGGGGGTTTACGACCTCGATGTTG-3') were designed to the same place as the Gogr-Cb-H and Gogr-16S-L primers and these primers were used to amplify the larger mitochondrial fragment, covering the NADH dehydrogenase, cytochrome oxidase and ATPase genes. Afim-16S-H (5'-GACCTGGATTACTCCGGTCTGAACTCAGAT-3') was also designed to the same place as Gogr-16S-H; however, only the 3' end of the primer was found in the available sequence, and the 5' end of the designed primer comes directly from Gogr-16S-H. Afim-Cb-L (5'-GATTAATCCGAAACATTCACGCTAACGGTG-3') was designed to a different portion of the Cytochrome B sequence, as Gogr-Cb-L did not align to the available Cytochrome B sequence. These two primers were used to amplify the small mitochondrial fragment, which contained the 12S ribosomal RNA and the D-loop. Both reactions were amplified with the Phire Hot Start II DNA polymerase (Finnzymes); 1x Phire reaction buffer, 0.2uM of each dNTP (Promega), 0.5uM of each primer (IDT), 50ng genomic DNA and 0.5ul enzyme in 25ul reactions. Reactions were loaded onto a Techne TC-412 at 98°C for 45 seconds, followed by 35 cycles of 98°C for 8 seconds, 63°C for 10 seconds and 72°C for 4 minutes. Following a final extension of 72°C for 8 minutes, samples were cooled to 4°C. After confirming successful amplification on a 1% agarose gel, the reactions were diluted 1/100, and used as templates for the secondary PCRs. The 30 pairs of fish-universal primers (Miya and Nishida 1999) and the 4 long primers were used to amplify small pieces of the mitochondrial genome in an overlapping fashion; when a pair could not successfully amplify a product (e.g. 5-L/5-H), an adjacent primer was used instead (e.g. 5-L/6-H); long primers were used when overlap between the two large fragments could not be obtained. 25ul reactions were used, containing 1x GoTaq Flexi Buffer (Promega), 2.5mM MgCl2(Promega), 0.2uM each dNTP (Promega), 0.5uM each primer (IDT) and 0.625U Hot Start GoTaq polymerase

(Promega). Samples were amplified on Techne TC-412 at 95°C for 3 minutes, followed by 30 cycles of 95°C for 15 seconds, 45°C for 15 seconds and 72°C for 45 seconds, with a final extension of 10 minutes at 72°C, and a final hold of 4°C. The PCR products were plate purified with Qiagen MinElute 96UF PCR purification; samples were eluted in 20ul DNase/RNase free water. Sequencing reactions were prepared containing 0.5 µl BigDye Terminator v3.1 (ABI), 0.5ul BigDye Terminator sequencing buffer, 0.64 µM forward or reverse primer and approximately 20ng purified PCR product in a 5ul reaction; these were run on a TC-412 (Techne) as follows: 1 min 95°C initial denaturation, followed by 30 cycles of 95°C for 30s, 50°C for 15s, and 60°C for 90s and a final extension of 5min at 72°C. Sequencing reactions were ethanol precipitated and re-suspended in 20 µl DNase/RNase free H_2O (Gibco). All sequencing was performed on an ABI 3730 DNA analyzer. Resulting sequences were assembled using the assembler in Geneious v5.1.7. Mitochondrial features were identified using DOGMA (Wyman et al. 2004) and by comparison to the available mitochondrial genomes of other Scorpaeniformes: *Aptocyclus ventricosus* (AP004443), *Clinocottus analis* (FJ848374), *Cottus hangiongensis* (EU332751), *Cottus poecilopus* (EU332750), *Cottus reinii* (AP004442), *Helicolenus hilgendorfi* (AP002948), *Satyichthys amiscus* (AP004441), *Sebastes schlegeli* (AY491978), *Sebastiscus marmoratus* (GU452728).

For phylogenetic analysis, the RAXML (7.0.4) protocol of Kawahara et al. (Kawahara et al. 2008) was followed. Along with 71 of the 75 mitogenomes previously analyzed, we added the 15 mitogenomes most recently uploaded to NCBI that were expected to be closely related to the Scorpaeniformes (see Additional file B8 for the list), as well as the sablefish mitogenome. Following the trimming of ambiguously aligned sections, gaps and loop structures, we focused on 13528 bases : 10436 in the 12 protein-coding genes (excludes ND6, see (Kawahara et al. 2009)), 1411 bases in the 22 tRNAs, and 1681 in the two rRNA sequences. For the 12 protein-coding genes, we re-assigned the third position in each codon as either purine (R) or pyrimidine (Y) as deemed the best estimate and conservation of signal to noise (Kawahara et al. 2008).

Primer Development and Marker Identification - SNPs

SNP primer design was performed with three different approaches. For the first batch, assembled contigs were analyzed for SNP variation in individual ESTs. Using primer3 (Rozen and Skaletsky 2000), primers amplifying products of 200-400bp containing the expected SNP were designed, with primers of 18-22 bp with $T_m \approx 55^oC$, a max 5^oC difference in T_m and a 2bp GC clamp. In the second batch, primers were designed to a random selection of EST contigs, with a focus on the 3' UTR. Again, using primer3, primers amplifying larger products, 400-550bp, were designed, with the same characteristics as above. Finally, one set of primers was designed to target each of *dmrt1, cdk5rap2*, myostatin, growth hormone receptor, growth hormone receptor hormone, SSR2, TGF2, DAX1, and SOX9 loci. Alignment of the loci from the species *Gasterosteus aculeatus*, *Oryzias latipes, Takifugu rubripes*, and *Tetraodon nigroviridis* were visualized using the UCSC genome browser and primers were designed to highly conserved sequence in the four species. PCR and sequencing reactions were performed using the SNP amplification and sequencing protocol in Messmer et al. (Messmer et al. 2011); all sequences were obtained using an ABI 3730 DNA analyzer. Sequences from each primer pair were aligned in Geneious 5.1.7 and SNPs identified and scored manually.

Primer Development and Marker Identification – Microsatellites

Using the program RepeatFinder (Benson 1999), repeats of 2-5 nucleotides length (min 4 repeats) were identified from the assembled EST transcripts. Using primer3, primers were designed to amplify products 75 – 200bp in size encompassing the expected repeat, with primers of 18-22 bp with $T_m \approx 55^oC$ and a max 5^oC difference in T_m. An additional 13 primer pairs designed from the same dataset were used, and are

described in (Consortium Molecular Ecology Resources Primer Development et al. 2011).

Microsatellite genotyping followed the amplification and scoring protocol in Messmer et al. (Messmer et al. 2011). In initial primer testing, successful amplification and polymorphism, were determined using 10% polyacrylamide gels (25ml of 40% 19:1 acrylamide:bis-acrylamide (BioRad); 100ul TEMED (Sigma); 1ml 10% w/v ammonium persulphate (Sigma); 74ml H_2O) run for 12-14 hours, stained using EtBr and imaged; half the microsatellites in family 1 were also analyzed by this method. The remaining microsatellites scored in family 1 and all microsatellites scored in family 2 included either a 6-FAM or HEX labelled fluorescent dye (IDT) attached to one of the primers. Following amplification, 0.5ul each PCR product (one HEX-labelled and one 6-FAM labelled) was added to 9.9ul Hi-Di™ Formamide (ABI) and 0.1 µl GeneScan™ -500 ROX™ Size Standard (ABI), and samples denatured by heating to 95 ^{o}C for 3 min and placed on ice for 5min. Amplifications were run on an ABI 3730 DNA sequencer, and electropherograms were analyzed using GeneMapper V4.0 (ABI).

Marker Naming

Nomenclature for newly developed markers follows a modified version of the microsatellite nomenclature of Jackson et al. (Jackson et al. 1998). The species was identified by the first letter of the genus followed by 2 of the species (Afi); marker type was represented by a two-letter identifier (MI = microsatellite, SP = SNP); a unique four-digit number was assigned to each of the individual markers; a four-letter identifier for the institute from which they were found (UVic).

With both microsatellite and SNPs, any occasion where multiple markers were utilized from the same locus, the markers were given the locus name for mapping purposes. Locus names were assigned with a short identifier of the marker type (SNP or MI) followed by the word locus and a unique contig number (see Additional files B2 and B3 for contig numbers). Furthermore, should two markers from the same contig be

useful in the same mapping parent, each marker was used to create a representative locus marker, with one marker filling in any missing data from the other. Each contig sequence used to design the SNP or microsatellite primers was input into a NCBI BLASTx search. Should the contig have a significant BLAST result, an abbreviation for the protein product was enclosed in square brackets and joined to the original name (either marker or locus as necessary) by an underscore.

Linkage Analysis

Analysis of linkage in this study was performed using programs contained within the LINKMFEX package, v 2.3 (R. Danzmann, University of Guelph, http://www.uoguelph.ca/~rdanzman/software/LINKMFEX) following the standard protocol (LOD threshold = 3.0). Markers with more than 15% of genotypes missing were omitted from the analysis. Merged sex-specific maps were produced followed by an estimated merged map using the MERGE programs in the LINKMFEX package. Should an insufficient number of common markers be mapped to linkage groups to be merged, the more complete linkage group was taken as a representative for the sex. Maps were visualized using the program MAPCHART (Voorrips 2002).

Anoplopoma fimbria vs *Gasterosteus aculeatus* (threespine stickleback)

Each of the contig sequences (minus extended repeats) used to design the primers for the markers linked to the map were input into the UCSC Genome Bioinformatics DNA Blat server (http://genome.ucsc.edu/ bin/hgBlat?command=start;(Kent et al. 2002, Fujita et al. 2011)) using the stickleback Feb. 2006 (Broad/gasAcu1) assembly. The hit with the highest score (min = 100, default BLAT score) was determined to be the most likely ortholog in the threespine stickleback, and chromosome number and position along the chromosome were retained. If contig

sequences showed no BLAT hits, contigs were mapped with BLASTn to TSA database, and the top TSA hit BLATed against the *G. aculeatus* genome (min score = 150).

Genotyping-by-sequencing (GBS) Library

The GBS library was prepared following the protocol of Elshire et al. (Elshire et al. 2011) utilizing the *Ape*KI restriction enzyme and the 96 barcodes suggested. 2 lanes of the sample were run on a HiSeq 2000 with 100bp paired-end reads (Illumina) and 48 individuals per lane. Resulting sequences were trimmed and markers were scored using the Stacks package (Catchen et al. 2011); see Additional file B7 for details. JELLYFISH v1.1 (Marçais and Kingsford 2011) was used to identify and count all 31mers in the dataset. The 31mers were searched for sequences found exclusively in males (family males >70% occurrence; unrelated males >60%; females <2) and females; the resulting kmers were mapped back to full-length reads and used to identify the potential male or female-specific polymorphisms.

Genome Walking

Genome walking was used to expand the available sequence around the male-specific markers, as well as the gene predicted to be between the male-specific markers using a protocol based on Siebert et al. (Siebert et al. 1995) with primers from Rebrikov et al. (Rebrikov et al. 2004). Following three separate digestions of male and female sablefish DNA with restriction enzymes EcoRI, NdeI and BbsI (NEB), the genome walking adapter was ligated overnight at 16°C using T4 DNA ligase (NEB) in a 10µl reaction. Following inactivation and dilution (1/10), the samples underwent PCR with primer P1 (5'-CTAATACGACTCACTATAGGGC-3') and a specific primer designed to the 100bp read of interest (Additional file B9); PCR mix of 1X GoTaq Flexi Colorless PCR buffer (Promega), 2.5mM $MgCl_2$ (Promega), 320µM each dNTP (Promega), 0.5 µM each forward and reverse primers (IDT), 0.75U GoTaq Hot Start polymerase (Promega), and

5ng of DNA template made up to 30µl with DNase/RNase free H_2O (Gibco), and cycled 95°C for 3min (without template), 72°C for 10minutes (including template), 21 cycles of 95C for 30s, 62°C -0.5°C per cycle for 30s and 72°C for 4min, 19 cycles of 95°C for 30s, 52°C for 30s, 72°C for 4min, with a final extension of 72°C for 10mins. PCR was diluted 1/10000 and used as a template for second round using nested primer NP1 (5'-TCGAGCGGCCGCCCGGGCAGGT-3') and a nested specific primer. PCR was cycled as above without the initial 72°C for 10 minutes. Following Exo/FastAP treatment (Fermentas), samples were sequenced following the SNP sequencing protocol (Messmer et al. 2011). Sequencing reactions were purified by ethanol precipitation while sequencing was performed on ABI 3730, and data analyzed using Geneious V5.1.7. Resulting sequences were BLATed against the *Gasterosteus aculeatus* genome.

Identification of Sex-specific markers and sequences

All primer sequences used in the identification of Sex-specific markers can be found in Additional file B9. PCR primers were designed to sequences containing sex-linked polymorphisms, as well as genes predicted to be adjacent and between these markers based on the stickleback BLAT. PCRs were performed to amplify the sequence between the markers and genes, looking for sex-specific amplicons; long PCRs were performed following the Phusion (NEB) protocol with 35 cycles and 15s/kb. Once sex-specific size differences were discovered, primers were designed to fully amplify the region, with amplification and sequencing performed following Genome Walking PCR protocol, with 2min extensions. Multiple bands were purified by gel extraction using the QIAquick Gel extraction kit (Qiagen). All 53 unrelated, mature sablefish were sequenced for the observed DNA inserts to confirm sex-specificity. Time-series samples were sexed using the X-insertion primers using the genomic DNA as a template, while amplification on cDNA samples utilized exon-specific primers; see Additional file B9 for sequences and reaction conditions.

Acknowledgements

This work was supported by the University of Victoria, GenomeBC, Sablefish Canada Ltd., and the Natural Sciences and Engineering Research Council of Canada.

Supporting Data

The data sets supporting the results of this article are available in NCBI where accession numbers are presented All other supporting information is available as supporting data within the article (Appendix B)

Chapter 3 :

The genome and linkage map of the northern pike (*Esox lucius*): Conserved synteny revealed between the salmonid sister group and the Neoteleostei

Abstract

The northern pike is the most frequently studied member of the Esociformes, the closest order to the diverse and economically important Salmoniformes. The ancestor of all salmonids purportedly experienced a whole-genome duplication (WGD) event, making salmonid species ideal for studying the early impacts of genome duplication while complicating their use in wider analyses of teleost evolution. Studies suggest that the Esociformes diverged from the salmonid lineage prior to the WGD, supporting the use of northern pike as a pre-duplication outgroup. Here we present the first genome assembly, reference transcriptome and linkage map for northern pike, and evaluate the suitability of this species to provide a representative pre-duplication genome for future studies of salmonid and teleost evolution. The northern pike genome sequence is composed of 94,267 contigs (N50=16,909 bp) contained in 5,688 scaffolds (N50=700,535 bp); the total scaffolded genome size is 878 million bases. Multiple lines of evidence suggest that over 96% of the protein-coding genome is present in the genome assembly. The reference transcriptome was constructed from 13 tissues and contains 38,696 transcripts, which are accompanied by normalized expression data in all tissues. Gene-prediction analysis produced a total of 19,601 northern pike-specific gene models. The first-generation linkage map identifies 25 linkage groups, in agreement with northern pike's diploid karyotype of 2N=50, and facilitates the placement of 46% of

assembled bases onto linkage groups. Analyses reveal a high degree of conserved synteny between northern pike and other model teleost genomes. While conservation of gene order is limited to smaller syntenic blocks, the wider conservation of genome organization implies the northern pike exhibits a suitable approximation of a non-duplicated Protacanthopterygiian genome. This dataset will facilitate future studies of esocid biology and empower ongoing examinations of the Atlantic salmon and rainbow trout genomes by facilitating their comparison with other major teleost groups.

Background

Of the nine currently available teleost genome assemblies (www.ensembl.org; April 2014), all fall into either the Neoteleostei or the Ostariophysi (Betancur-R et al. 2013). This is not necessarily surprising as combined these two clades comprise more than 70% of the approximately 26,840 extant teleost species (Nelson 2006). While not as large, other groups contain many commercially important species and thus receive a significant amount of scientific attention; this observation is particularly true of the Protacanthopterygii (the salmonids, esocids and marine smelts) (Nelson 2006). The Protacanthopterygii are placed in an interesting evolutionary position, having diverged from the Neoteleostei approximately 200 million years ago (MYA) (Betancur-R et al. 2013) following an even more ancestral split from the lineage containing the Ostariophysi (250 MYA; see Figure 3.1). Thus, the generation of a genome sequence representing the Protacanthopterygii provides an intermediate phylogenetic branch of significant scientific interest.

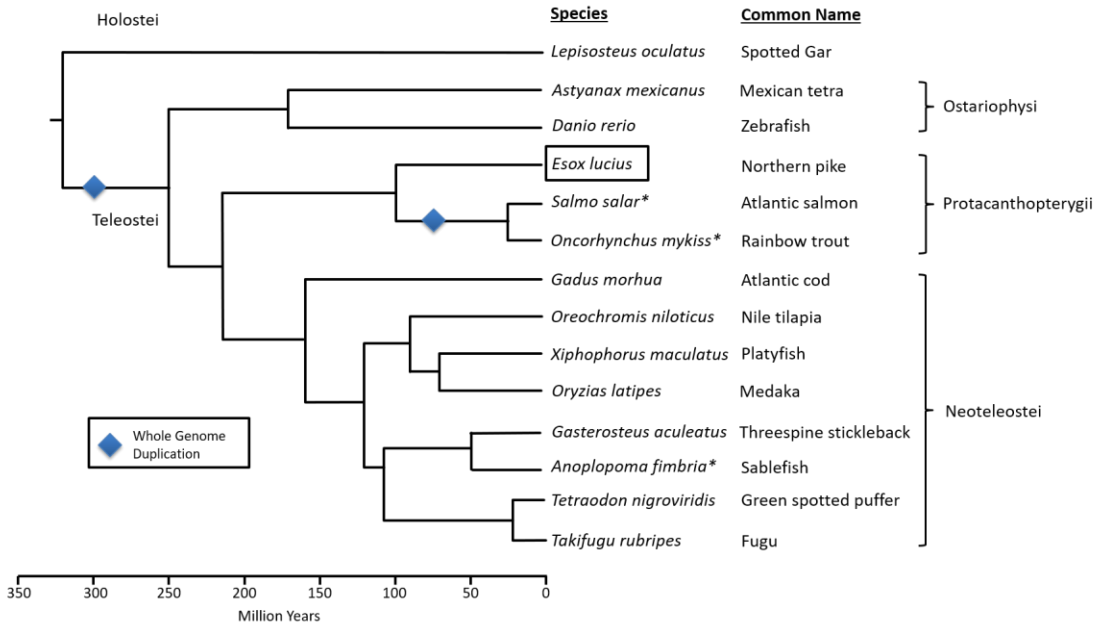


Figure 3.1: The phylogenetic relationship of teleosts with publicly available genome assemblies, as well as select species with ongoing genome projects (*)
Divergence times and Subcohort naming based on Betancur et al. (Betancur-R et al. 2013). Diamond shapes represent presumed whole-genome duplications. *Lepisosteus oculatus* (Spotted gar) used as an outgroup.

While not conclusively resolved (Li et al. 2010a, Near et al. 2012, Zou et al. 2012, Betancur-R et al. 2013, Campbell et al. 2013) the Protacanthopterygii generally include the orders Salmoniformes, Esociformes (pikes and mudminnows) and Argentiniformes (marine smelts). The Salmoniformes (grayling, ciscoes, whitefish, trout, char, and salmon) are the most studied of the Protacanthopterygii, and nearly all constituent species have significant economic, ecological, conservation, environmental and societal value (Thorgaard et al. 2002, Davidson et al. 2010). In the last 60 years, over 70,000 publications (Davidson et al. 2010) have examined fundamental and applied questions relating to their fisheries, aquaculture, physiology, ecology, evolution, disease, reproduction, growth, tolerance to physical factors, and general immunity.

The Salmoniformes is a rapidly evolving order, with most species exhibiting in excess of 92% similarity in the coding DNA (Leong et al. 2010). The biological diversity and complexity observed within salmonids is undoubtedly a reflection of their complex genomes. While two rounds of whole-genome duplication (WGD) are thought to have preceded the origin of vertebrates and a third round (3R) is believed to have occurred in basal teleosts (Taylor et al. 2003, Christoffels et al. 2004, Jaillon et al. 2004, Vandepoele et al. 2004, Blomme et al. 2006), the common ancestor of salmonids has purportedly experienced an additional, fourth round (4R) WGD approximately 60-100 MYA (Allendorf and Thorgaard 1984, Thorgaard et al. 2002, Shiina et al. 2005, Davidson et al. 2010, Macqueen and Johnston 2014). Modern salmonids are considered pseudotetraploid, and various independent re-diploidizations and genome re-stabilizations have occurred in different salmonid lineages as they revert to a stable diploid state (Allendorf and Thorgaard 1984, Thorgaard et al. 2002, Shiina et al. 2005, Davidson et al. 2010). Although this recent duplication makes the salmonids difficult to use in the analysis of major teleost lineage evolution, it does make them ideal organisms for examining some of the early impacts of gene and genome duplications. These processes are thought to have played pivotal roles in promoting increased genetic diversity, the development of novel functions and functional specialization. The way in which a genome is reorganized to cope with duplicated chromosomes and the importance of gene duplications for evolution and adaptation represent long-standing biological questions that remain unresolved (Sémon and Wolfe 2007, Conant and Wolfe 2008, Van de Peer et al. 2009, Davidson et al. 2010, Dittmar and Liberles 2010). Additionally, the role of repetitive DNA in facilitating post-duplication genomic reorganization is also of increasing interest in light of the recent expansions of transposable elements which have been identified within salmonids (de Boer et al. 2007). The salmonid-specific WGD, the resulting pseudotetraploid genomic state, and recent waves of transposable element activity all contribute to the increased complexity of relating genetic mechanisms to physical characteristics in salmonid species. The

presence of duplicated genes has made it more difficult to clearly identify genetic mechanisms underlying physiological, immunological, ecological and evolutionary traits within the commercially and culturally valuable salmonids. Presently, the International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG) represents researchers, funding agencies and industry from Canada, Chile and Norway. It has the goal of generating a robust salmonid reference genome sequence that facilitates the study of all salmonids (Davidson et al. 2010). This work in Atlantic salmon (*Salmo salar*) (Davidson et al. 2010) is ongoing. During the review of this work, the genome assembly for the rainbow trout (*Oncorhynchus mykiss*) was released (Berthelot et al. 2014) providing an additional dataset with which to analyse the results of genome duplication in the salmonids.

The genetic diversity in the Protacanthopterygii is reflected in their variable genome sizes and highly variable karyotypes that frequently differ from the presumed ancestral karyotype of 48 or 50 chromosomes (Mank and Avise 2006). This primitive karyotype is seen in a few members, including within two genera of the Esociformes. Total DNA content comparisons (Gregory 2005) and several gene studies (Koop et al. 2008, Leong et al. 2010) suggest that Esociformes diverged from Salmoniformes prior to the salmonid WGD and that a genome of the Esociformes may closely approximate the ancestral genomic state. Therefore, it is important to determine how closely the organization of a representative Esociformes genome may reflect that of the ancestral Protacanthopterygiian. This information will help determine if comparisons to the Esociformes are appropriate in helping to resolve patterns of genome recombination and changes in gene expression that are implicated in the re-diploidization of the salmonids.

Esociformes, an old lineage with fossils dating back to 70 MYA, contains a limited number of species (~13) of which most literature concerns the northern pike (*Esox lucius*). The karyotype for the *Esox* genus is 2N = 50 (NF = 50) (Beamish et al. 1971, Ráb and Crossman 1994), similar to the presumed karyotype of the teleost ancestor (2N =

48-50) (Mank and Avise 2006). This compares to the larger Salmoniformes karyotypes, which range from 2N = 52 (*Oncorhynchus gorbuscha*) to 2N = 102 (*Thymallus thymallus*) (Gregory 2005). Esociformes outside the *Esox* genus have varying karyotypes of 2N = 22-78, with only *Novumbra hubbsi* exhibiting an ancestral-like diploid chromosome number of 2N=48 (although the fundamental number indicates there are 62 chromosomal arms, suggesting significant rearrangement may have occurred (Crossman and Ráb 2001)). A previous assessment of the suitability of northern pike as an outgroup for the study of duplicated genes in the salmonids examined 408 genes in Atlantic salmon (*Salmo salar*) and *E. Lucius* (Leong et al. 2010) and obtained results consistent with northern pike having a pre-WGD genome and being much more closely related to salmon than to other teleost species (zebrafish, medaka, fugu or stickleback). Such data suggest that the esocids provide an important intermediate comparison between salmonids and other more distantly-related fish genome sequences. For the present study, *E. lucius* was chosen to represent Esociformes because of its availability, general interest and literature support.

The northern pike is a northern hemisphere, circumpolar, freshwater species that is typically the top predator in many lakes (Craig 2008, Nilsson et al. 2008). It is an important northern recreational fishery and forms the basis of small commercial fisheries in some locations (FAO 1988). Northern pike are morphologically diverse and highly adaptable, they play a major role in structuring freshwater communities and are used in stocking programs to improve water quality (FAO 1988). Despite their wide distribution, northern pike populations have very low levels of polymorphism and divergence (Nilsson et al. 2008), indicative of a recent common ancestor. Genetic studies suggest that 1-3 post-glacial refugia in North America and several significant refugia in Europe (Miller and Senanan 2003, Craig 2008) contributed to the reduced levels of genetic diversity within this group.

To complement ongoing genome studies of Atlantic salmon and rainbow trout, we sought to develop an extensive set of genomic resources for northern pike. To this

end, we have sequenced a comprehensive set of northern pike transcripts and characterized their expression levels in various tissues. Further, we have produced a genome assembly that establishes the structure of genes, gene order, and repeat element content in northern pike. We have constructed a first-generation linkage map using highly variable microsatellite markers and mapped the relative position of both genomic scaffolds and expressed transcripts to linkage groups. Finally, we have examined the syntenic relationships between this representative esocid and distantly-related model organisms within the Neoteleostei and Ostariophysi to determine if northern pike represent an appropriate model of a pre-duplication genome for use in the analysis of the salmonids. This work presents a significant base for future studies of northern pike and provides the basis for invaluable comparisons in future salmonid studies, particularly once salmonid genomes become available.

Results and Discussion

A high-quality genome assembly for *Esox lucius*

Libraries from 180 bp DNA fragments were Illumina paired-end sequenced (465.9 million 100 bp overlapping reads, ~50X genome coverage) (Michael Smith Genome Sciences Centre, Vancouver, BC, CANADA). Libraries from 2 kb (448 million reads, ~45X coverage) and 5 kb fragments (59.7 million reads, ~6X coverage) were mate-pair sequenced (BGI). Assembly using ALLPATHS-LG (Gnerre et al. 2011) produced 94,275 contigs with 823,910,316 bp. Subsequent filtering (vector, duplicates, mitochondrial sequences) produced a finalized dataset comprising 823,673,596 bp in 94,267 contigs (mean length = 8,738, median length = 4,729, contig N50 = 16,909). Contigs were further joined by ALLPATHS-LG to include the estimated gaps between contigs based on mate-pair data, which produced 5,688 scaffolds with a total scaffolded genome size of 877,777,613 bp. The longest scaffold was 5.1 Mb, and the overall scaffold assembly had an N50 of 700,535 bp. In all, 199 scaffolds were over 1 Mb in

length, and 1,600 scaffolds were over 100,000 bp. These sequences have been uploaded to NCBI under bioproject PRJNA221548, accession GenBank: AZJR00000000, as well as to web.uvic.ca/grasp/pike in FASTA format. As the genome assembly improves, updates to the assembly will continue to be provided on the website. Major genome assembly statistics are summarized in Table 3.1.

Table 3.1: *Esox lucius* genome assembly statistics.

Minimum contig and scaffold size ≥200bp.

	Number (≥200bp)	**Bases (Mb)**	**N50 (Sequence length)**	**N50 (number of sequences)**	**Maximum Length (bp)**
Contigs	94,266	824	16,910	13,483	232,364
Scaffolds	5,688	878	700,535	318	5,140,982

The most recent figures based on the c-value (Vinogradov 1998, Hardie and Hebert 2003, 2004) estimate the *Esox lucius* haploid genome size to be approximately 1.09-1.12 Gb; if correct, our scaffold assembly covers approximately 78-79% of the genome. As there are known or suspected biases in next-generation sequencing (NGS) library construction (see (Poptsova et al. 2015) for example), it is possible that a small portion of the genome is missing from the sequencing libraries constructed. Published genome size estimates, however, range from 0.83-1.37 Gb (Gregory 2005), implying scaffold coverage could range anywhere from 64-100% of the genome. Therefore, because of the wide range of genome size estimates, a further estimate of genome completeness based on published EST transcript coverage (Leong et al. 2010) was performed. Of 11,662 EST contig sequences considered, 11,263 (96.6%) were successfully mapped to the scaffold assembly. Further, ALLPATHS-LG reported an estimated genome size of 911,413,085 bp based on the kmer distribution of raw

genomic reads (k=25), or 96.3% coverage. Together, these results suggest that the assembly represents a very large percentage of the genome.

While we cautiously expect that some chimeric scaffolds occur within our assembly, none were observed in our analysis of the largest 50 scaffolds (see synteny section below). Therefore, this initial version of the *E. lucius* genomic sequence provides an excellent framework to assess gene structure, gene content and general genome organization. The assembly statistics for both contigs and scaffolds are in-line with recently published fish genomes sequenced using NGS technologies, such as Atlantic cod (*Gadus morhua*; Contig N50 = 2,311 bp, Scaffold N50 = 393,166 bp) (Star et al. 2011) and African coelacanth (*Latimeria chalumnae*; Contig N50 = 12,671 bp, Scaffold N50 = 924,513 bp) (Amemiya et al. 2013). This assembly is the first genome assembly for a non-salmonid member of the Protacanthopterygii and represents an important evolutionary bridge between the two largest teleost subcohorts, Neoteleostei and Ostariophysi.

The genome sequence was examined for low-complexity repeat and transposable element content. This analysis allowed for the identification of major transposable element families present in the genome, and more importantly, provides a consensus library of transposable element sequences for use in filtering non-genic RNA sequences. An estimated 18.1% of the genome sequence is derived from transposable elements. This repetitive fraction is dominated by class II elements of the Tc1-Mariner superfamily (at least 9.42% of the genome is derived from Tc1-Mariner elements). The genomic abundance of observed transposable element taxa is reported in Additional File C3. An additional 2.0% of the genome consists of low-complexity sequences for a total of 20.1 % of the genome annotated as repetitive sequence. Additional studies are currently underway to more thoroughly profile the transposable element sequences in *E. lucius* and to compare them to those of the Atlantic salmon. As a relative observation, 20% repeat-derived genome content is within the range represented in other teleost genomes; while repeat elements represent ≥50% of the well-characterized zebrafish

genome (Howe et al. 2014), only 25% of the stickleback (Jones et al. 2012) and Atlantic cod genomes (Star et al. 2011), 17.5% of the medaka genome (Kasahara et al. 2007), and <5% fugu genome (Aparicio et al. 2002) are TE-derived.

Transcriptome assembly and Gene Identification

A total of 677,321,182 100 bp of reads were obtained from 13 different tissues of a one-year-old individual and assembled into 413,679 contigs using the Trinity assembler (Grabherr et al. 2011). Potential gene candidates were identified from the Trinity raw assembly. Those that were potential full-length annotated genes had homology to a UniProtKB/SwissProt or Gene Ontology protein, or those that possessed an unannotated ORF ≥ 300 bp were retained. All potential gene candidates were further filtered for the longest unique, non-overlapping location in the genome sequence. To reduce isoforms, alleles, recent duplicates, and possible sequencing errors, the potential gene candidates were screened for redundancy (≥ 98%, ≥ 300 bp). 38,696 potential gene candidates are represented in this curated non-redundant set, comprised of full-length annotations (candidates=9,553), UniProtKB/SwissProt homology (candidates=14,538), Gene Ontology homology (candidates=604), and unannotated ORFs ≥ 300 bp (candidates=14,001). These gene candidates represent our *E. lucius* reference transcriptome set. These sequences have been submitted to NCBI under project accession GenBank: GATF00000000; the assembly described in this work is the first version, GenBank: GATF01000000.

An additional dataset incorporating all available transcriptome (including ESTs) and genome sequence data was prepared using the gene-modelling program MAKER2 to produce an *ab initio* dataset of 19,601 putative transcripts. It can be noted that the number of transcripts predicted by modelling is less than those from the de novo assembly. In some instances, this is a result of predicting a single gene model that spans multiple, non-overlapping, RNA-seq contigs. In other cases, it appears to be due to limits in the data used to predict the models such as the size of the genomic contigs used; the

short sequences limit the ability to model a transcript successfully. As the genome assembly is improved, so should the gene-modelling results. Datasets have thus been made available for download at web.uvic.ca/grasp/pike, and any future updates will be distributed here as well; the current MAKER2 dataset is additionally presented as Additional File C10.

With previous EST studies and the current RNA-seq data, we attempt to provide a comprehensive view of the number and identity of genes that make up the genome of *Esox lucius*. In combination with transcriptome data from other fish species, these gene lists will help facilitate future estimates of the identity and number of genes in the ancestral Protacanthopterygii lineage, before the whole-genome duplication occurred in the salmonid ancestor.

A snapshot of tissue-specific gene expression

To determine the expression levels of hypothesized genes in each of the 13 tissues examined reads from each tissue library were mapped back to the reference transcriptome and normalized expression values (fragments per kilobase of transcript per million mapped reads = FPKM) were determined for each sequence. All FPKM values from the non-redundant protein-coding transcriptome are presented in Additional File C4, while results based on the MAKER2 transcriptome can be found in Additional File C5. The primary intent of this initial characterization of gene expression is to provide a basic resource that can empower future work in *E. lucius* while also providing a baseline comparison for research in salmonids that seeks to understand the impact of gene duplication on transcriptional regulation and gene network organization.

To illustrate the general utility of gene expression values, the genetic complexity of the different tissues was investigated by analysing the magnitude and specialization of expression in each tissue. We discuss results based on the Trinity-assembled transcriptome, although findings using the MAKER2 transcriptome generally reflect the same trends. Unsurprisingly, the most common highly-expressed genes are NADH-

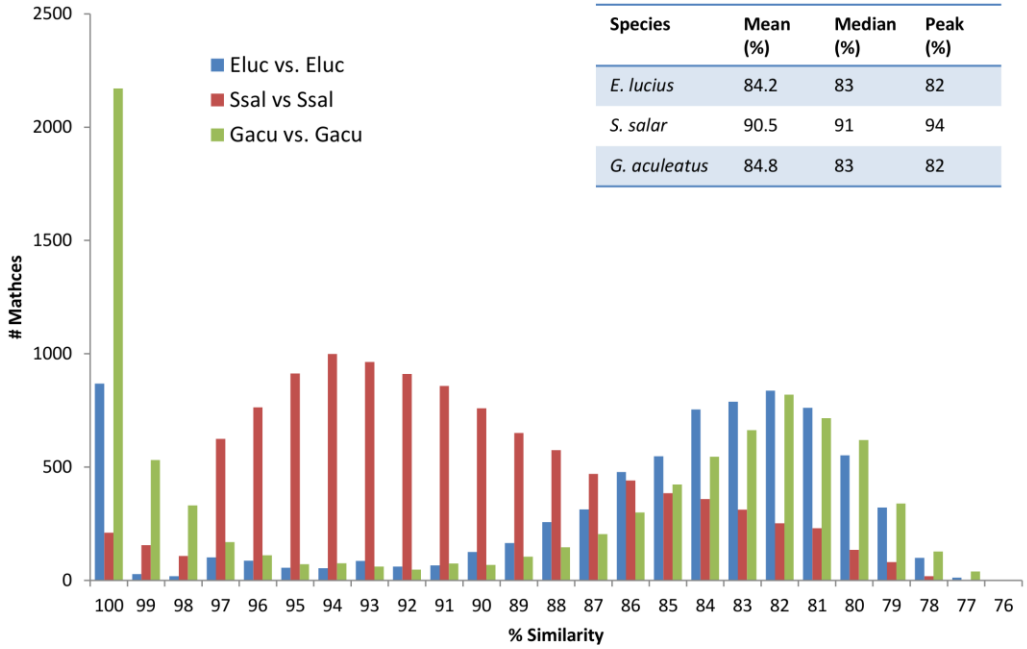
ubiquinone oxidoreductase, actin, cytochrome c and the ribosomal protein-coding genes (Additional File C4); these occur within the top 10 highest-expressed genes in almost all tissues. We also looked at specialized expression patterns for each gene by identifying any tissues in which FPKM values were more than three standard deviations greater than the average across all tissues. The number of specialized transcripts was counted for each tissue (Additional File C6) and the order of tissues by decreasing complexity is: brain (5,408 transcripts), testis (2,784 transcripts), spleen (1,100 transcripts), eye (1,055 transcripts), nose (804 transcripts), gut (659 transcripts), heart (630 transcripts), muscle (430 transcripts), kidney (406 transcripts), gill (374 transcripts), head kidney (338 transcripts), liver (332 transcripts) and stomach (140 transcripts). The brain represents the most complex/specialized tissue with the largest number of genes expressed and the largest number of specialized genes. We found that the testis was the second-most complex and specialized tissue, a result that is in accord with the work of others (He et al. 2012). The spleen, eye and nose regions have the next highest number of specialized genes expressed.

In examining the ten most specialized genes in each tissue (Additional File C6) it was found that chymosin, gastricsin, and pepsin were highly expressed in the stomach tissue, rhodopsin and gamma-crystallin in the eye, immunoglobulins in the head kidney and complement in the spleen, genes often associated with these tissues. The muscle was the least complex tissue in terms of the number of transcripts expressed. However the stomach tissue expressed the smallest number of specialized genes. Perhaps one surprise was the elevated level of ice-structuring protein expressed in the gut. Though antifreeze protein (AFP) gene expression is known to occur in many tissues, it is typically produced in the liver and transferred to the blood (Fletcher et al. 2001). However, as the gut is a possible route whereby nucleating ice might enter the fish, the finding of high transcription of AFP in the gut is informative and may help to explain some of the protective physiology of survival in high northern climates.

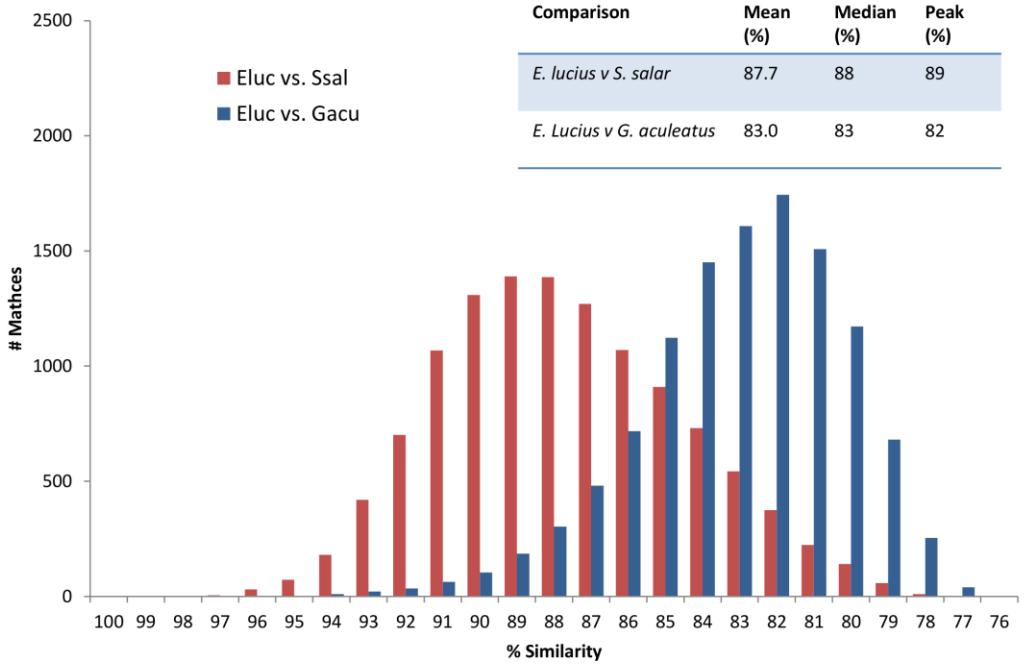
Documenting the expression values for each of the transcripts from the different tissues provides a single snapshot of expression in one set of conditions and is therefore bound to vary in different circumstances. This snapshot does, however, provide a basal atlas of gene expression estimates that can be used by the fish community for many experimental and comparative purposes.

E. lucius homologues suggest a pre-duplication genome

To investigate patterns of gene duplication in *E. lucius,* we plotted the percent similarity of transcript homologues between and within northern pike, Atlantic salmon and threespine stickleback (*Gasterosteus aculeatus*) (Figure 3.2). To facilitate this analysis, three transcript datasets were used: 1) 38,696 reference transcripts identified within our RNA-seq assembly; 2) previously published Atlantic salmon EST data (Koop et al. 2008, Leong et al. 2010), and; 3) *G. aculeatus* transcripts available from the UCSC genome browser. Transcript homologues were identified within and between datasets using reciprocal best BLASTn hits following a similar approach previously used to analyse Atlantic salmon homologues (Koop et al. 2008). In this analysis percent similarity was used as a proxy for evolutionary time; lower percent similarity between two paralogues suggests more time has passed since their origin by duplication.



Species	Mean (%)	Median (%)	Peak (%)
E. lucius	84.2	83	82
S. salar	90.5	91	94
G. aculeatus	84.8	83	82



Comparison	Mean (%)	Median (%)	Peak (%)
E. lucius v S. salar	87.7	88	89
E. Lucius v G. aculeatus	83.0	83	82

Figure 3.2: The percent similarities identified through transcriptome blastn comparisons.

Blast results from non-redundant, repeat mask transcripts >300bp with minimum 100bp alignment. Results plotted by 1% intervals, grouped into bins. Results are graphed as A) within species (paralogues) and B) between species (orthologues). Eluc = *Esox lucius*, Ssal = *Salmo salar*, Gacu = *Gasterosteus aculeatus*.

Two peaks of percent similarity are apparent in the northern pike paralogue comparisons (Figure 3.2a): one above 97% and one centred at 82%. The analogous plot of stickleback paralogue similarity exhibits a similar profile (with a nearly identical peak at 82% identity). If a similar rate of mutation is assumed within the northern pike and stickleback lineages, the overlapping peaks occurring between 77% and 89% suggest a common origin. These peaks may represent the remnants of the 3R basal teleost genome duplication, which would have occurred in a distant common ancestor of both lineages. Atlantic salmon paralogue comparisons, in contrast, show much higher percent similarities and exhibit a notable peak at 94% identity, a region in which very few northern pike or stickleback paralogues appear. This upward shift in the peak of percent similarity is most easily explained by the fourth-round (4R), salmonid-specific WGD. While Atlantic salmon does not exhibit a distinct peak in the region identified in the other two fish species (77-89% similarity), the right tail of its percent distribution is considerably drawn out. The notably extended tail may be a result of a) paralogues produced following the 4R WGD with a much-increased rate of sequence divergence relative to the average, or b) remnants of the 3R duplication in which one copy of each paralogous gene pair produced following a salmon-specific 4R duplication has been lost. Together, these observations support the hypothesis that an Atlantic salmon ancestor experienced a comparatively recent whole-genome duplication event that did not occur in the pike or stickleback lineages, and supports the use of the northern pike as a WGD outgroup.

The peaks occurring above 97% similarity in all within-species comparisons imply very recent gene duplication events that are unique to each species. Although gene duplication is known to occur at very high rates, the vast majority of those duplicate pairs that are fixed in a lineage are thought to experience rapid loss of one or both copies, occurring within a few million years (Lynch 2000, Katju and Bergthorsson 2013). The high-similarity peaks observed in our within-species comparisons likely result from this dynamic of creation and loss; many paralogues are almost identical, indicating

recent duplication; however, fewer paralogues are observed as similarity decreases. These young paralogues may have been created by a recent series of segmental duplications (Lu et al. 2012), or may simply be an artefact of sequencing errors; we cannot with the current data distinguish between the two possibilities.

The most divergent peak observed in the northern pike within-species comparisons occurs from 77% to 89% similarity and contains 5,885 total transcripts. The annotations of these transcripts were examined to identify significantly over-represented functional categories (Additional File C7). Many of the terms identified fall into the development, signalling and regulatory classifications Brunet et al. (Aury et al. 2006) found to be enriched in *Tetraodon nigroviridis* gene duplicates likely to be from the 3R duplication. Of the Gene Ontology (GO) terms identified in the presumed 3R orthologues in northern pike, however, the most striking are phosphate metabolism biological processes and nucleoside-binding and nucleotide-binding molecular functions. Interestingly, nucleotide-binding molecular function category was significantly under-represented in 3R-duplicated genes examined in the *Tetraodon* pufferfish. Examining more recent duplicates in carp, Wang et al. (Wang et al. 2012) identified immune-related terms and pathways as the predominantly enriched biological functions in the recently duplicated common carp (*Cyprinus carpio*) genome, relative to the zebrafish; immune-related terms were not a significant component of the GO terms identified in this study. The diversity of enriched GO categories implied by these studies suggests that different taxonomic groups may retain gene duplicate pairs from different functional categories, thus contributing to overall taxonomic diversity. Alternatively, the observed differences could indicate that different functional categories of gene duplicates are retained over different lengths of time.

Utilizing the same approach as that used for within-species comparisons, northern pike transcripts were compared with their orthologues in both the Atlantic salmon EST dataset and the threespine stickleback transcriptome (Figure 3.2b). Unsurprisingly, Atlantic salmon and northern pike orthologues are on average more

similar (distribution maxima at 89%) than northern pike transcripts are to their respective orthologues in threespine stickleback (distribution maxima at 82%). These data are consistent with the expectation that the pike is more closely related to the salmonids than to species in the Neoteleostei, including stickleback. As northern pike and Atlantic salmon are estimated to have diverged from one another 100-130 million years ago (MYA) (Near et al. 2012, Betancur-R et al. 2013, Macqueen and Johnston 2014), the value of 89% provides a very broad reference point with which to estimate the timing of the orthologue divergence in the transcriptomes of the Salmoniformes and Esociformes.

Restriction-site associated DNA (RAD) tag based linkage mapping inefficient in pike

Our attempts to produce a RAD-tag based genetic map from a single full-sib family of *E. lucius* were unsuccessful. From 33,588 *Sbf*I-associated sequence loci across all family members, 351 informative polymorphic SNP markers were identified, or about 175 informative SNPs per sex. Further, the combination of very few useful markers and the bi-allelic nature of the SNPs did not allow for an efficient merge of the two sex-specific maps. Attempts to produce a useful linkage map from this data proved fruitless. Our initial decision to pursue this approach was based on its comparative success in other fish species. Other studies using the same protocol and enzyme produced 8,406 mapped SNPs in the spotted gar (Amores et al. 2011), 6,458 mapped SNPs in Atlantic salmon (Gonen et al. 2014) and 5,703 mapped SNPs in Atlantic halibut (Palaiokostas et al. 2013a). The mediocre performance of this method in our study is almost certainly due to the very low degree of polymorphism observed in the northern pike data; the ALLPATHS-LG genome assembler reported an estimated polymorphism rate of 1 SNP for every 10,830 bp and an independent analysis of our RAD-tag dataset indicated an average polymorphism level of approximately 1 SNP for every 6,000 bp. The polymorphism rate estimates from our data are consistent with the very low levels of

genetic polymorphism and heterozygosity noted in previous genetic studies of North American pike populations (Miller and Senanan 2003). It is possible that future attempts utilizing a restriction enzyme which cuts more frequently could restore the usefulness of this RAD-tag approach. Such changes, however, would significantly increase the number of sequencing lanes required, thereby reducing some of the cost advantages of the RAD-tag technique.

A first-generation, Microsatellite-based linkage map

To increase the detection and integration levels of polymorphism, we utilized a half-sibling family (48 progeny from one father and 44 from a second) in combination with microsatellite markers. Utilizing 32,833 previously published northern pike EST sequences (Koop et al. 2008, Leong et al. 2010), we identified long near-perfect di- and tri-nucleotide repeats from which we generated 776 microsatellite primer pairs, of which 187 (24%) were found to be polymorphic in at least one of the parents. Of these polymorphic primer pairs, 180 were successfully mapped back to a single *E. lucius* genomic scaffold. To further extend our database of polymorphic primer pairs we similarly screened the largest 900 genomic scaffolds for long near-perfect di- and tri-nucleotide microsatellites. This approach allowed us to screen an additional 835 scaffold-designed primer pairs against the parents of our half-sib family. We identified 378 primer pairs that were polymorphic in at least one of the parents with between 1 and 6 markers identified per scaffold. From a total of 1,611 screened primer pairs we identified 565 polymorphic markers that cover 429 genomic scaffolds; 7 such markers could not be mapped to a scaffold.

The 565 polymorphic primer pairs we have identified represent a significant increase in the available markers for *E. lucius*; prior to this work, only 70 markers were available (Miller and Kapuscinski 1996, Launey et al. 2003, Wang et al. 2011, Gotoh et al. 2013, Ouellet-Cauchon et al. 2014). Given the low level of polymorphism observed in northern pike, particularly in North American populations, the additional marker

availability should be useful to future studies that seek to delineate population structure or trace unauthorized introductions of this fish species.

Analysis of the genotypes found within our half-sib family allowed 526 of the 565 polymorphic primer pairs (and 406 scaffolds) to be mapped to 25 linkage groups spanning 1289.3 cM, the expected number of linkage groups based on previous *E. lucius* karyotyping (Beamish et al. 1971). The final merged linkage map is presented in Figures 3.3-3.5. An additional 12 markers were polymorphic but remained unlinked at the LOD cut-off of 3.0. Of the 28 markers with more than 15% missing data, all but one was dropped from the linkage analysis: u0028c was maintained in the data set to join two groups of markers into the single linkage group LG04. This join was further verified at LOD 3.0 using u0046a, u0028c and u0330a in an unrelated family of northern pike (Whiteshell hatchery, MB, Canada). While most of the mapping procedure was straightforward, some difficulty was observed in the merging of the map due to a low level of polymorphism across multiple individuals. Of the 526 mapped markers, 301 were only informative in a single parent, leaving relatively few common markers with which to merge the maps. Therefore, true marker order can only be determined from the individual maps in Additional File C1; the merged map is presented as the best interpretation of the available data.

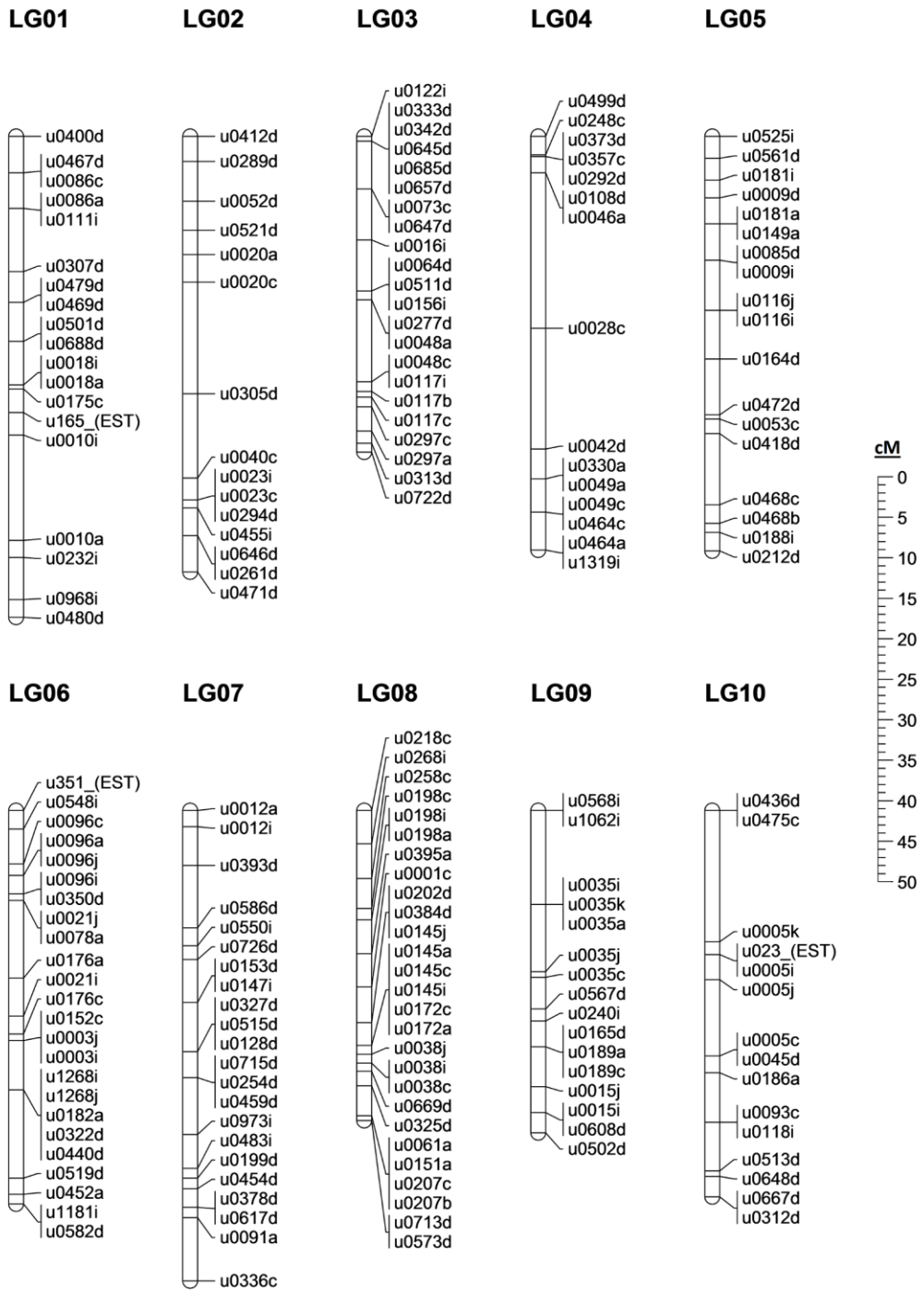


Figure 3.3: The northern pike merged linkage map: LG01 to LG10.

Genomic scaffold identified by first 4-digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Additional File C1. Marker information in Additional File C9.

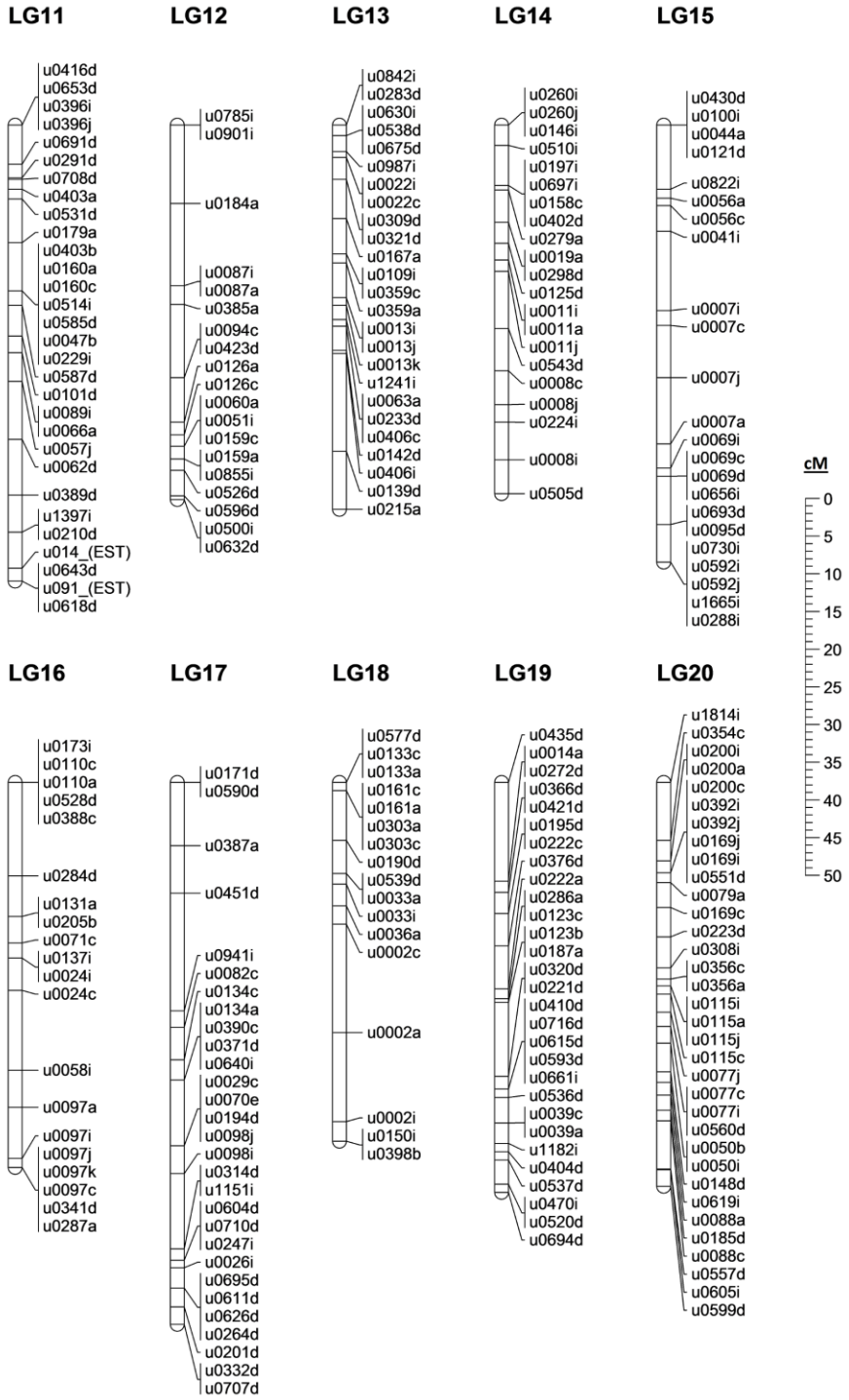


Figure 3.4: The northern pike merged linkage map: LG11 to LG20.

Genomic scaffold identified by first 4-digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Additional File C1. Marker information in Additional File C9.

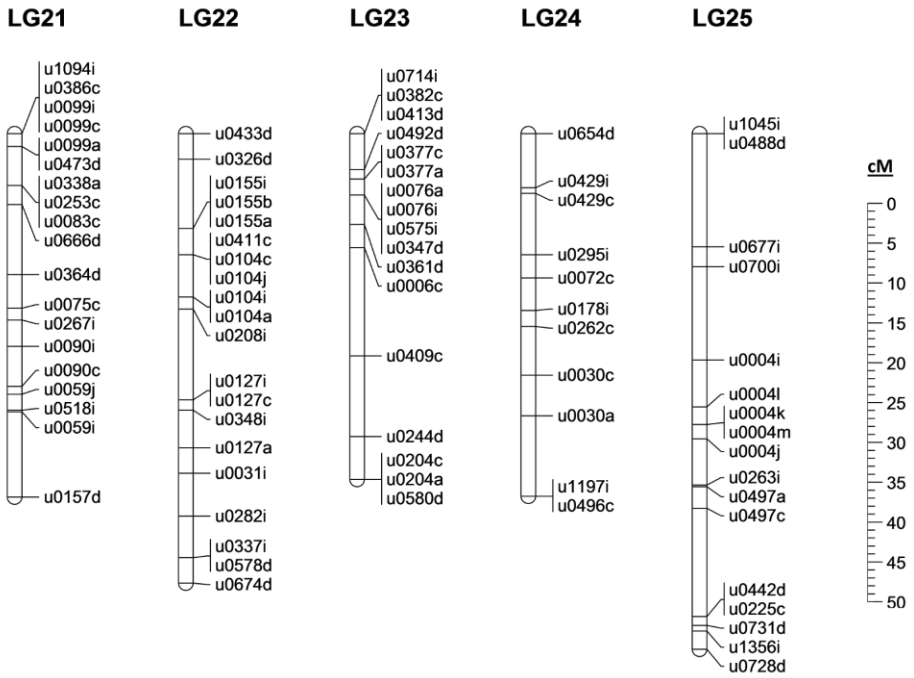


Figure 3.5: The northern pike merged linkage map: LG21 to LG25.

Genomic scaffold identified by first 4-digit number, with multiple scaffold markers distinguished by final letter. Markers not mapped to a scaffold identified by "(EST)"; sex-specific linkage maps in Additional File C1. Marker information in Additional File C9.

Sex-specific maps spanned 1245.4 cM in the female and 1166.0 cM in the male, an average recombination ratio of 1.07:1 between females and males. Regional recombination differed across and within linkage groups between the sexes, with one region standing out. In the male maps, the top of LG14 shows strongly repressed recombination in the males relative to the females, with ~11X greater rate of recombination between scaffold 260 and scaffold 11 in the female map relative to the male. While the average recombination rates diverge slightly from even between the two sexes in pike, the recombination ratio is not nearly as skewed as in other teleost species; in Atlantic salmon, recently published ratios ranging from 1.3-1.7:1 (female:male) (Lien et al. 2011, Gonen et al. 2014), in rainbow trout (*Oncorhynchus mykiss*) 1.68:1 (Rexroad et al. 2008), and in zebrafish (*Danio rerio*) 2.74:1 (Singer et al. 2002).

In one of the fathers we noted the unusual observation that of the 29 markers mapped to LG19, none were found to be polymorphic (individual male-1 in Additional File C1). In early karyotyping of northern pike, various chromosome counts were observed (Beamish et al. 1971), particularly in the range of 48-50 chromosomes; in concert with our data, these observations suggest that the father was possibly missing a single chromosome, or transmitted only a single chromosome to his progeny. It is also possible, depending on the degree of inbreeding within our sampled population, that this father inherited a separate but nearly identical chromosome from each of his parents; no observable polymorphism would be displayed in this situation. We are unable to determine the biological source of this unusual observation, as the available tissues are insufficient for karyotyping.

The low level of polymorphism implied by our data produces an interesting trade-off: genome assembly proved relatively straightforward due to a lack of sequence ambiguity introduced by polymorphic sites, while this same lack of polymorphic diversity ultimately limited the number of scaffolds that could be placed and oriented on the linkage map. Originally, we identified and targeted at least two microsatellite markers on each of the longest genome scaffolds; however, very few scaffolds ultimately possessed more than one such marker that was usefully polymorphic. Only 41 marker pairs derived from the same scaffold were mapped in such a way as to imply definitive scaffold orientation in at least one of the sex-specific linkage maps. Generally, such marker pairs were from the largest scaffolds, and almost all were found on scaffolds greater than 1 Mb in length. The 406 scaffolds mapped by at least one microsatellite marker to this first-generation linkage map contribute to a total of 46% of scaffold assembly bases being associated with a linkage group. Additional northern pike families, further polymorphic marker identification, directed sequencing of scaffold gaps, and/or additional whole-genome sequencing will be required to improve this value in future work.

Synteny analyses suggest Neoteleostei-like genome in northern pike

Given that most teleost species fall into either the Ostariophysi or the Neoteleostei subcohorts, it is understandable that all nine teleost genome sequences available in the Ensembl genome browser (December 2013) fall into one of these two clades. While the Ostariophysi are thought to have diverged from the Euteleosteomorpha approximately 250 MYA, the next branch point possessing a representative genome assembly occurs within the Neoteleostei, approximately 160 MYA when the Paracanthomorphacea (represented by the Atlantic cod) diverged from the Euacanthomorphacea (Betancur-R et al. 2013). Usefully, the Protacanthopterygii are thought to have branched off from other Euteleosteomorpha approximately 200 MYA, and as such, they represent a phylogenetically intermediate branch point between the fish species with currently available genomic data. When considering the reconstruction of ancestral genomes for evolutionary studies, the addition of branch points and the decreasing of branch lengths can only improve the quality of the results. If a phylogenetically useful Protacanthoptyerygiian genome is desired, the selection of an order that did not experience the salmonid-specific genome duplication is desirable to limit complications; such an order could be reasonably expected to provide the most straightforward and ancestral-like genome that can be obtained from this group. The Esociformes, represented in the present study by northern pike, are such a suitably placed order; the northern pike genome will be an excellent intermediate resource for use in phylogenetic studies of teleosts.

Using the genetic map, we set out to determine if the gene distribution and genome structure of northern pike imply synteny with the hypothesized genome of the ancestral teleost, as is tentatively implied by karyotype morphology. To this end, the syntenic relationships were examined between four species representing three major subcohorts of the infraclass Teleostei; northern pike (*Esox lucius*) was chosen to represent the Protacanthopterygii, zebrafish (*Danio rerio*) was chosen to represent the Ostariophysi, and both threespine stickleback (*Gasterosteus aculeatus*) and medaka

(*Oryzias latipes*) were selected to represent the Neoteleostei. First, transcripts from northern pike were mapped to those northern pike genomic scaffolds that had previously been assigned to a group on the linkage map. Putative orthologues for these mapped transcripts were then identified in the transcriptomes of the three other representative teleosts using a reciprocal BLAST approach. A linkage group was subsequently inferred for each mapped transcript based on that of its parent scaffold. Finally, this *E. lucius* transcript linkage group was compared to the chromosomal location of the transcript's orthologue in stickleback (Figure 3.6a), medaka (Figure 3.6b) and zebrafish (Figure 3.6c); see Additional File C8 for specific numbers.

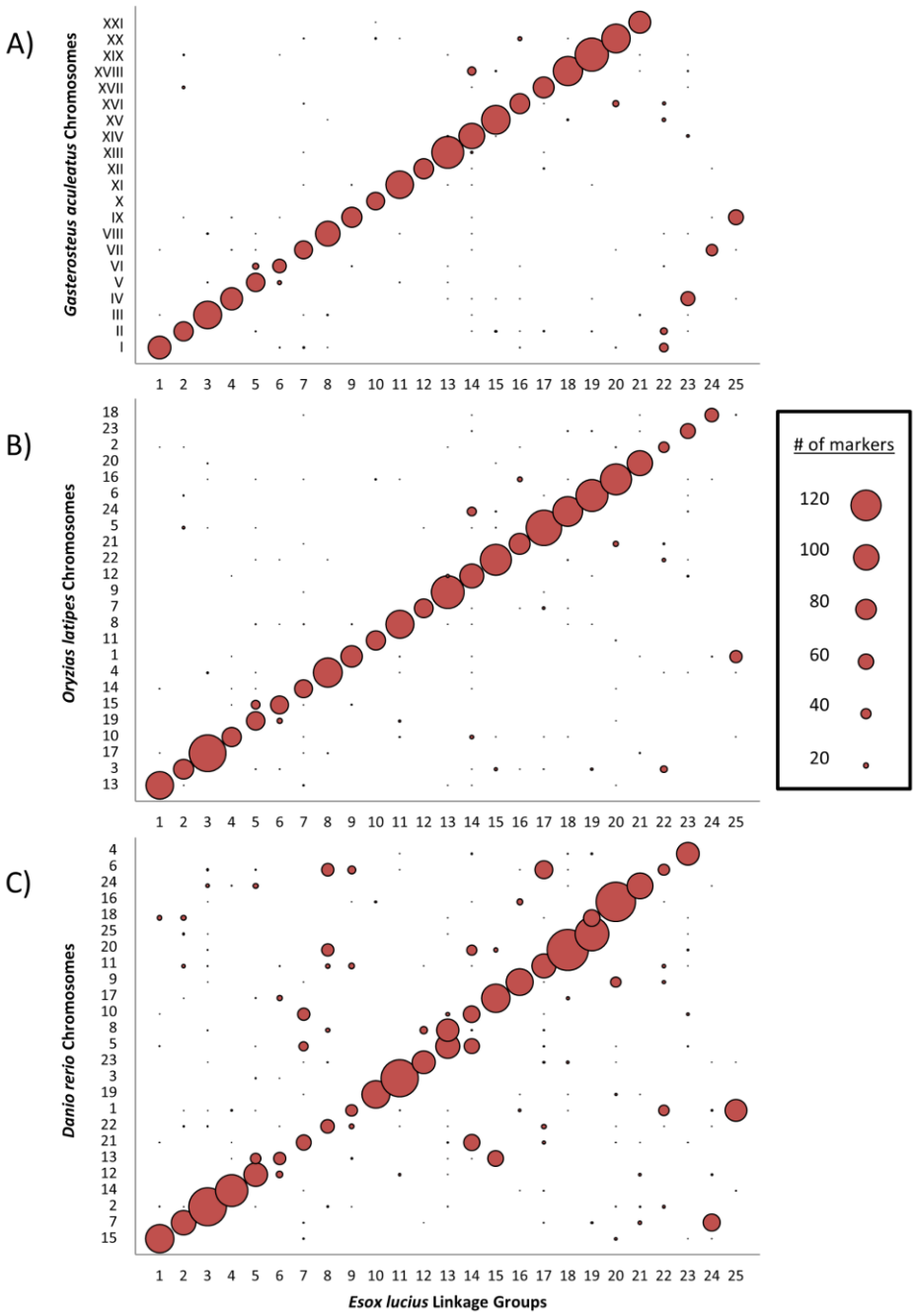


Figure 3.6: The synteny between the northern pike and model teleosts.

Reciprocal best hit results (≥50% of total length, ≤ 1e-5) displaying synteny between the 25 pike linkage groups and three model teleost genomes, threespine stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), and zebrafish (*Danio rerio*). Sphere diameter reflects the number of matched transcripts. For numerical results, see Additional File C8.

Considering the time since their last common ancestor there exists a clear association between linkage groups in *E. lucius* and chromosomes in *G. aculeatus* and *O. latipes*. In general, each northern pike linkage group predominantly associated with a single stickleback chromosome. In four cases there were two *E. lucius* linkage groups that associated with only a single stickleback chromosome. Of these four cases, only one was analogously observed in *Oryzias latipes* (Figure 3.6b), a species that possesses a karyotype equivalent to the presumed ancestral karyotype in the Neoteleostei (Mank and Avise 2006). Specifically, only Chr1 of *O. latipes* (orthologous to *G. aculeatus* ChrIX) shows a strong association with two separate *E. lucius* linkage groups. This observation suggests that the other three instances in which a single stickleback chromosome corresponds to two northern pike linkage groups (i.e. *G. aculeatus* Chrs. I, IV and VII) are best described as stickleback lineage-specific fusions of three pairs of ancestral-like chromosomes. This interpretation is further supported by recent work using sablefish (*Anoplopoma fimbria*), a species which is much more closely related to stickleback than is northern pike or medaka (Rondeau et al. 2013).

As previously discussed, transcripts from northern pike linkage groups LG09 and LG25 predominantly possess orthologues on stickleback ChrIX and medaka Chr1; an analogous situation is observed with zebrafish Chr1. These observations suggest that the common ancestor of the Neoteleostei and the Protacanthopterygii most likely possessed a haploid karyotype of 24 chromosomes; this is in contrast to the haploid karyotype of 25 chromosomes observed in members of the *Esox* genus (Beamish et al. 1971) and in the majority of the members of the Ostariophysi (Mank and Avise 2006). This analysis further implies that the two *E. lucius* chromosomes represented by LG09 and LG25 were formed because of the fission of a single ancestral chromosome and that this fission occurred after the divergence of the Protacanthopterygii lineage. Syntenic analyses of orthologous sequence shared by stickleback and Atlantic salmon (Lien et al. 2011) suggest that the ancestral chromosome that eventually evolved into stickleback chromosome IX had already split in two prior to the salmonid-specific genome

duplication. Therefore, this chromosomal fission likely occurred prior to the Esociformes/Salmoniformes divergence and it can be reasonably concluded that the common ancestor of the Esociformes and Salmoniformes possessed a 2N=50 karyotype which was secondarily derived from the more ancestral karyotype of 2N=48.

While the distribution of orthologous genes to chromosomes was particularly well conserved between the Protacanthopterygii and the Neoteleostei, the gene order along chromosomes was not as consistently similar. The positions of mapped genes in northern pike and the three model teleost genomes were plotted for the largest 50 scaffolds (Additional File C2). In some instances, such as scaffolds 4 and 6, syntenic blocks more than three 3 million bp can be found where gene order appears to be conserved between medaka, stickleback and northern pike. In other cases, such as those observed in scaffolds 13 and 15, there are only small blocks of 500,000 bp or less where conservation of gene order can be observed. There are qualitatively many fewer occurrences of conserved gene order between zebrafish and pike. However there do exist large tracts of conserved order in certain scaffolds; one such scaffold is scaffold 1, where sections of 1-2 million bp maintain consistent ordering. In fact, gene order appears to be more consistently similar between zebrafish and pike than between pike and medaka in scaffold 1.

While the relatively small number of reciprocal best blast hits per scaffold makes quantifying the degree of gene order conservation difficult, the number of consistently ordered transcripts from a single syntenic chromosome lends confidence to the quality of the assembled scaffolds. In three of the largest 50 scaffolds (scaffolds 8, 9 and 31), however, northern pike transcripts are consistently mapped across two or three chromosomes in all three model teleosts; this raises the possibility of chimeric scaffold assembly. An examination of the scaffolds mapped to either side of scaffold 9 in LG05 implies that the discrepancy in this scaffold represents a genuine chromosomal translocation, in which the breakpoint falls in the middle of the scaffold. Scaffolds to one side of LG05 exhibit blast hits predominately to *G. aculeatus* ChrV, while scaffolds on the

other show blast hits to *G. aculeatus* ChrVI. Similarly, scaffold 8 (mapped to LG14) also appears to contain a genuine chromosomal translocation rather than a chimeric assembly artefact. This is inferred because of a) most other scaffolds on LG14 possess the best blast hits to *G. aculeatus* ChrXIV and b) scaffold 224, mapped adjacent to the ChrXVIII 'end' of scaffold 8 in LG14, exhibits blast results dominated by hits to *G. aculeatus* ChrXVIII. Given the pronounced split between these regions, it seems likely these translocations are quite recent and that the genes involved have not yet had time to be re-distributed to other regions of the chromosome. The third scaffold exhibiting potential assembly chimerism, scaffold 31 (located on LG22), is not flanked by scaffolds that support a similarly distinct breakpoint. However, scaffolds mapped to LG22 possess genes with significant numbers of orthologues on at least four different chromosomes in *O. latipes* and *G. aculeatus* (Figure 3.6a and b; Additional File C8); genes from different chromosomal sources are also mixed and distributed over the entire length of this linkage group. Together, these observations suggest that LG22 possesses an old collection of pieces from multiple ancestral chromosomes that have had considerable time to redistribute themselves throughout the linkage group. This characterization of LG22 supports the idea that scaffold 8 reflects the biological reality and is not the result of inappropriately linked assembly contigs. While we are unable to rule out the possibility of chimeric joins having occurred in our genomic assembly, the three large scaffolds deemed most likely to have been chimeric show evidence suggesting that they are indeed correct.

Although gene order is highly variable, chromosomal synteny remains high between the northern pike and the Neoteleostei representatives. While small sections appear to have undergone chromosomal translocations, and LG22 may be highly reorganized, there is generally a 1:1 relationship between linkage groups identified in northern pike and the chromosomes in the Neoteleostei. This implies the genome of *E. lucius* is indeed ancestral-like and will provide a good overall representative of the pre-duplication Protacanthopterygii genome structure in wider studies of genome evolution.

The usefulness of salmonid duplication analysis

While inferring ancestral states is made easier with an ancestral-like genome, the examination of more recent events, such as the WGD event and subsequent re-diploidization in the salmonid lineage, can be made much easier using this genetic dataset. Although numerous rearrangements have undoubtedly had a substantial effect on genome organization, many of the greatest genomic changes must be seen as the consequence of the formation and evolution of paralogous genes resulting from a WGD (such gene pairs are termed homeologues or ohnologues (Wolfe 2000)). As the genome begins to revert to a stable diploid state, ohnologues are presumed to evolve like individual gene duplicates through subfunctionalization, neofunctionalization or loss of one of the gene copies. Determining the mode and rate of gene and genome evolution requires an understanding of where the two copies originated. Gaining this understanding would be most easily accomplished using the ancestral, pre-duplicated sequence of the gene or genome segment; since this sequence is unavailable, it is most readily inferred using an appropriate pre-duplication sister group. Even given the significant amount of time that has passed since their divergence from each other over 100 MYA (Macqueen and Johnston 2014) the Esociformes remain the most closely related order to the Salmoniformes and therefore represent the best system of species from which a pre-duplication genomic state can be inferred.

Nucleotide BLASTn alignments between the flanking sequence of 5,918 mapped genomic markers from the Atlantic salmon linkage map (Lien et al. 2011), and the northern pike genome scaffolds produced 3,241 high-quality alignments. Of these alignments, 1,803 involved scaffolds assigned a linkage group on the northern pike genetic map. An analysis of these alignments clearly indicates that each *E. lucius* linkage group possesses significantly conserved synteny with two Atlantic salmon chromosomes (Table 3.2). This observation is consistent with a historical duplication of each chromosome in the salmonid lineage that did not occur in an esocid ancestor. These

results corroborate previous analyses which compared Atlantic salmon to stickleback (Lien et al. 2011). Indeed, the increased relatedness and shared Protacanthopterygiian-specific rearrangements between Atlantic salmon and northern pike demonstrate an even greater example of shared synteny between these two species that are separated by a genome duplication.

Table 3.2: Sequences from mapped Atlantic salmon SNPs against the *Esox lucius* genome (≤1e-10)

The Atlantic salmon chromosome linked to each mapped SNP was plotted against the linkage group of the top *E. lucius* scaffold blast match. Results < 3 for a given LG/chromosome comparison are removed for clarity.

		***Esox lucius* Linkage Groups**																								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
***Salmo salar* Chromosomes**	**Ssa01**					8	**28**								**44**	**41**										
	Ssa02									**45**							13				**40**					
	Ssa03			**37**								**43**														
	Ssa04							**28**																		**24**
	Ssa05				**28**												10				**47**					
	Ssa06											**52**							**43**							
	Ssa07																		3					**17**	**10**	
	Ssa08																									**15**
	Ssa09	**32**			**34**											**29**										
	Ssa10								**68**	3										**36**						
	Ssa11		**35**					10							**31**											
	Ssa12		3							**18**								**55**								
	Ssa13							**20**					**28**		11											
	Ssa14			**18**							**25**															
	Ssa15												**13**	5	8				**33**							
	Ssa16	3							22											**20**			**23**			
	Ssa17																						**22**	**23**		
	Ssa18					4	**29**																		**16**	
	Ssa19					**17**	3															**21**				
	Ssa20	**18**												**36**												
	Ssa21																**25**				6					
	Ssa22		4															**56**								
	Ssa23								**33**											12						
	Ssa24													**39**												
	Ssa25																**25**									
	Ssa26		**46**									3								3						
	Ssa27										**41**															
	Ssa28					**24**	8																			
	Ssa29																					**24**				
Total blast hits		**57**	**89**	**58**	**65**	**59**	**71**	**63**	**131**	**67**	**72**	**108**	**43**	**87**	**103**	**73**	**77**	**118**	**84**	**74**	**97**	**46**	**49**	**45**	**27**	**40**

This work in northern pike provides a base from which to examine genome evolution following whole-genome duplication. The reference transcriptome allows for the functional analysis of ohnologue evolution while the genome assemblies facilitate the study of intron and promoter evolution as well as post-WGD gene and genome rearrangements. The role of transposable elements in both esocid and salmonid genome evolution and speciation can be further explored by inferring pre-duplication genomic repeat content, by dating periods of transposable element activity and by characterizing the abundance and character of transposable elements in this pre-duplication genome.

Conclusions

This work represents the largest currently available collection of genomic data for a member of the Esociformes and the first genome assembly for a non-salmonid member of the Protacanthopterygii, an economically important superorder which bridges the evolutionary gap between the Ostariophysi and Neoteleostei. We present a genome assembly of 824 Mbp across 94,267 contigs, and 878 Mbp across 5,688 scaffolds, representing an estimated 96 % of the genome; 50% of the assembly is represented in the largest 318 scaffolds. Approximately 20% of the genome sequence is made up of repeat elements. Two transcriptome assemblies from multi-tissue RNA-seq are presented, and FPKM values provide a baseline dataset of expression levels across 13 individual tissues. The microsatellite-based linkage map places 526 markers across 25 linkage groups, the expected number of linkage groups based on previous karyotypes. These markers further facilitate the mapping of 46% of the scaffold assembly sequence onto linkage groups. Comparisons between linkage-mapped scaffolds and three teleost genomes show that northern pike has a chromosomal distribution of genes very similar to that of the Neoteleostei, suggesting relatively low chromosomal translocation activity and an ancestral-like gene distribution. Gene order along individual chromosomes, however, exhibits significant reorganization. This dataset represents a valuable resource

for future ecological and evolutionary work in northern pike and will assist in understanding the effects of whole-genome duplication on genome evolution, especially once the upcoming salmonid genomes become available.

Methods and Materials

Animal Care

In accordance with the Canadian Council on Animal Care Guidelines, the Animal Care Committee at the University of Victoria did not require ethical review as only archival tissue or tissue from fish that were harvested by various government agencies for their own purposes, was used during the experiments.

Whole-Genome Sequencing

For whole-genome sequencing, genomic DNA was extracted from the spleen of a single male pike individual from a population introduced to Charlie Lake, British Columbia, CANADA (Nordin and Pommen 1985), using pooled DNA from both DNeasy (QIAGEN) and phenol (Sambrook et al. 2000) extractions. DNA libraries of 180bp were prepared and sequenced (paired-end) by the Michael Smith Genome Sciences Centre; libraries of 2 kb and 5 kb fragments were constructed and sequenced by BGI (http://www.genomics.cn/en/index); all sequences were produced using Illumina HiSeq2000 instruments (San Diego, CA).

Following ALLPATHS-LG genome assembly, contigs were checked for long runs of ambiguous bases and trimmed/split if necessary. We removed contigs that were primarily vector sequence (BLASTn 1e-50). Furthermore, all contigs were trimmed internally and at terminal ends for any vector sequence according to NCBI's VecScreen search recommendations (www.ncbi.nlm.nih.gov/tools/vecscreen/). Duplicate contigs were removed and mitochondrial contigs were removed (BLASTn 1e-5). This resulting contig set was compared to GenBank's non-redundant nucleotide database for contamination (BLASTn 1e-25). Potential non-vertebrate sequence matches were

flagged and manually inspected to determine classification. These contig sequences including scaffold information have been uploaded to NCBI under BioProject ID PRJNA221548, accession AZJR00000000.

Transposable element library creation and genomic repeat annotation

The process used to create the northern pike transposable element (TE) library was based on that used by Wegrzyn et al. (Wegrzyn et al. 2013) to identify TE sequences in the Loblolly Pine genome. First, a custom repeat 'seed' library (CSL) containing 33,304 TE sequences was assembled from three sources: 1) a quality-checked repeat consensus library produced by the *de novo* repeat-finding program REPET (version 1.3.9.1; default settings; (Flutre et al. 2011)); 2) the RepBase database of TE sequences (REPET-formatted v18.08; (Jurka et al. 2005)); and 3) TE sequences previously identified in Esociformes or Salmoniformes and deposited in GenBank.

Following CSL creation, the CENSOR program (Jurka et al. 1996) was used to identify all non-redundant instances of CSL sequences in the northern pike genome contigs. Subsequent filtering removed all CENSOR hits that were smaller than 80% of the length of the respective CSL query sequence. Sequences from long CENSOR hits were clustered using the uclust.global program of USEARCH (Edgar 2010). USEARCH parameters were such that a sequence was only added to an existing cluster if it possessed at least 80% nucleotide similarity over 80% of its length to the cluster's representative 'centroid'. Following clustering, a representative sequence was obtained for each cluster. If there were fewer than three sequences in the cluster the longest sequence was chosen as a representative. Otherwise, a consensus sequence was generated from a multiple sequence alignment (MSA). MSAs were built using the T-COFFEE program (Notredame et al. 2000), which used alignment data from two sources: global alignment information from MUSCLE (Edgar 2004a, 2004b) and local alignment information from USEARCH.

The representative sequences from all clusters were combined, and sequences with strong BLASTx alignments (E-value ≤ 1e-10) to the UniProtKB/SwissProt protein

database (July 2013 release) were manually examined and, if appropriate, removed as non-TE host genes. The remaining library sequences were classified in accordance with the Wicker et al. classification system (Wicker et al. 2007) based on the occurrence of structural motifs and/or similarity to previously identified TEs. The classification process relied on both manual and automated processing of output from NCBI's BLAST+ software and the PASTEClassifier.py script included with REPET. The final northern pike TE library contained 26,628 sequences of which 14,820 were classified to at least the subclass level under Wicker's taxonomic system. Many of the in-house scripts used to create and validate the TE library utilized modules included in the BioPython library (Cock et al. 2009, Talevich et al. 2012).

The total amount of low-complexity DNA and TE-derived sequence within the genome was estimated using RepeatMasker (v4.0.3 (Smit et al. 2010)) ; RepeatMasker was configured to use RMBlast (version 2.2.23+) as its internal search engine. The abundance estimations for individual TE taxa were aggregated from RepeatMasker's annotation output file.

RNA-seq and assembly

All tissues for RNA-seq were extracted from a single, juvenile, male pike caught November 1, 2011, in the Portage Diversion/Assiniboine River Floodway near Portage la Prairie, MB, Canada (N50.00747 W98.38181) as the water was beginning to freeze. The animal was euthanized and then frozen on dry ice until sampling, and subsequently at -80°C. RNA from all 13 tissues – brain, eye, gill, hindgut, head kidney, heart, kidney, liver, muscle, nose, stomach, spleen and testis - were extracted by mixer-mill (Retsch) homogenization in TRIzol (Invitrogen), followed by column clean-up using the RNeasy kit (QIAGEN). Total RNA was submitted directly to BGI for Illumina sequencing; indexed samples were pooled (4 per RNA-seq lane, pooled along with other projects) and sequenced on an Illumina HiSeq2000.

Paired-end raw reads were quality-trimmed using Trimmomatic (Lohse et al. 2012). Trimmed reads from all tissues were combined, and a *de novo* transcriptome

assembly was created using the Trinity assembler (Grabherr et al. 2011). Trinity used a minimum kmer count of 3 to reduce noise introduced by sequencing errors.

To generate a non-redundant reference set of transcripts, the resulting set of Trinity transcripts were filtered. These putative raw transcripts were reduced by retaining those that were characterized as full-length (Leong et al. 2010), those that had a significant BLASTx (Altschul et al. 1997) match to the UniProtKB/SwissProt or Gene Ontology protein databases (≤ 1e-5) without being transposable element annotations, and those that did not show any sequence homology to a known protein but had a predicted open reading frame ≥ 300 bp. This reduced set was mapped to our genome assembly using BLAT (Kent 2002). Transcripts mapping uniquely to a single locus were retained. In cases where multiple transcripts (from alternatively-spliced mRNAs) were mapped to the same genomic location, only the longest transcript was retained. We masked repeats using RepeatMasker (Smit et al. 2010) and our *Esox lucius* repeat library, removing open reading frames that were no longer ≥ 300 bp. To remove possible alleles, recent duplicates, and sequencing errors in our de novo assembly, we took a single representative of transcripts that were ≥ 98% similar over a minimum length of 300 bp, as determined by BLASTn. This curated set represents our RNA-seq reference transcriptome.

We produced an *ab initio* transcript set from our genome contig assembly using MAKER2 (Cantarel et al. 2008, Holt and Yandell 2011). An Augustus (Stanke and Waack 2003) gene model specific for *Esox lucius* was generated using contigs from the RNA-seq assembly and the previously published EST assembly as evidence. MAKER2 was run for two rounds to produce an *Esox lucius* SNAP model. A final round of MAKER2 was run using the Augustus gene model, SNAP model, and the mRNA evidence.

Homology analysis of protein-coding mRNA data utilized the non-redundant protein-coding northern pike transcriptome (repeat-masked), Atlantic salmon EST data and *Gasterosteus aculeatus* transcripts obtained from the UCSC genome browser. Intra-transcriptome comparisons for northern pike, Atlantic salmon and threespine

stickleback, as well as between-species comparisons (northern pike vs Atlantic salmon and northern pike vs threespine stickleback), were performed using a reciprocal best BLASTn approach. Only BLASTn alignments longer than 100 bp and possessing an E-value ≤ 1e-5 were considered in establishing homology. DAVID analyses (Huang et al. 2008) were performed using paralogues from *Esox lucius* that exhibited 77-89% similarity in the aforementioned BLASTn alignments. Results were categorized by molecular function and biological process, and sorted by EASE-Score (p-value).

RAD-tag based linkage mapping

A RAD-tag-based approach for linkage map construction was attempted following the protocol of Amores et al. (Amores et al. 2011) using the *Sbf*I restriction enzyme. Tissues from a family consisting of a single father and mother (fin clips) and 94 progeny (yolk-sac larvae) from the Hackettstown Fish Hatchery, New Jersey, USA were extracted and used to construct three barcoded Illumina libraries, each containing data from 32 individuals. These libraries were paired-end sequenced using an Illumina HiSeq-2000 sequencer (Michael Smith Genome Sciences Centre, Vancouver, BC, CANADA). Subsequent analysis to identify and score SNPs was performed with Stacks (Catchen et al. 2011).

Microsatellite Primer Design

A second family was obtained from the Hackettstown Fish Hatchery, New Jersey. Tissues stored in 95% ethanol from 103 progeny (either yolk-sac larvae or fin clips from 4-month old juveniles) of a half-sibling family produced from 3 fathers and a single mother were extracted by placing 2 mm^3 of tissue in 5% Chelex 100 (BioRad), 0.2% SDS and 0.27 mg/ml proteinase K (Invitrogen) for 2 hours at 55°C followed by 95°C for 10 minutes (Messmer et al. 2011). Extractions were diluted 1/100 in water for use in PCR. Offspring were screened using microsatellite primers to determine parentage (see

below for primer design); progeny was distributed 106:45:0 between the three fathers. The family was therefore used as a two-father/single mother half-sibling family in further analysis, split 48:44 progeny per father to fit a 96-well format.

The first 776 primer pairs were designed to target di- and tri-nucleotide repeats identified by Tandem Repeats Finder (Benson 1999) from *Esox* EST sequences (Leong et al. 2010), with an additional 835 primer pairs designed to target microsatellite repeats identified from genome scaffolds (1-3 primer pairs evaluated per scaffold). The unlabelled EST-designed primer pairs, u001-u296 were evaluated on a 5% polyacrylamide gel prior to re-ordering HEX or 6-FAM directly-labelled forward primers. The forward primer from each of the remaining primer pairs (u297-u776) and all scaffold primers were ordered with one of three tails, Tail-A, Tail-B or Tail-C for fluorescently labelling products with a labelled universal third primer, adapted from Blackett et al. protocol (Blacket et al. 2012). The universal primers were labelled: Tail-A with 6-FAM, Tail-B with HEX, Tail-C with NED, 6-FAM and HEX (Blacket et al. 2012).

For directly labelled primers, each PCR reaction contained: ~2.5ng template DNA, 200 µM each dNTP (Promega), 0.5 µM each forward (HEX or 6-FAM labelled) and reverse primer (IDT), 0.25U Hot Start *Taq* DNA polymerase (Promega), 1X GoTaq Flexi Colorless PCR buffer (Promega), 2.0 mM $MgCl_2$, to a final volume of 10 µl with DNase/RNase free H_2O (Gibco). Reactions were cycled on a TC-412 thermocycler (Techne) following the profile: 3 min at 95°C followed by 35 cycles of 30 s at 95°C, 30 s at 52°C and 30 s at 72°C, and a final extension of 72°C for 10 min with a hold at 4°C. For "Tailed" reactions: ~2.5ng template DNA, 200 µM each dNTP (Promega), 0.15 µM forward primer (IDT), 0.5 µM reverse primer (IDT), 0.2 µM labelled universal tail primer (IDT), 0.25U Hot Start *Taq* DNA polymerase (Promega), 1X GoTaq Flexi Colorless PCR buffer (Promega), 2.0 mM $MgCl_2$, and final volume of 10 µl with DNase/RNase free H_2O (Gibco). Tailed reactions were cycled on the same profile as above, except the 52°C annealing step was replaced with 56°C for 1 minute 30 seconds.

Following PCR, genotyping reactions consisted of 1.0µl each PCR product, 9.9 µl Hi-Di Formamide (Life Technologies) and 0.1 µl GeneScan -500 ROX Size Standard (Life Technologies); two or three PCR reactions with different dyes were pooled into a single genotyping reaction. Following denaturation for 3-5 minutes at 95°C, samples were cooled on ice for 5-10 minutes before analysis on a 3730 DNA analyzer (Life Technologies). The generated electropherograms were analysed using GeneMapper V4.0 (Life Technologies), with automated scoring reviewed manually for errors in all samples. Samples for which clear peaks were not identified were removed from the dataset in subsequent analyses. All primer pairs were evaluated on the mother and two fathers of the family. Markers identified as polymorphic in the parents were used to amplify and score the distribution of alleles in the progeny. Polymorphic marker information is summarized in Additional File C9.

Linkage Mapping

Linkage mapping was performed using the LINKMFEX package, v 2.3 (R. Danzmann, University of Guelph, http://www.uoguelph.ca/~rdanzman/software.htm) following the standard protocol; markers with more than 15% of genotypes missing were omitted from the analysis (apart from u0028c; see below). Linkage groups were identified at a LOD threshold of 4.0; linkage groups in the individual maps were further joined at a LOD of 3 if in another individual the linkage was recognized using the stricter cut-off. Further estimation of distances below LOD = 3 was used to create merged, sex-specific maps; these are represented by dotted lines in Additional File C1. Merged sex-specific and final merged maps were produced using the MERGE programs in the LINKMFEX package. Maps were visualized using the program MAPCHART (Voorrips 2002). The linkage of u0028c (with more than 15% missing data in the original family) to adjacent markers in LG04 was confirmed using a separate, unrelated family from the Whiteshell fish hatchery (West Hawk Lake, Manitoba, CANADA), extracted, amplified and analysed using the previously described procedures.

Synteny analyses

Four datasets were used for synteny analyses: 1) the reference *Esox lucius* transcriptome (see 'RNA-seq and assembly' section of Methods); 2) stickleback protein sequences from the Feb. 2006 Broad/gasAcu1 release; 3) zebrafish protein sequences from the Jul. 2010 Zv9/danRer7 release and; 4) medaka protein sequences from the Oct. 2005 NIG/UT MEDAKA1/OryLat 2 release. Stickleback, zebrafish and medaka sequences and their associated genomic location information were obtained from the UCSC Genome Browser (Rosenbloom et al. 2015). Scaffold locations for northern pike transcripts were obtained through mapping using GMAP (Wu and Watanabe 2005); linkage group assignments followed if the host scaffold had been previously mapped to a group in the genetic map. Using the BLASTx and tBLASTn programs, BLAST alignments (E-value ≤ 1e-5) were obtained between the northern pike transcripts and the proteins of each other fish species. Orthology between northern pike transcripts and other fish protein sequences was determined using the reciprocal best hit (RBH) paradigm requiring at least 50% of each sequence was covered in non-overlapping BLAST alignments (HSPs) from the other. Synteny between two species (Figure 3.6) and scaffold continuity (Additional File C2) was examined by plotting the genomic locations of each sequence in a relevant orthologue pair.

The analysis of synteny between northern pike and Atlantic salmon (Table 3.2) was performed by obtaining the flanking sequence of chromosome-associated SNPs in Atlantic salmon and identifying the strongest BLASTn hits (E-value ≤ 1e-10) between these sequences and northern pike scaffolds with a known linkage group.

Acknowledgements

We would like to thank Doug Watkinson for tissues used in RNA-seq and the staff at the Whiteshell hatchery (West Hawk Lake, Manitoba, CANADA) and the Hackettstown

Fish Hatchery (Hackettstown, New Jersey, USA) for their work in producing and collecting the pike families used in linkage mapping.

Supporting Data

The data sets supporting the results of this article are available in NCBI as part of NCBI BioProject GenBank: PRJNA221548, with the assembly, described under GenBank: GCA_000721915.1. All other supporting information is available as supporting data within the article (Appendix C)

Chapter 4 :

A Chromosome-level assembly of the northern pike (*Esox lucius*) genome traces the evolution of the salmonids following whole-genome duplication

Abstract

The northern pike (*Esox lucius*), a circumpolar freshwater predator, is the most studied member of the Esociformes, a small order of fish that includes pikes, pickerels and mudminnows. The Esociformes and Salmoniformes last shared a common ancestor approximately ~125 million years ago, prior to the whole-genome duplication event that occurred in the ancestor of all salmonids. Given the phylogenetic position and an ancestral karyotype of 2n=50, it was thought pike might be well-suited to act as an outgroup to studies of salmonid genome evolution post-duplication. Indeed, the preliminary northern pike genome presented in 2014 showed very little inter-chromosomal rearrangement in comparison to model species of the Neoteleostei, justifying its use as a proxy for the pre-duplication genomic state in evolutionary analyses of salmonid genomes. To improve the pike genome's ability to act as a reference for salmonid genome analyses we have significantly expanded to a chromosome-level assembly of the northern pike genome. This assembly is now composed of 18,750 contigs (N50 = 126 kb) and 1,211scaffolds (N50 = 7.95 Mb) with half of the genome now represented by the 29 largest scaffolds. Using linkage mapping data as well as synteny mapping to the recently published Atlantic salmon, we have now placed 94% of the sequenced genome into chromosomes. We demonstrate the utility of northern pike to help order and orient small scaffolds in regions that are difficult to assemble in salmonid genomes due to residual tetrasomy. This genome provides an invaluable resource linking ongoing studies of salmonid genomes with other major

teleost groups and should serve as a suitable outgroup in studies of evolution following genome duplication in salmonids.

Background

The northern pike is a fresh and brackish water species that can be found through much of the northern hemisphere (Craig 2008). Used widely in physiological, toxicological and ecological studies, it has been proposed that the pike is approaching the status of a model organism (Forsman et al. 2015). Despite their circumpolar distribution, pike show very little polymorphism throughout their range. Explanations for this suggest population bottlenecks during the previous ice age due to a limited number of refugia (Seeb et al. 1987). This limited variation allows for genome sequence reconstructions without the complexity of multiple copies represented in divergent alleles.

The northern pike belongs to the small order Esociformes, of which there are ~10-13 species (Nelson 2006, Bianco and Delmastro 2011, Lucentini et al. 2011, Denys et al. 2014). The Esociformes sit in an interesting phylogenetic position, being most closely related to the much larger and diverse Salmoniformes (Ishiguro et al. 2003, Campbell et al. 2013). The Salmoniformes, the economically important order comprising fish commonly known as salmon, trout, charr, whitefish, ciscoes and grayling, all share an unusual genome feature: the common ancestor experienced a whole-genome duplication (WGD) approximately 88-103 million years ago (mya) (Macqueen and Johnston 2014). This has made sequencing very difficult, but given the level of importance of the group, two genomes have been published, the rainbow trout (*Oncorhynchus mykiss*) (Berthelot et al. 2014) and the staple of the salmon farming industry the Atlantic salmon (*Salmo salar*) (Lien et al. 2016). The genome of the Atlantic salmon is quite robust and should serve as the reference genome in a variety of ongoing salmonid genome projects currently underway. With the availability of two reference-

quality genomes, we can start to ask larger questions about the genome changes that have occurred post-duplication; using the northern pike outgroup allows a proxy for the state of the genome prior to the duplication event, and allows us to link an approximation of the pre-duplicated salmonid genome to the wider teleost lineages. A preliminary northern pike genome was published in 2014 and was based exclusively on short-read Illumina sequencing technology, using libraries of various insert sizes (Rondeau et al. 2014).

The steady generation of genome assemblies in the teleosts can be linked to rapid advances in DNA sequencing technologies and decreasing cost. Short-read technologies, in particular Illumina-based protocols and instruments (Bentley et al. 2008), have revolutionized our understanding of genome-wide genetic variation, but genome assembly contiguity had not necessarily approached what had been possible with older technologies such as Sanger sequencing. While short-reads are sufficient within unique regions of the genome, in areas where there is a greater level of variation or repetition, short-reads can fail to assemble correctly leaving the genome in heavily fragmented pieces; resolving structural variation in such fragmented regions is extremely difficult. To bridge such regions, sequencing the ends of libraries of increasing fragment sizes allows for bridging and ordering of contigs into scaffolds with gaps of estimated sizes but this strategy often fails to fill in the gaps between contigs, and insert size can often be insufficient to resolve long repetitive elements. Though the majority of coding sequences can be mapped to these new assemblies, new, longer sequencing technologies are required to fully characterize genes, repeat regions and chromosomal structural elements (reviewed in (Chaisson et al. 2015)).

New technologies such as Pacific Bioscience's SMRT (Eid et al. 2009) or ONT's single-molecule nanopore sequencing (Stoddart et al. 2009) allow a slightly different approach to de novo genome assembly. While the higher error rate of both technologies requires a certain level of error-correction with short-reads (e.g. (Koren et al. 2012, Goodwin et al. 2015)) or through hierarchical approaches that rely on

consensus of multiple long-reads (Chin et al. 2013), long-reads have a significant advantage in that they allow for direct characterization through all but the longest repeat regions while also filling gaps between contiguous sequences. These technologies promise improvements in genomes where repetitive content is high, and in polyploid species where identifying differences between duplicated copies may require increased length to identify subtle changes between duplicated copies.

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

In this work, we present a new high-quality northern pike genome with the addition of sequence from large-insert Illumina mate-pair libraries and Dovetail Genomics Chicago/HiRise data for improved scaffolding, as well as long-read PacBio RSII sequencing for gap filling. This additional data has been integrated to produce an assembly, with N50 scaffold sizes exceeding 7.9Mbp and many scaffolds approaching complete chromosome lengths. We utilize this information to compare patterns of evolution in the Protacanthopterygiian lineage to that of the Neoteleostei to evaluate how closely the northern pike represents a pre-duplication state. In addition, we look at the evolution of the recently published Atlantic salmon genome from duplication to the

pseudo-tetraploid state in which it exists today and show how the availability of the northern pike genome may aid in refining the salmonid genomes.

Results and Discussion

Esox lucius Genome assembly

Genome assembly improvements were accomplished using new data generated from a combination of three sequencing technologies. From the addition of 2 lanes of large insert (9-13kb) Illumina mate-pair assembled with ALLPATHS-LG (Gnerre et al. 2011) and 24 SMRT cells of PacBio sequencing incorporated with PB Jelly (English et al. 2012), a total of 1,708 scaffolds were obtained, with an N50 of 4.23 Mb representing just 45 scaffolds (Table 4.1b). Utilizing the linkage map from Rondeau et al. (Rondeau et al. 2014), the scaffolds were anchored to linkage groups. In all, 209 scaffolds were linked by at least one marker to the 25 linkage groups; This version of the genome was submitted to NCBI through WGS accession AZJR03000000.2 and Assembly accession GCA_000721915.2.

Table 4.1: Genome sequencing stats

Genome sequencing stats for version 1-3 of the northern pike genome as well as the resulting chromosomes built from v3.0.

		Number	Bases (Mb)	N50 (Sequence length)	N50 (number of sequences)	Maximum Length (bp)
A) Version 1	**Contigs**	94,266	824	16,910	13,483	232,364
	Scaffolds	5,688	878	700,535	318	5,140,982
B) Version 2	**Contigs**	18,738	892	126,056	1,985	1,089,236
	Scaffolds	1,708	904	4,231,106	45	23,611,278
C) Version 3	**Contigs**	18,750	890	125,635	1,991	1,089,236
	Scaffolds	1,211	904	7,945,253	29	25,961,733
D) Mapped	**Unplaced Scaffolds**	993	108.02	423,854	68	2,378,070
	Chromosomes	25	796.44	-	-	-

Additional scaffolding was performed using the Chicago/HiRise method from Dovetail Genomics (Putnam et al. 2016). Following one lane of sequencing, 537 joins, and 43 breaks were performed by HiRise, with 29 gaps closed. Contig N50 remained effectively unchanged as the number of breaks roughly balanced the number of gaps closed, but Scaffold N50 nearly doubled to 7.945Mb. Details of version 3 of the assembly are summarized in Table 4.1c. The linkage map from Rondeau et al. (Rondeau et al. 2014) was used to order and orient the scaffolds into chromosomes. In all, 218 scaffolds could be linked to a chromosome by the linkage map, accounting for 796 Mbp or 88% of the assembly (Table 4.1d). The addition of new sequencing overall increased scaffold N50 tenfold and nearly doubled the percentage of the assembly assigned to chromosomes, bringing the genome of the northern pike from preliminary assembly to high-quality reference.

Improvements in the assembly of the northern pike genome continue. While this assembly places much of the genome into chromosome sequences, a considerable number remain as unlinked scaffolds. In many cases, these scaffolds are not small, with

53 exceeding 500 kbp and another 206 scaffolds falling between 100-500 kbp. Significant advances continue to be made as we incorporate both sequencing and assembly technologies and programs such as SSPACE-LR (Boetzer and Pirovano 2014), LINKS (Warren et al. 2015) and OPERA-LG (Gao et al. 2016b) . Future directions including long-read only (e.g. in Gorilla (Gordon et al. 2016)) or hybrid rather than tiered assemblies (e.g. *Medicago trunculatus* (Miller et al. 2017)) may result in even greater future improvements to the genome.

An improved genetic map for pike would likely generate even greater improvements. While additional sequencing could help in linking some of these scaffolds to the current chromosomes, the most logical scenario would seem to be to increase the number of markers on the linkage map using one of the Genotyping-By-Sequencing technologies. Unfortunately, the very low level of polymorphism which has been so fortuitous for genome assembly is an impediment to increasing marker numbers through random marker identification. It could be possible with this new assembly to identify and target new highly polymorphic markers in unlinked scaffolds, but the development of these markers would be both time-consuming and expensive; the benefits of definitively linking the remaining scaffolds would need to be clear, particularly if conserved synteny to another teleost could be exploited to make reference-guided approximations (discussed further below). Utilizing northern pike families from outside North America may be an additional option, and indeed work on European northern pike linkage map is in preparation (Pan et al., personal communication). As additional genomic information becomes available from the larger geographic range, it may become possible to identify population-specific genome rearrangements as well as improving the contiguity of the pike genome.

Synteny in Teleostei

In keeping with observations of extensive chromosomal synteny within much of the Teleostei (Mank and Avise 2006), there is a conservation of chromosomal gene

content between northern pike and the well-characterized threespine stickleback. Figure 4.1 shows a one-to-one chromosome-to-chromosome match through most of the chromosomes between the two species, although within-chromosome structure can be substantially rearranged. Most exceptions to the one-to-one relationship seem to be species or family-specific rearrangements, such as the fusion of three pairs of chromosomes in stickleback (stickleback I, IV and VII). Exceptions, however, include a split relative to stickleback IX which appears as two chromosomes in northern pike (Eluc09 and Eluc25), as previously identified in the preliminary published version of the pike genome (Rondeau et al. 2014). It appears that this change occurred specifically along the lineage leading to the Protacanthopterygii as the homeologous duplicates in Atlantic salmon (Eluc09 ~ Ssal2q and Ssal12qa; Eluc25 ~ Ssal4p and Ssal8q) also appear on separate arms (Figure 4.2). Two additional translocations (Eluc05 – GacuV + GacuVI; Eluc14 – GacuXIV + GacuXVIII) also appear to have been previously identified correctly as lineage-specific to pike, as both stickleback and Atlantic salmon display the same breakpoint relative to pike. With this new assembly, we can also identify previously unknown rearrangements. Eluc06 appears to have been rearranged relative to GacuV + Gacu VI prior to the split from the salmonids, while Eluc16 (GacuXVI + GacuXX; given as an example of a stickleback/pike/Atlantic salmon comparison in figure 4.3), Eluc17 (GacuXII + GacuXVII) and Eluc23 (GacuIV + GacuXIV) all appear to be specific to the pike lineage. Full-genome synteny results are summarized in Table 4.2. A more comprehensive pike genome enables a complete characterization of Neoteleostei interchromosomal rearrangements.

Table 4.2: Conserved synteny between northern pike, threespine stickleback, Atlantic salmon and rainbow trout

Eluc	***Gacu***	***Ssal***	***Omyk***
1	1	9qc, 20qb	24,27
2	2	11qa, 26	26,6q
3	3	3p, 14qa	28,8q
4	4	5p, 9qb	14q,25
5	5,6	1qa, 19qb, 28	23,11p,3p
6	6,5	1qa, 18qa, 19qb, 28	23,1q,16p,20q
7	7	4q, 11qb, 13qb	10p,29,12p
8	8	10qa, 16qa, 23	5q,1p,4p
9	9	2q, 12qa	13p,17p
10	10	14qb, 27	14p,18q
11	11	3q, 6p, 23	12q,13q,4p
12	12	3p, 13qa, 15qb	28,16q,9q
13	13	15qa, 20qa, 24, 26	8p,11q,6p,6q
14	14,18	1qb, 6q, 11qb, 13qb, 15qa	5p,4q,29,12p,8p
15	15	1p, 9qa, 11qa	19q,25,26
16	16,20	21, 25	22q,22p,3q
17	17,12	12qb, 22	17q,7q
18	18	6q, 15qa	4q,8p
19	19	10qb, 16qa, 23	2q,1p,4p
20	20	2p, 5q, 21, 25	3p,2p,22q,22p,3q
21	21	19qa, 29	11p,15p
22	2	16qb, 17qa	18p,7p
23	4,14	1qb, 6q, 7q, 15qb, 17qb	5p,4q,21q,9q,15q
24	7	7p, 18qb	21p,9p
25	9	4p, 8q	10q,19p

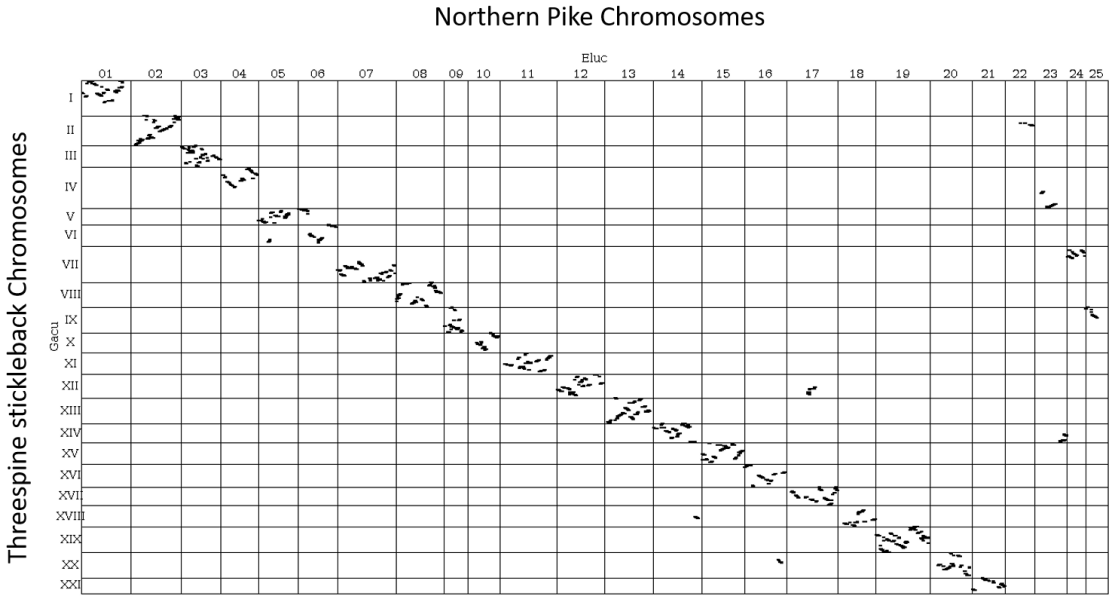


Figure 4.1: Threespine stickleback genomes aligned to northern pike.

Using symap4.2 (Soderlund et al. 2011), the genome of threespine stickleback (gasAcu1) was mapped to the v3.0 assembly of northern pike with default settings. Hits not assigned to syntenic blocks (generally repeat elements) were removed to simplify visual analysis of the results.

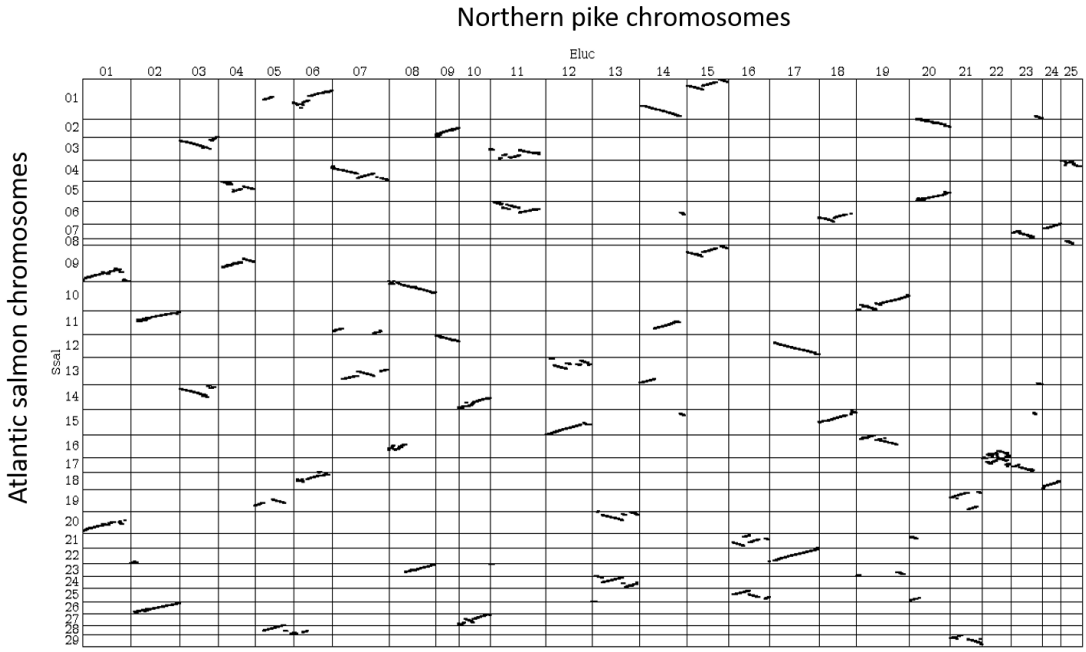


Figure 4.2: Atlantic salmon genome aligned to northern pike.

Using symap4.2, the genome of Atlantic salmon (ICSASGv2) was mapped to the v3.0 assembly of northern pike with default settings. Hits not assigned to syntenic blocks (generally repeat elements) were removed to simplify visual analysis of the results.

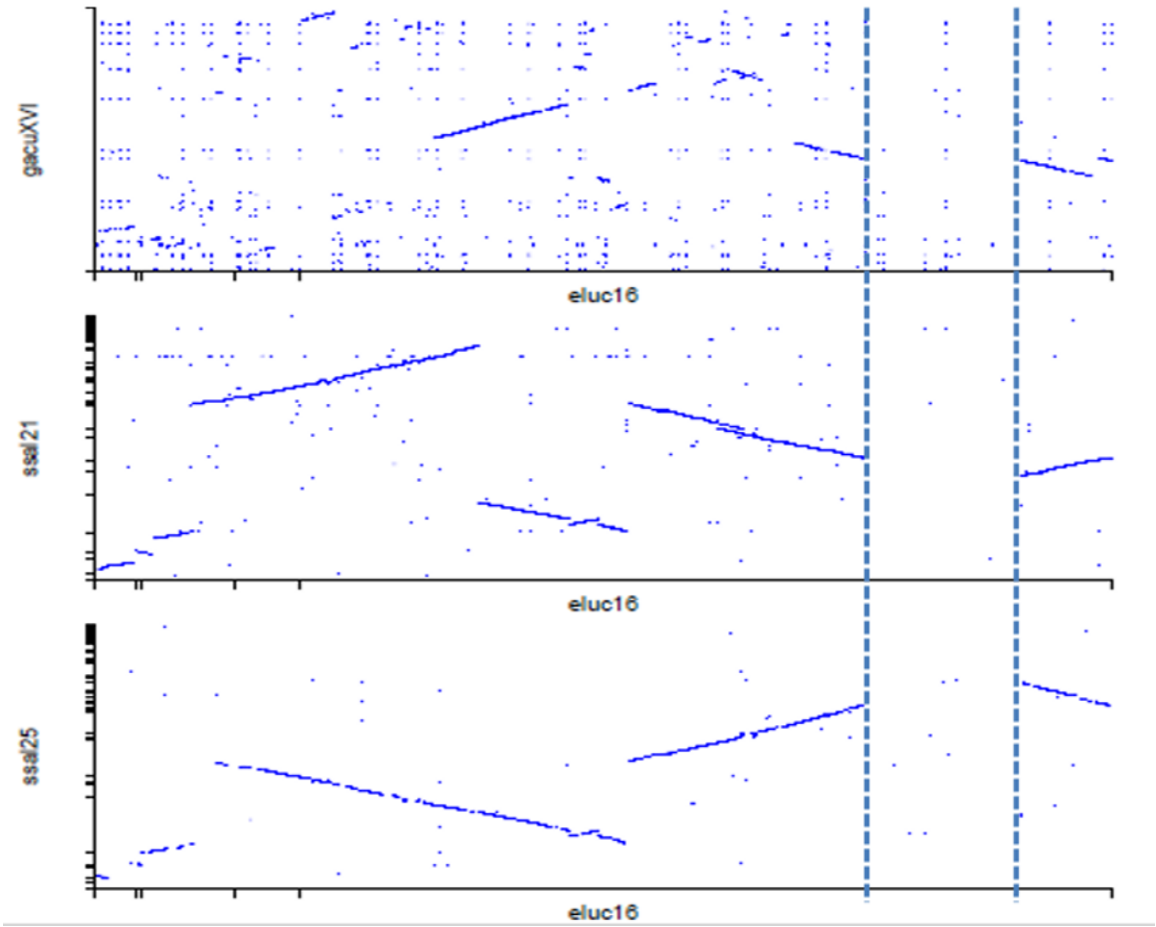


Figure 4.3: Northern pike-specific sequence through mapping to Atlantic salmon and threespine stickleback

Using LastZ, chromosomes ssa21 and ssa25 of Atlantic salmon and GacuXVI of threespine stickleback (y-axis) were mapped to chromosome Eluc16 in northern pike v3.0 (x-axis). Tick marks on the x and y-axis refer to boundaries of scaffolds; dashed line through the plots represent the boundaries for loss of conserved synteny (i.e. pike-specific rearrangement).

Intra-chromosomal rearrangements are frequent as compared to the lack of inter-chromosomal rearrangements demonstrated by the high degree of synteny between northern pike and the Neoteleostei (represented by stickleback in Figure 4.1). Figure 4.4 displays comparative mapping between threespine stickleback chromosome XII (GacuXII) and northern pike chromosome 12 (Eluc12). 61 distinct linear blocks (>50,000bp in pike) can be identified between these two chromosomes alone, with over

1,000 such blocks identified in the full genomes. It can be noted the length of the largest such blocks is quite significant – in the example in Figure 4.4, the longest linear block is just under 3Mbp and contains just over 200 genes (NCBI Annotation release 102). Symap v4.2 identified 76 blocks between threespine stickleback and northern pike that exceed 1Mbp in length. While it is outside the aim of this work, the size of the blocks that remain between two species suggests that reconstructing the ancestral chromosomal gene content by following inversions and translocations through to a common origin should be possible as has been demonstrated in angiosperms (Murat et al. 2017) once additional Protacanthopterygiian genome sequences become available. Therefore, while chromosomal gene content is largely conserved over the ~200 million years (Near et al. 2012) that separate these two species, gene order within chromosomes is conserved in blocks less than 3Mb, with an average block size of 100 kbp.

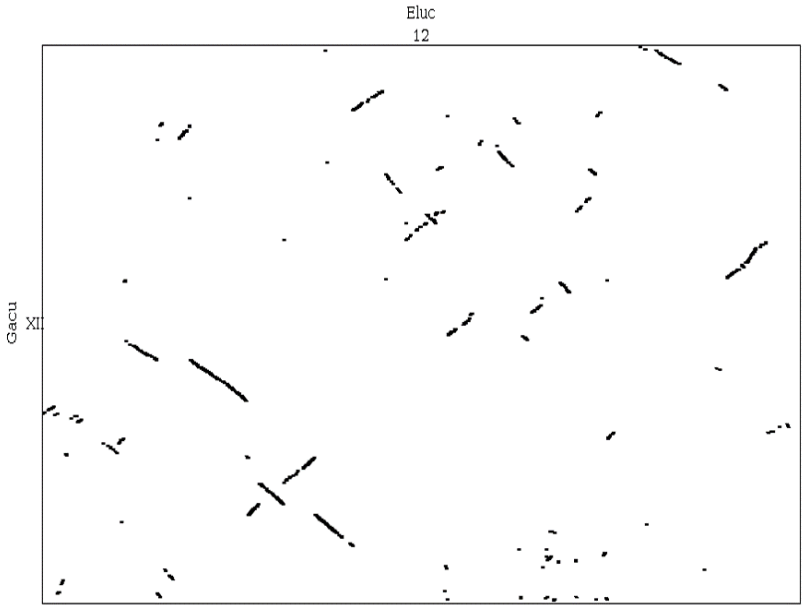


Figure 4.4: Threespine stickleback genomes aligned to northern pike – GacuXII vs Eluc12.

Using symap4.2, chromosome GacuXII of threespine stickleback (gasAcu1 – Y-axis) was mapped to the v3.0 Eluc12 of northern pike (X-axis) with default settings. Hits not assigned to syntenic blocks (generally repeat elements) were removed to simplify visual analysis of the results.

Synteny with the Atlantic salmon:

Northern pike as an outgroup to Atlantic salmon can be used to pronounced effect in the analysis of gene evolution within salmonids. Pike has been used to relate the salmonid-specific whole genome duplication (Ss4R) to extensive gene duplications in chemokine receptors (Grimholt et al. 2015). It has been used to anchor and identify regions of lineage-specific or delayed re-diploidization in the salmonids (Robertson et al. 2017). Expression levels of transcripts generated as part of the previous pike genome have been used to suggest regulatory neofunctionalization as a major fate of retained duplicated post-Ss4R genes (Lien et al. 2016, Sandve et al. 2017). With the new genome assembly in northern pike, it also now becomes possible to examine the detailed chromosomal structure, effectively identifying conserved synteny between linkage groups (Rondeau et al. 2014). With the improvement in assembly contiguity and scaffold integrity, and the refinement of order and orientation of scaffolds into linkage groups, we can now begin to tackle full chromosomal comparisons between the pike genome and the post-Ss4R duplications found in recent genomes of Atlantic salmon and rainbow trout.

The chromosome-wide comparison of northern pike and Atlantic salmon is shown in Figure 4.2 (and see Additional file D1 for single-chromosome comparisons). Immediately obvious, as would be expected from a WGD is a general 2:1 chromosome relationship between salmon and pike. Perhaps less expected is the extensive conserved linearity between the two genomes, with most of the rearrangements limited to 1-4 major intra-chromosomal inversions (see Additional file D1 for a chromosome-by-chromosome breakdown of major syntentic relationships). Fusions and fissions events between northern pike and Atlantic salmon largely fit low-density genetic map comparisons examined by Sutherland et al. (Sutherland et al. 2016). However, the ability to look at the whole genome sequence data allows for identification of several inversions that are associated with fusions. We can also identify several subtle re-arrangements that involve the movement of smaller portions of chromosomal arms

(following naming scheme in Sutherland et al.: 8-19-22; 12-14-18-23; 2-17; 7-12-14-23; 2-7-14). The ability to trace all the fine-scale rearrangements, within Atlantic salmon requires additional salmonid genomes and comparative analyses; however, most of the major rearrangements since the last common ancestor between pike and salmon can now be identified.

Guiding assemblies

An emerging trend in several genome projects is the use of reference guided assembly to help order and orient genome assemblies, either through direct integration of the reference into the assembly procedure (e.g. (Wang et al. 2014)) or by post-assembly integration into chromosomes (e.g. (Shen et al. 2016)); most examples utilize very closely related species, ideally between species at the genus level, or at most at the level of the family. However, given our difficulty in generating a marker-dense linkage map, the unavailability of additional genomes within the Esociformes, and the observed co-linearity identified between Atlantic salmon and northern pike, we sought to determine how well a salmonid genome could serve to guide additional pike scaffold integration into chromosomes. Using the largest (>50 kbp) unlinked scaffolds, we mapped the longest two hits to the Atlantic salmon genome using LastZ, and visually analysed the results, looking for unlinked scaffolds (or a series of unlinked scaffolds) that linearly connect two adjacent scaffolds by way of both Atlantic salmon homeologues, as demonstrated by example in Figure 4.5. In all, 211 additional scaffolds could be incorporated in such a way, increasing overall genome incorporation from 88% to 94% of assembled bases. While we debated as to whether to incorporate such comparative-mapping into the published NCBI reference genome, we ultimately decided the evolutionary distance between the two was too far to justify incorporation without direct linkage-based support within the species. However, in instances where such reference-guiding may be justified (e.g. gene candidates in QTL or GWAS), this “synteny”-based genome shall be available on our project site,

http://web.uvic.ca/grasp/pike. The long-term conservation of genome structure within the salmonids following genome duplication, therefore, provides significant information for improvements to the northern pike genome, even given the ~125 million years since divergence.

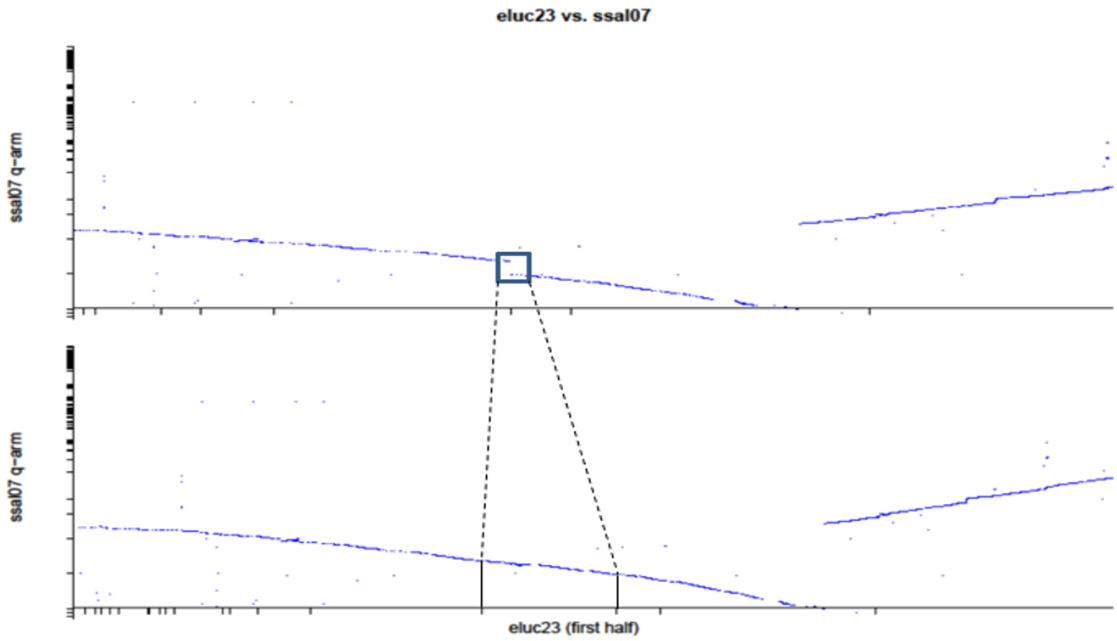


Figure 4.5: Adding additional scaffolds through comparative mapping – Eluc23 vs Ssal07q

Using LastZ (Harris 2007), the q-arm of chromosome Ssal07 of Atlantic salmon (Lien et al. 2016) was mapped to chromosome Eluc23 in northern pike v3.0. The added scaffold was identified through parsing lastZ results mapping unlinked scaffolds to Atlantic salmon chromosomes. Scaffolds were added when both ends created linear continuity through a previous scaffold breakpoint. Tick marks on the x and y-axis refer to boundaries of scaffolds.

The release of the Atlantic salmon genome represents an incredible resource to aid in the assembly of other ongoing salmonid projects (as demonstrated in (Lien et al. 2016) using rainbow trout assembly v1.0 (Berthelot et al. 2014)). However, it also has some unique difficulties that persist or are a result of the genome duplication. The high repetitive content in the salmonid genomes remains a significant hurdle to closing scaffold gaps. Regions of residual tetraploidy or delayed re-diploidization remain

particularly difficult to resolve. This problem can be demonstrated in that almost 1/4 of assembled bases have yet to be incorporated into chromosomes in Atlantic salmon despite almost 200K markers available for high-density linkage mapping.

In genomics, regions that are highly repetitive or regions with highly similar, subtle differences between duplicates tend to collapse into short, fragmented contigs (Margarido and Heckerman 2015). While increasing read-lengths will aid in regions where there are structural differences between copies, the current error-rate will likely mean error correction algorithms, particularly those that involve short-read correction, will fail to resolve the duplicates correctly. Other emerging sequencing technologies may help resolve such regions, although results thus far have been inconclusive. Chromosome conformation capture techniques have been used successfully in the allopolyploid duplicated genome of *Xenopus laevis* (Session et al. 2016), although in our hands the current version of the Chicago/HiRise has had difficulty scaffolding regions of delayed re-diploidization in autopolyploid salmonids (Rondeau and Koop, unpublished Results). Such differences may be a result of the origin of the duplication (allo vs auto-tetraploidy), repetitive content in the genomes (~40 vs ~60%) or significant delays in re-diploidization (Robertson et al. 2017). We are encouraged by the possibility of long-range haplotyping of double-haploid individuals (effectively diploid in regions of high retained tetraploidy) using 10X chromium, although this may require significant adaptation as current protocols are not validated in non-diploid species. Thus, improvements in such difficult to resolve regions, particularly as new salmonid genomes are released, may not be simple.

One potential path forward may be to use northern pike as a reference in difficult regions of the Atlantic salmon genome. As these sequences are not recently duplicated, assembly contiguity tends to remain high in the northern pike assembly. Given the degree of conserved-linkage observed between the two species, it suggests that by guiding salmon assemblies of particularly difficult regions with pike data (a “pike-ized” configuration), in conjunction with a genetic map and other linking data,

pike could inform the most likely order and orientation of additional unmapped contigs into an approximated genomic position. In figure 4.6, we use northern pike to identify scaffolds which are effectively duplicated on a single chromosome (ssa26) in the Atlantic salmon reference, likely due to misassignment of highly similar SNPs to the alternate homeologue. In such a way, potential misassignments could be identified and relocated to a more likely position – in this case by filling in the missing sequence in ssa11. In figure 4.7, we demonstrate the possibility of reference-guiding a cloud of small contigs (figure 4.7a). While all such contigs have been physically linked to a chromosome by way of the high-density linkage map, resolving true order and orientation, particularly when based on the mapping of a single marker, is unrealistic. By re-sorting and orienting scaffolds by way of a pike reference, we can begin to resolve a more likely placement in these difficult chromosomal regions. While small contigs composed primarily or completely of repeat elements will not be possible to place in such a way, this method does suggest a practical path forward in resolving gene order and general structure in such highly similar regions. A combination of long-reads bridging and capturing regions of unique repetitive content with reference-guided order and orientation by way of a non-duplicated outgroup such as pike should at a minimum allow the placement of most genes into a chromosomal context, with future technologies allowing for eventual confirmation and validation.

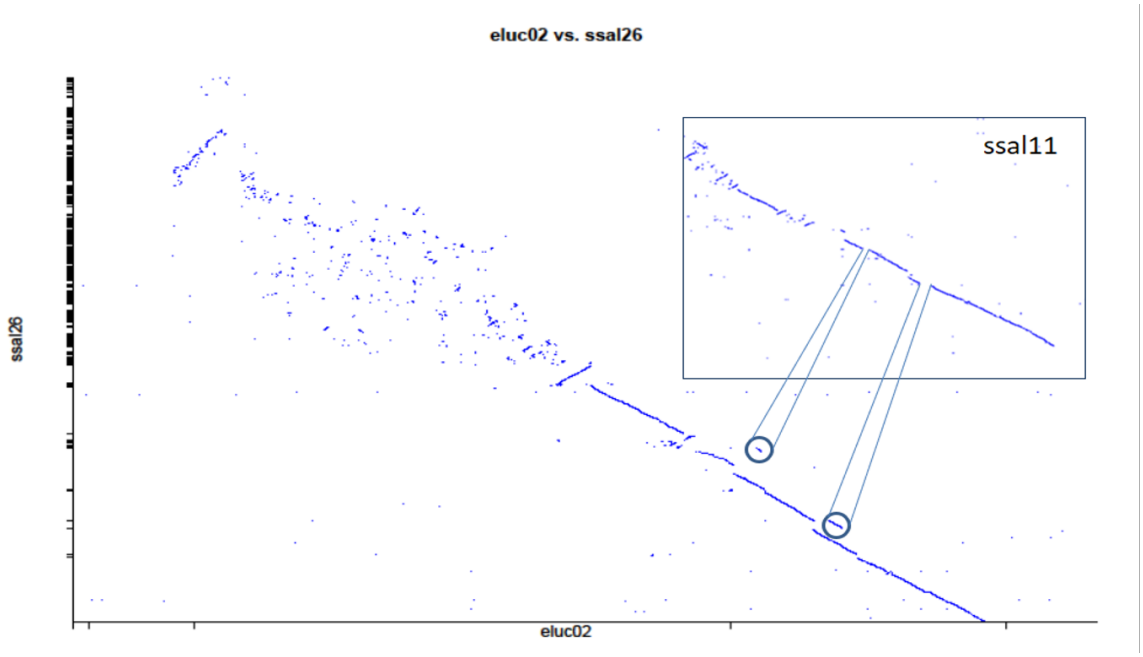


Figure 4.6: Identifying mismapped scaffold candidates in Atlantic salmon – Eluc02 vs Ssal26 and Ssal11

Using LastZ, chromosome Ssal26 of Atlantic salmon was mapped to chromosome Eluc02 in northern pike v3.0. Potentially misplaced scaffolds were identified by looking for instances where two best scoring, overlapping hits were associated with the same Atlantic salmon chromosome. Evaluation of missing sequence in homeologous chromosome was identified visually from LastZ plots, displayed here as an inset of a portion of Ssal11. Tick marks on the x and y-axis refer to boundaries of scaffolds.

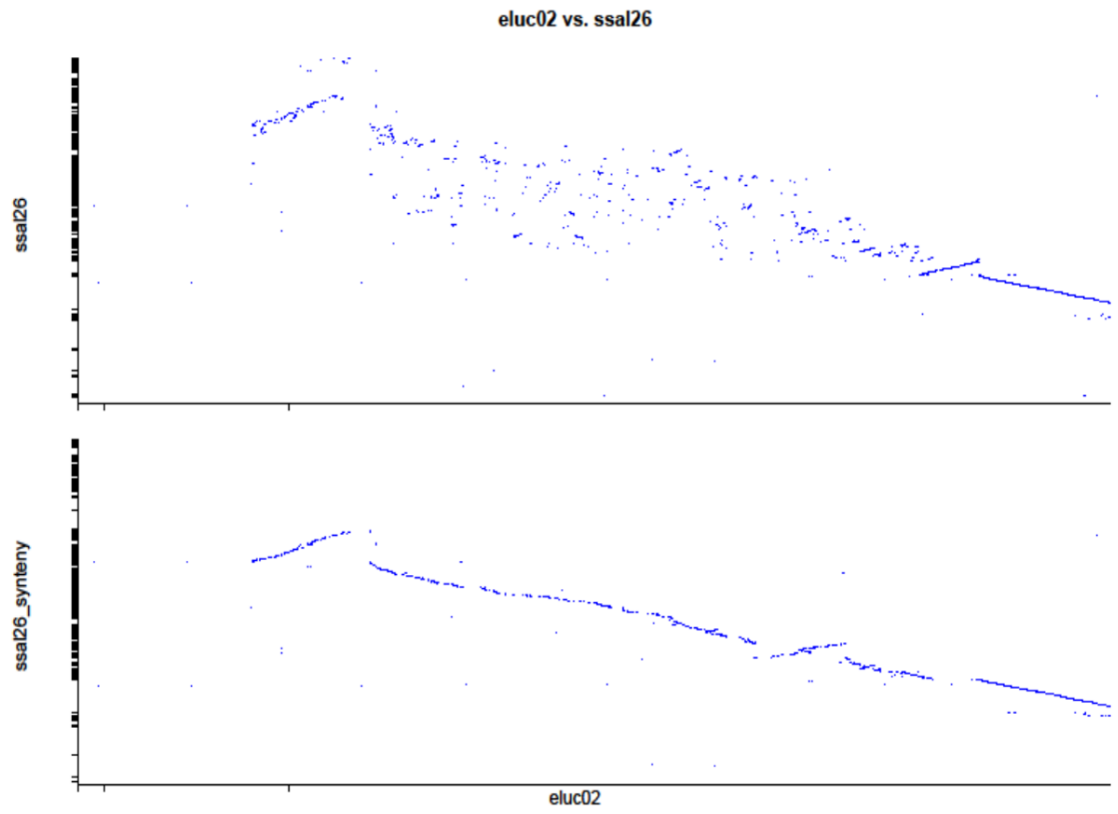


Figure 4.7: "Pike-ized" regions of residual tetraploidy – Eluc02 vs Ssal26

Using LastZ, chromosome Ssal26 of Atlantic salmon was mapped to chromosome Eluc02 in northern pike v3.0. Small contigs were sorted based the order and orientation of the best-scoring LastZ hit relative to pike. Tick marks on the x and y-axis refer to boundaries of scaffolds.

Conclusions

The new genome assembly of the northern pike provides chromosome-level resolution with which to bridge the duplicated Salmoniformes to the remaining Teleosts. Scaffold N50 has increased from 700 Mbp to 7.9 Gbp using a combination of new sequencing technologies, with the length of the longest assembled scaffold exceeding the length of the shortest chromosome. The new genome confirms the retention of conserved synteny between the Neoteleostei and the chromosomes of pre-duplicated Protacanthopterygii and identifies blocks of up to 3Mb of conserved gene order. While a small number of lineage-specific re-organizations can be found in pike,

this study largely reconfirms that pike represents a largely ancestral organization of the genome as had been suggested by the karyotype. Comparative mapping between pike and Atlantic salmon shows an extensive conservation of co-linearity between the two genomes and allows for both genome improvements in pike from Atlantic salmon as well as to the Atlantic salmon reference using pike. Future work on incorporating additional sequencing technologies along with reference-guided assembly should make significant improvements in genome assembly in difficult to resolve regions of the re-diploidizing genome of modern salmonid.

Methods

Sequencing and Assembly

All sequencing was performed with archived material from a single male individual utilized in the first version of the northern pike genome (Rondeau et al. 2014), and was one of three individuals originally collected for EST sequencing (Leong et al. 2010); this individual was sourced from an introduced population in Charlie Lake, BC, CANADA (Nordin and Pommen 1985). DNA was extracted from spleen, brain and head kidney tissues stored at -20°C in RNAlater utilizing a standard phenol extraction (Sambrook et al. 2000). Mate-pair sequencing targeting 9-13kb size fragments were sequenced to overcome gaps found in the original northern pike assembly. 4 libraries pooled into two lanes with two libraries each were sequenced on an Illumina HiSeq2000 (Genome Sciences Centre, BC cancer agency). 467,818,663 paired-end reads were generated and added to the sequences of the previous assembly (used in assembly v1.0; Table 4.1a; GenBank: SRX494131). All reads were utilized for a new ALLPATHS-LG genome assembly that was performed following the same protocol in Rondeau et al. (Rondeau et al. 2014). Contigs were examined for long runs of ambiguous bases and trimmed/split if necessary. Contigs were removed that were primarily vector sequence (BLASTn 1e-50) and contigs were trimmed internally and at terminal ends for any vector

sequence according to NCBI's VecScreen search recommendations (www.ncbi.nlm.nih.gov/tools/vecscreen/). Duplicate and near duplicate (>99% identity; >95% length) contigs were removed and mitochondrial contigs were removed (BLASTn 1e-5). This resulting contig set was compared to GenBank's non-redundant nucleotide database for contamination (BLASTn 1e-25). Potential non-vertebrate sequence matches were flagged and manually inspected to determine classification.

Additional libraries were constructed and sequenced by Pacific Biosciences (PacBio) RSII using P6/C4 SMRT chemistry (Norwegian Sequencing Centre, University of Oslo). Size selection was >5kb (4 libraries, 19 cells), >7kb (1 library, 1 cell) and >9kb (1 library, 4 cells). Across 6 libraries and 24 SMRT cells 5,414,742,446 bases in 1,205,909 sequences were produced. PBJelly (English et al. 2012) was used in combination with the ALLPATHS-LG scaffolds primarily to fill in scaffold gaps as well as to join scaffolds. In analysing resulting scaffolds, five cases were identified where markers from one scaffold mapped to two different linkage groups. The results of the syntenic comparison to *Salmo salar*, it became clear that these were most likely due to chimeric scaffolds. Thus, these five scaffolds were broken in two at the scaffold gap closest to the syntenic break. These scaffold sequences have been uploaded to NCBI under BioProject ID PRJNA221548, accession AZJR00000000; this version is the second version, AZJR00000000.2.

The third round of sequencing was undertaken using the Chicago/HiRise sequencing and assembly protocol from Dovetail Genomics. From one Chicago library, 84 million read pairs were sequenced using 2x101bp Illumina HiSeq. 57.8X estimated coverage from 1-50kb pairs were input into HiRise scaffolder (Dovetail genomics: 02-July-2016) along with version 2.0 of the assembly described in the previous section. These scaffold sequences have been uploaded to NCBI under BioProject ID PRJNA221548, accession AZJR00000000; this version is the third version, AZJR00000000.3.

Assembly into Chromosomes:

Primers used to amplify microsatellites in the first-generation linkage map were mapped against the assembled scaffolds (≤1e-01). Matches from both primers were used to place markers onto the scaffold. Scaffolds were ordered and oriented (when possible) into chromosomes utilizing the sex-specific linkage maps; when order or orientation was ambiguous (i.e. a single marker linked it to a chromosome), comparisons to the *Salmo salar* genome (Lien et al. 2016) with LastZ were utilized to predict the most likely placement (see below); sequences submitted to NCBI were oriented such that the most likely orientation was in the forward direction, but only for scaffolds where two or more markers could confirm direction was unambiguous orientation described in the agp files accompanying the submission. Likely errors in orientation identified in v2.0, primarily in LG11 and LG25 (Yann Guiguen, Qiowei Pan and Rene Guyomard, Personal Communication) were also corrected at this stage.

Genome Alignments

Repeat masking for northern pike, and Atlantic salmon follows the protocol described in detail in the methods of the Atlantic salmon genome (Lien et al. 2016); equivalent custom repeat libraries were built for northern pike (Minkley et al., In preparation). Symap 4.2 (Soderlund et al. 2006, 2011) was used to align repeat masked genomes of northern pike and Atlantic salmon (GenBank:GCA_00023375.4) as well as masked versions of medaka (oryLat2) (Kasahara et al. 2007), stickleback (gasAcu1), (Jones et al. 2012) and zebrafish (danRer10) (Howe et al. 2014) from the UCSC genome browser (Rosenbloom et al. 2015) in order to identify homologous syntenic regions, identify conserved linkage blocks as well as generating images in Figure 4.1 and 4.2; for the figures, hits that did not fall into blocks identified by symap were removed. With homologous regions identified, LastZ version 1.03.66 (Harris 2007) was used to align individual chromosomes across species. The flags used in generating the images were –nochain –gfextend –ungapped –both –ambiguous=IUPAC –identity=75.0..100.00 and –

matchcount=100; where feasible, overlapping hits, mostly from residual repeat elements not otherwise masked were removing using the package GenomicRanges (Lawrence et al. 2013) in R (R Core Team Computing 2017). Re-ordering of the small scaffolds in Atlantic salmon was performed by taking the position and orientation of the longest, non-overlapping hit of a repeat-masked genome as identified by LastZ and sorted to reflect the order in the northern pike genome.

Acknowledgements

The authors would like to acknowledge Frank Koop for the original collection of the northern pike tissue used in this work. This work was supported by the University of Victoria, Compute Canada, and the Natural Sciences and Engineering Research Council of Canada.

Supporting data

The data sets supporting the results of this article are available in NCBI as part of NCBI BioProject GenBank: PRJNA221548, with the assembly, described under GenBank: GCA_000721915.3. All other supporting information is available as supporting data within the article (Appendix D)

Chapter 5 :

Discussion, Future Directions and Conclusions

Mapping methods

Until advances in genome sequencing allow full chromosomal sequences, even the generation of new, reference quality sequence assemblies needs to be oriented and ordered into chromosomal sequences using a map. Mapping techniques can be broadly separated into two groups (Brown 2002): those that query the structure of the DNA are defined as physical maps (e.g. optical mapping (Schwarz et al. 2009)) and those that examine the inherited linkage of genetic features are genetic maps. As genetic maps can more easily be linked to genes or traits in non-model organisms, the basis of mapping in this work relies on the latter.

As introduced in Chapter 1, while microsatellites represented a major advance in generating linkage maps, advances in next-generation genotyping procedures allowed the use of relatively large numbers of SNPs without prior knowledge of the genome. These advances, therefore, beg the question – given that microsatellites were used in both studies, how would we approach marker generation and genotyping in the two maps presented were we to begin again? For sablefish, my answer would be relatively straightforward. Sablefish represents the perfect candidate for restriction-site associated DNA (RAD) mapping, in that there was limited prior knowledge on SNP locations and no closely related species from which large numbers of microsatellites could have been obtained. While the Genotyping-By-Sequencing protocol we performed was meant to fulfil this purpose, the enzyme and pooling depth chosen left individual loci too dilute. Ultimately, subsequent studies with RAD mapping in sablefish populations have proven the utility of such an approach (Jasonowicz et al. 2017). One potential downside is that SNPs identified by this method are randomly distributed

within the genome, unlike our microsatellite approach through which we can target conserved sequences such as genes. However, the standard RAD sequencing protocol (Amores et al. 2011) also allows a randomly-sheared end to maximize available sequences for comparative mapping. If these randomly-sheared ends could then be linked to the contigs of a preliminary genome assembly as an intermediary (Rondeau, Leong, Campbell and Koop; AWGY01000000.1), the comparative mapping would likely yield enough links to stickleback to refine our analysis of conserved linkage. While utilizing microsatellites at the time allowed the construction of the linkage map and ultimately was sufficient to associate an important trait through comparative mapping, the cost and time savings today would likely have made RAD mapping the optimal choice.

For northern pike, the solution is not quite as clear. As demonstrated in Chapter 3, we attempted the standard RAD mapping protocol, but a lack of polymorphism rendered this technique insufficient for generating the number of necessary markers. This was particularly problematic as individual SNPs were unlikely to be shared across maps from each sex or between different families. The relatively high polymorphism of the microsatellite markers and the shared polymorphism across both sexes allowed for the generation of a map of sufficient coverage to link 88% of our final genome to chromosomes, and no alternative method would have necessarily offered obvious benefits. Future improvements may lie in generating more polymorphic families by utilizing European populations with greater polymorphism. Alternatively, full-genome approaches to characterize genome-wide variation could allow the construction of a SNP chip, allowing a more targeted analysis of the most polymorphic SNPs. Such a characterization of genome-wide variation through whole-genome re-sequencing is currently underway. The relative benefit of investing in such mapping projects, versus the immediate value of the reference-guided synteny approach demonstrated with Atlantic salmon in Chapter 4, remains to be determined.

Continued progress with genomic tools developed for sablefish

While my work on sablefish was primarily completed with the publication of Chapter 2, significant advances have been made by researchers in Washington state, particularly by J. Adam Luckenbach and colleagues with NOAA Northwest Fisheries Sciences Center. Molecular characterization of sex-specific genes and pathways (Guzmán et al. 2013, Smith et al. 2013, Fairgrieve et al. 2016) have advanced our understanding of sex-determination pathways, not only in sablefish but in teleosts in general. Characterization of the gonadotropin pathway has led to the identification of the role seasonal capture plays in gametogenesis while in captivity. This understanding potentially paves the way for an effective solution to problems of reproductive dysfunction in sablefish aquaculture (Guzmán et al. 2013, 2015a, 2015b, 2017). The primary advances of work in sablefish, however, related to my work on the identification of sex-determination markers and how these markers help to develop sex-reversed sablefish and an all-female aquaculture line. With the sexual size dimorphism in growth rate (Echave et al. 2012), the generation of an all-female line would more efficiently and rapidly generate individuals of market size while also potentially allowing the production of reproductively sterile fish (Donaldson 1996, Piferrer 2001). Through a series of studies using the genotypic sex markers developed in Chapter 2, Luckenbach and colleagues were able to show that temperature alone is not enough to modify phenotypic sex, which supports a strictly genetic sex -determination system (Lee et al. 2017). It was further demonstrated that hormonal treatment, using methyltestosterone (Luckenbach and Fairgrieve 2016, Luckenbach et al. 2017) over a sufficient period of time, allowed the generation of genetically female and phenotypically male individuals that could be used as broodstock. While a direct sex-reversal using estrogen hormone was also possible (Luckenbach et al. 2017), the exposure of fish to be consumed as food to hormones is controversial and, in some jurisdictions, banned outright (Budd et al. 2015). Thus, the ability to generate all genetically female sablefish using sex-reversed parents, bypassing direct hormone application to food fish, is a significant step forward.

The ability to detect such neomale brood is probably the most practical benefit to industry that comes from the work presented in Chapter 2. As strategies to induce sexual maturation in cultured sablefish are refined and a stable, multi-generational broodstock is developed, mapping additional, multi-genic traits will become possible using much of the additional resources developed in this work.

Much of the scientific attention generated from this work has been directed at the identification of the sex-determination gene candidate. Sex-determination genes in fish have not been well-characterized, with only a handful of such genes being identified across the diverse teleost lineages. Furthermore, different sex-determination genes have been identified in different species, often with little conservation within a taxonomic group (Heule et al. 2014, Pan et al. 2016). While the rapid turnover in these genes ultimately leads to requiring individual identification of sex-determining mechanisms in each individual species, the pool from which these genes are drawn appears to be relatively small. Continued re-emergence of the same or similar genes in divergent lineages have now been described (reviewed in (Pan et al. 2016)). Such a system, in which several candidates are known, but the exact gene in each species remains to be identified, represents an ideal system with which to demonstrate the power of conserved synteny. The association with sex and *gsdf* in sablefish may seem a niche topic at first glance, but once associated with a conserved syntenic region of a model organism, such as stickleback ChrXIV, it provides evidence in support of other such species for which sex-determination can be associated with a homologous chromosome. Such gene candidate associations can be made with ChrXIV in a lineage of rockfish (Fowler and Buonaccorsi 2016) and in Atlantic halibut (Palaiokostas et al. 2013b), further demonstrating the power of conserved synteny in non-model systems.

An additional use of the genomic data produced as part of the sablefish work has been in the ability to design cross-species conserved primers to interrogate specific genes in species for which little to no sequence data has been produced. The resources generated as part of Chapter 2 continue to represent the second largest pool of

assembled transcripts for the Perciformes (NCBI – September 2017). In combination with the threespine stickleback, this conservation has allowed me to design primers to highly conserved portions of many potential sex-determination genes and to screen related lineages for potential sex-determining gene candidates. This approach has been successful in linking duplications of the Anti-Müllerian hormone (*amh*) in the lingcod (Rondeau et al. 2016) and yelloweye rockfish to male phenotypic sex (figure 5.1). As *amh* is a known sex-determination gene from other teleosts (Hattori et al. 2012, Eshel et al. 2014, Li et al. 2015), this cross-species primer design has allowed the identification of a strong sex-determination gene candidate in two species with no prior genomic information.

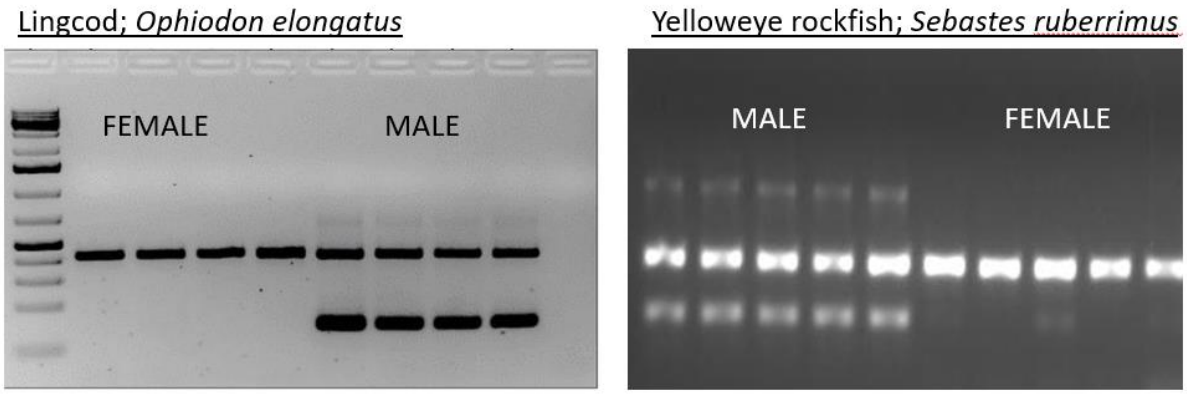


Figure 5.1: Sex-specific PCR assays designed for lingcod and yelloweye rockfish using highly-conserved domains between threespine stickleback and sablefish.

Fate of gene duplicates – pike as a proxy for last common ancestor

The origin and fate of new genes is a fascinating subject lying at the heart of the genomics era. Along with the increasing complexity of eukaryotic species through evolution comes an increasingly complex network of functions and pathways, driven by the emergence of genes that can perform new functions. The origin of such genes can be separated into five categories. The first is completely *de novo*, in which gene origination emerges from a non-coding region of DNA (e.g. (Levine et al. 2006, Begun et

al. 2007)). Second is gene fusion or fission whereby two genes become one (Shtivelman et al. 1985) or whereby one becomes two (reviewed in (Snel et al. 2000)). The third, domestication of transposable elements, can incorporate the protein product generated by the “selfish” element into controlling regulatory factors functioning within the genome (reviewed in (Volff 2006)). The fourth, lateral gene transfer, while relatively common in prokaryotes (Ochman et al. 2000), is rare in multicellular eukaryotes; this passing of genes between species can have major phenotypic effects where it occurs (Wybouw et al. 2016). The final category, gene duplication, involves the duplication of one gene into two copies. This last category is thought to be the most frequent method of new gene generation (Lipinski et al. 2011). While gene duplicates can be generated in a number of ways, the most rapid increase is through whole-genome duplication (WGD), which immediately doubles all of the genes within a genome (Ohno et al. 1968, Ohno 1970). It is this doubling of the genome that has been the driving force behind the work presented in Chapters 3 and 4.

The fate of gene duplicates has been a central question in genomics. It is generally well accepted that most genes that arise through gene duplication are lost, often quite quickly (Lynch 2000). Following genome duplication, in which a doubling of all genes initially necessitates a dosage balancing effect, such loss and reorganization can be delayed (Conant et al. 2014), although the ultimate fate remains the same (Bowers et al. 2003, Paterson et al. 2004, Aury et al. 2006). Of interest, though, is the fate of genes that are not lost. There are two generally proposed models of retention. The first is neofunctionalization (Ohno 1970), whereby one copy of a gene maintains ancestral function while the second is free to undertake a new function (e.g. (Zhang et al. 1998)). The second is subfunctionalization (Nowak et al. 1997, Lynch and Force 2000), whereby two duplicates each maintain a share of the original function (e.g. (Force et al. 1999)). Although alternative fates have also been proposed (e.g. (Des Marais and Rausher 2008)), it is generally the disentanglement of the two fates above that have fascinated biologists through the genomic era.

The primary fate of retained gene duplicates following WGD remains unclear. In *Drosophila*, recent work implicates neofunctionalization as the fate of 65% of retained duplicates, with subfunctionalization detected in only 1% of cases (Assis and Bachtrog 2013). This has tended to be the primary result of retained duplicates following single-gene duplication, but results from whole-genome duplication have tended to give more conflicting results. While general support for neofunctionalization has been described following duplication in some lineages (Sémon and Wolfe 2007), support for subfunctionalization has been reported in others (Gout and Lynch 2015). An additional wrinkle that may contribute to conflicting results is the method of genome duplication (either allopolyploidy (between taxon) or autopolyploidy (within taxon)). The functional difference between gene duplicates from allopolyploidy or autopolyploidy methods remains unknown (Glover et al. 2016). The effect of the type of genome duplication on gene duplication, therefore, remains an open question.

With our co-development of genomic resources in both Atlantic salmon (Lien et al. 2016) and northern pike, we had the opportunity to look at the differential regulation of expression across our common gene atlas presented in Chapter 3 and an equivalent atlas produced in Atlantic salmon. Gene duplicates that were maintained from the salmon-specific autotetraploid 4R duplication and for which a single orthologous gene could be found in pike were examined for the fate of the duplicates. Specifically, correlation of expression to the ancestral proxy, northern pike, was examined between duplicates such that they could be separated into “more conserved” and “more diverged” duplicates based on expression patterns and by looking for expression within or across tissue-specific co-expression clusters. Such an analysis identified 3,028 triplets for which evidence implicated neofunctionalization as the primary candidate in retained-duplicate fate versus only 167 for which evidence of subfunctionalization could be well characterized (Lien et al. 2016). This was based on evidence of duplicates with expression patterns that clustered with two different tissue types one of which was the ancestral pike-type. This result appeared to be at odds with the fate of duplicates

observed following the teleost-specific duplication (Braasch et al. 2016) using spotted gar as outgroup, although a recent reanalysis of both sets implies the same fate would have been observed given the same statistical methods, with neofunctionalization best matching simulations of both potential fates (Sandve et al. 2017). Thus, the resources produced for northern pike allowed for the determination that, in the Salmon-specific 4R autopolyploid genome duplication, neofunctionalization predominates. Extending such an analysis to the functional regulation of expression remains to be determined (see future directions).

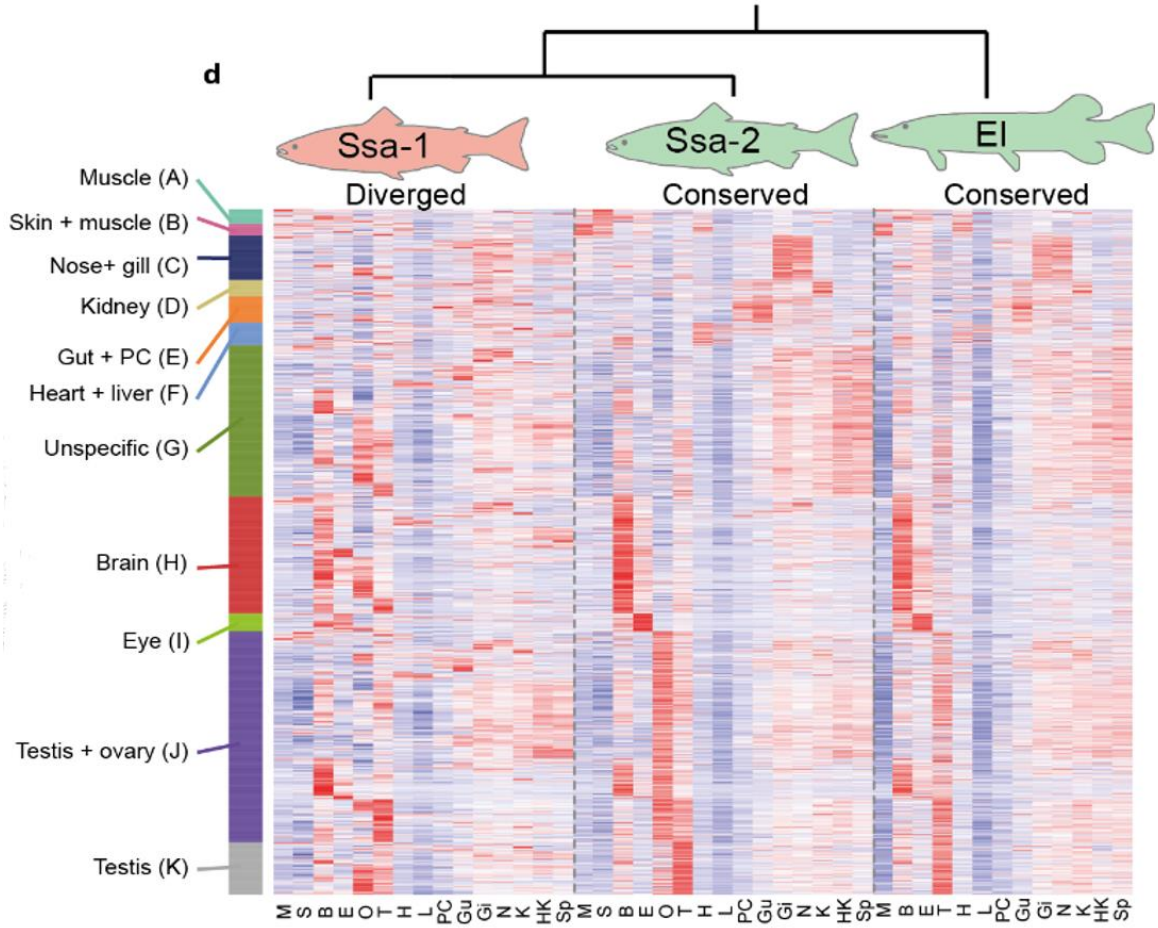


Figure 5.2: Heatmap of 2,272 triplets (two salmon homeologues and a pike orthologue), in which one of the Atlantic salmon homeologues has diverged in gene expression regulation (Figure 4b, Lien et al. (Lien et al. 2016)**).**

The X-axis identifies individual tissues, aligning with those sequences in Chapter 3; M=muscle, S=Skin, B=Brain, E=Eye, O=Ovary, T=Testis, H=Heart, L=Liver, PC=Pyloric Caecum, Gu=Gut, Gi=Gill, N=Nose, K=Kidney, HK=Head Kidney, Sp=Spleen. Y-Axis displays genes, sorted into clusters based on expression

While the focus on gene duplicates following duplication has been primarily aimed at the overall fate, it can be noted that specific examples of duplication can often have significant consequences. As discussed in the previous sections, the continued gene duplication in the gene *amh* is the primary candidate for master sex-determinant in at least four fish lineages (Hattori et al. 2012, Eshel et al. 2014, Li et al. 2015, Rondeau

et al. 2016). The emergence of *gsdf* is thought to come from the vertebrate-specific 1R and 2R whole-genome duplications of a TGF-β family genes (Yan et al. 2017); paralogs from these duplications, *gdf9* and *bmp15*, play an integral role in proper ovarian development (Nef et al. 2005, McIntosh et al. 2008). Gene and genome duplication therefore plays a prominent role in all of the studies presented in this dissertation.

Genome assembly Improvements – *Esox lucius* Chapter 3 to Chapter 4

Improvements to the genome assembly can be difficult to communicate effectively. The original genome released in 2014 (Rondeau et al. 2014) was sufficient to provide an annotated transcriptome in the analysis of the fate of gene duplicates as described above. It has also allowed us to identify conserved synteny across the Neoteleostei in a general context. However, this genome assembly was reflective of the technology available at the time in which it was generated, as a *de novo* assembly based on standard Illumina short and medium insert libraries. While a number of similar genomes were released at the time (Amemiya et al. 2013, Berthelot et al. 2014), such preliminary assemblies could not reveal much about the structure of the genome, as they remained far too fragmented. As discussed in Chapter 4, adding additional scaffolds through mapping was not a practical solution, given the limited polymorphism observed in northern pike. Thus, as genome analysis transitioned from a focus on gene content towards genomic structure, major improvements to genomic contiguity was required.

Approaches to resolving contiguity have varied, from supplementing with longer, high-quality, but expensive Sanger reads (Lien et al. 2016) to complete re-assembly with emerging new sequencing technologies (Jansen et al. 2017). The speed at which innovative technologies are being released, combined with the frequent improvements to existing technologies, means that were I to begin again, my choice of technology would be completely different from four years ago (when I began the work reported in Chapter 3) or even 1 year ago (when the last library for Chapter 4 was sequenced). The

hybrid approach we chose in Chapter 4 reflects the yearly improvements to technology and recommended approaches.

The step-by-step incorporation of new sequencing libraries to the northern pike genome successfully generated a significant improvement to the assembly. The increase in scaffold N50 from 700 Mbp to 7,945 Mbp allowed us to more reliably order and orient scaffolds onto chromosomes by utilizing more markers per scaffold. It allowed us to physically assign a greater proportion of the genome to chromosomes, from 46 to 88 percent, without the challenging task of improving the genetic map in a low polymorphism species. It provided the opportunity to assign a majority of genes directly to chromosomes, and it allowed a measure of conserved linkage that can further indirectly link genes on unlinked scaffolds to the likely position. In short, this new genome allowed for analysis of the structural organization of the genome that had not been possible with the original preliminary assembly and provided a high-quality reference to use in ongoing salmonid and esocid projects for years to come.

While many genome assemblies are compared on their N50 statistics, a measure of the scaffold or contig size from largest to smallest at which 50% or more of the bases are represented, it is imperfect. While such a standardized and simple to understand term has allowed relative understanding of contiguity of an assembly, as we approach full chromosome contiguity, maximal N50 values will never exceed the length of mid-sized chromosomes. In humans, for example, N50 could never exceed the length of chromosome 8, as the total bases in chromosomes 1-8 exceed the bases in the remaining chromosomes (GCR38; female). A slightly unusual trend, therefore, is the emergence of N50 based on the sequence anchored to chromosomes (e.g. (Park et al. 2015, Reichwald et al. 2015)). Under such a definition, we could determine an equivalent N50 of 32.94 Mbp in our current northern pike assembly. As we transition towards the ability to generate relatively full-length representations of chromosomes, it will be likely that representations and comparisons between N50 statistics will drift away from describing scaffolds or chromosomes and towards contigs and contiguity

directly. No matter how one quantitatively measures genome quality, the version presented in Chapter 4 represents a near-complete, high-quality reference representing chromosomal structural organization for the northern pike.

Future Directions

Future projects stemming from this work generally head in two primary directions. The first is the continuation of genome sequencing, particularly in expanding the number of salmonid species for which chromosome-level genome assemblies are constructed. The second is in the annotation of genomes. This latter direction includes documenting variation across the species through genome re-sequencing, developing tools necessary to analyze variation, and extending annotation of the salmonid genomes beyond simply identifying genes and their expression patterns.

Towards our first direction, we recently released assembly 1.0 for the coho salmon (*Oncorhynchus kisutch*), the first chromosome-level release for a Pacific salmonid as part of the larger Genome Canada project EPIC4 – Enhanced Production in Coho: Culture, Community, Catch. Further, we have collaborated on the version 2.0 release of the rainbow trout genome and have ongoing genome assembly projects in collaboration with Dr. Willie Davidson at Simon Fraser University on Arctic charr (*Salvelinus alpinus*) and in collaboration with Drs. Robert Devlin and Kris Christensen at Fisheries and Oceans Canada on Chinook salmon (*Oncorhynchus tshawytscha*) and sockeye salmon (*Oncorhynchus nerka*). Plans also include preliminary genome characterization of pink salmon (*Oncorhynchus gorbuscha*), chum salmon (*Oncorhynchus keta*), and cutthroat trout (*Oncorhynchus clarkii*). Worldwide, additional salmonid projects covering the Thymallinae (graylings) and Coregoninae (whitefish), as well as additional European and freshwater Salmoninae are in various stages of completion (Macqueen et al. 2017). As the salmonid genome assemblies become available, reconstruction of ancestral salmonid common ancestors will allow for an analysis of genome rearrangements and inversions separating individual species,

expanding genome structural analysis from the level of phylogenetic order to the genus. This will allow for a full analysis of the effects of individual rearrangements that may have contributed to speciation or sub-species isolation. While such reconstructions have been attempted in the past using a handful of teleost species (e.g. (Jaillon et al. 2004)), this represents an excellent opportunity to observe the various paths to re-diploidization after genome duplication utilized in different salmonid lineages. Whole genome duplication has played a fundamental role in the evolution of animals and plants, and the recency of such an event in salmonids provides a unique opportunity to examine the mechanism of this fundamental evolutionary event.

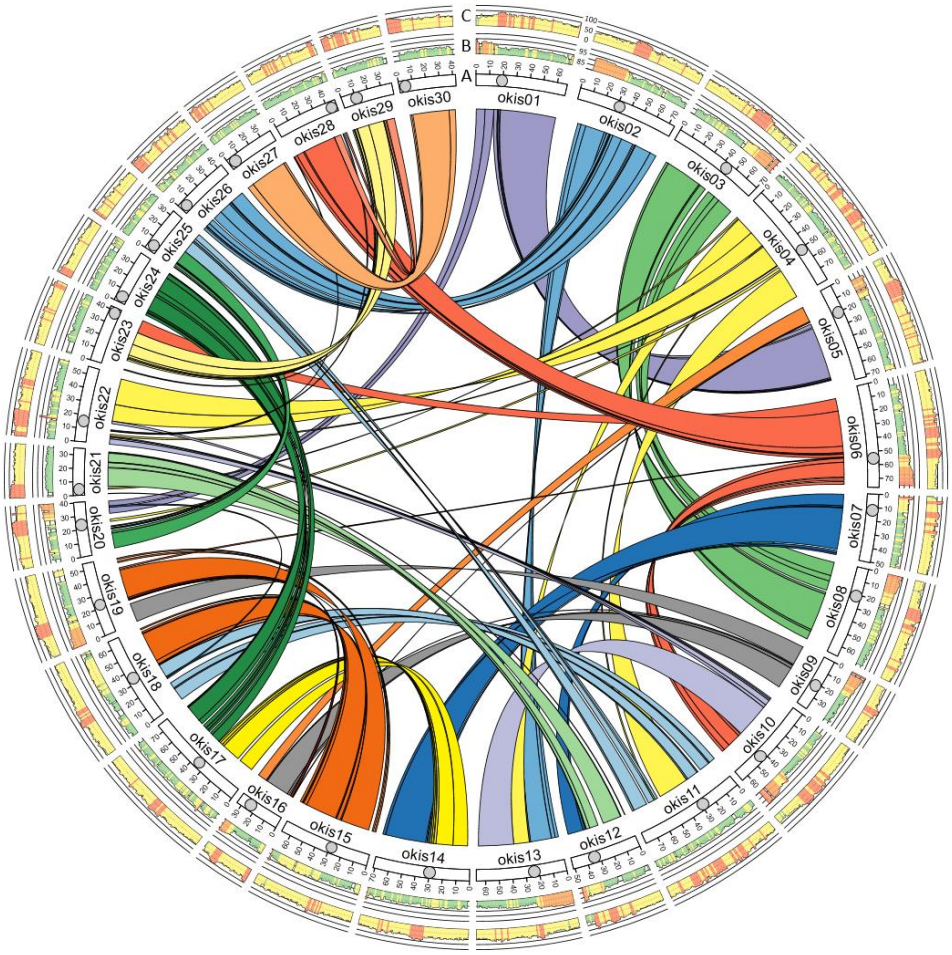


Figure 5.3 Circos plot of homeologous regions in coho salmon

Homeologous regions in the coho salmon genome subdivided into collinear blocks along the 30 chromosomes. The inner track shows genomic similarity (in 1 Mb intervals) between duplicated regions (red = high, yellow = medium, green = low sequence similarity). The outer track shows frequency of repetitive elements in the genome.

As introduced in Chapter 4, the technology of both sequencing and assembly is evolving rapidly. Particularly exciting to me, as it relates to salmonids, is the development of long-read technologies, allowing for long, linear contiguity of the

sequences without large-scale reassembly. Pacific Biosciences RSII or Sequel sequencing has been widely adopted by many of the core sequencing centres. We sequenced a relatively low-coverage of PacBio reads in this work, primarily for gap-filling in a more traditional Illumina context. The advances in error-correction and assembly, however, particularly in the development of dedicated assemblers such as Canu (Koren et al. 2017), have resulted in highly contiguous, long-read only assemblies in a number of species (Jiao et al. 2017, Badouin et al. 2017). The relative ability of long-read assemblers to identify such significant structural variations should allow for the characterization of regions of extended repetitive content or regions of residual tetraploidy. Nanopore-only assemblies have also been released with similar contiguity to those generated through PacBio-only protocols, but with two additional advantages. The first is the capital cost of the smallest device, termed the MinION, which is $1,000 USD. This device allows in-lab sequencing rather than through a core-facility, thereby allowing more flexibility and control over the output. The second advantage is that there is no advertised theoretical upper size limit, and indeed individual read lengths have approached 1Mbp in publicly released data with an N50 read-length in excess of 100 kbp (Jain et al. 2017). While the initial results I have generated in coho salmon have produced reads more than 125kbp, the read-length distribution has been unfortunately low (read-length N50 of 5 kbp). As the technology advances and long-read protocols become more consistent, such long-reads could allow for efficient assembly through regions of subtle structural differences or the high-density of long repeats.

The second area of interest is in functional annotation of the salmonid genomes (Macqueen et al. 2017). Phylogenetic conservation of genomic structure between species as primarily discussed in this dissertation is just one modification to the genomic landscape. Population-level variation in the structural organization (e.g. (Tuttle et al. 2016)), loss-of-function mutations (e.g. (MacArthur et al. 2012)),or genetic polymorphism in general (1000 Genomes Project Consortium et al. 2015) can produce a significant degree of variation between individuals within a species. With decreasing

sequencing costs, it becomes increasingly effective to look at full genomic variation across individuals rather than microsatellite or other reduced-representation methods. Indeed, I am currently pursuing such full genome re-sequencing approaches, and I am generating such data from coho salmon, rainbow trout, Arctic Charr, and northern pike.

Our earlier work in the regulation of expression is also being extended to the functional annotation of the elements controlling expression. While regulatory neofunctionalization was proposed as a major fate of retained duplicates, the mechanism of control remains a mystery. In other systems, it has been proposed that epigenetics may be part of the mechanism (Conant et al. 2014), and as such some of our current efforts have shifted from gene expression to the epigenetic control associated with expression. This includes looking at transcription factors, histone modifications, and methylation. In the same way that a triplet approach was utilized to analyze the regulation of expression in Atlantic salmon relative to the non-duplicated outgroup (Lien et al. 2016), I am looking at conservation of epigenetic controls between the northern pike and coho salmon and how well they correlate to control and expression in maintained gene duplicates.

Finally, as demonstrated with ongoing work in Atlantic salmon testis (von Schalburg et al. 2017), current Illumina short-read technologies that require reconstruction through assembly make an evaluation of true alternative splicing enormously difficult. Exploring alternative splicing using full-length RNA sequencing platforms (Tseng and Underwood 2013, Byrne et al. 2017) should further extend the study of epigenetic control from the genome to pre-mRNA splicing (Luco et al. 2010, Curado et al. 2015).

Conclusions

This work details the degree to which conserved synteny and linkage are maintained among the teleosts. I demonstrated the ability to identify specific genes of interest in non-model organisms by harnessing conserved linkage, thereby detailing a

practical path towards answering fundamental genetic questions in species with limited genomic resources. I defined the degree to which conserved synteny is maintained across some teleost lineages and demonstrated that while intra-chromosomal rearrangement is high, inter-chromosomal rearrangements are extremely low. This limited inter-chromosomal rearrangement allows a newly sequenced pike genome to serve as an outgroup to a whole-genome duplication event. Finally, by improving the genome to chromosome-anchored reference quality, I demonstrated that while major structural changes occur following re-diploidization, long stretches of conserved linkage allow reference-guided improvements across phylogenetic Orders by anchoring additional sequences for a genome with limited options of genetic mapping improvements. I demonstrated that by guiding error-correction, ordering, and orientation of assembly collapsed scaffolds within regions of delayed re-diploidization, the pre-duplicated genome suggests a practical path towards improved chromosomal anchoring in many salmonid projects until full-length chromosomal sequencing becomes available.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. 2015. A global reference for human genetic variation. Nature **526**(7571): 68–74. doi:10.1038/nature15393.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.-H.C., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Miklos, Abril, J.F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P. V, Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., Pablos, B. de, Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N. V, Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D.C.,

Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.-Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.-F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., and Venter, J.C. 2000. The genome sequence of *Drosophila melanogaster*. Science **287**(5461): 2185–2195.

Ahn, S., and Tanksley, S.D. 1993. Comparative linkage maps of the rice and maize genomes. Proc. Natl. Acad. Sci. **90**(September): 7980–7984.

Alderdice, D.F., Jensen, J.O.T., and Velsen, F.P.J. 1988. Preliminary trials on incubation of sablefish eggs (*Anoplopoma fimbria*). Aquaculture **69**(3–4): 271–290. doi:10.1016/0044-8486(88)90335-3.

Allen, M.J., and Smith, G.B. 1988. Atlas and zoogeography of common fishes in the Bering Sea and northeastern Pacific. NOAA Tech. Rep. **66**(April): 151.

Allendorf, F.W., and Thorgaard, G.H. 1984. Tetraploidy and the Evolution of Salmonid Fishes. *In* Evolutionary Genetics of Fishes. *Edited by* B. Turner. Springer US. pp. 1–53. doi:10.1007/978-1-4684-4652-4_1.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17): 3389–3402. doi:10.1093/nar/25.17.3389.

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J.J., Robinson, M., Dorrington, R.A., Gerdol, M., Aken, B., Biscotti, M.A., Barucca, M., Baurain, D., Berlin, A.M., Blatch, G.L., Buonocore, F., Burmester, T., Campbell, M.S., Canapa, A., Cannon, J.P., Christoffels, A., De Moro, G., Edkins, A.L., Fan, L., Fausto, A.M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J. V.,

Haerty, W., Hahn, M.E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S.I., Kuraku, S., Lara, M., Levin, J.Z., Litman, G.W., Mauceli, E., Miyake, T., Mueller, M.G., Nelson, D.R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C.P., Prohaska, S.J., Przybylski, D., Saha, N.R., Ravi, V., Ribeiro, F.J., Sauka-Spengler, T., Scapigliati, G., Searle, S.M.J., Sharpe, T., Simakov, O., Stadler, P.F., Stegeman, J.J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J.-N., Tabin, C.J., Shubin, N., Schartl, M., Jaffe, D.B., Postlethwait, J.H., Venkatesh, B., Di Palma, F., Lander, E.S., Meyer, A., and Lindblad-Toh, K. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature **496**(7445): 311–316. doi:10.1038/nature12027.

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q., and Postlethwait, J.H. 2011. Maps for the masses: Genome evolution and meiotic maps by massively parallel DNA sequencing. Genetics **188**: 799–808. doi:10.1534/genetics.111.127324.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D.S., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science **297**(5585): 1301–1310. doi:10.1126/science.1072104.

Assis, R., and Bachtrog, D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. Proc. Natl. Acad. Sci. **110**(43): 17409–17414. doi:10.1073/pnas.1313759110.

Aury, J.-M., Jaillon, O., Laudet, V., and Robinson-Rechavik, M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol. **23**(9): 1808–1816. doi:10.1093/molbev/msl049.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière,

C., Owens, G.L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N.C., Bowers, J.E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., Boniface, M.-C., Brunel, D., Catrice, O., Chaidir, N., Claudel, C., Donnadieu, C., Faraut, T., Fievet, G., Helmstetter, N., King, M., Knapp, S.J., Lai, Z., Le Paslier, M.-C., Lippi, Y., Lorenzon, L., Mandel, J.R., Marage, G., Marchand, G., Marquand, E., Bret-Mestries, E., Morien, E., Nambeesan, S., Nguyen, T., Pegot-Espagnet, P., Pouilly, N., Raftis, F., Sallet, E., Schiex, T., Thomas, J., Vandecasteele, C., Varès, D., Vear, F., Vautrin, S., Crespi, M., Mangin, B., Burke, J.M., Salse, J., Muños, S., Vincourt, P., Rieseberg, L.H., and Langlade, N.B. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature **546**(7656): 148–152.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One **3**(10): e3376. doi:10.1371/journal.pone.0003376.

Barbazuk, W.B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J.A., McPherson, J.D., and Johnson, S.L. 2000. The synthenic relationship of the zebrafish and human genomes. Genome Res. **10**(9): 1351–1358. doi:10.1101/gr.144700.

Bargelloni, L., Franch, R., Patarnello, T., Tsalavouta, M., Sarropoulou, E., Magoulas, A., Kotoulas, G., Chatziplis, D., Georgoudis, A., Louro, B., and Power, D.M. 2007. A genetic linkage map of the hermaphrodite teleost fish *Sparus aurata* L. Aquaculture **272**: S243–S243. doi:10.1016/j.aquaculture.2007.07.029.

Beamish, R.J., Merrilees, M.J., and Crossman, E.J. 1971. Karyotypes and DNA values for members of the suborder Esocoidei (Osteichthyes: Salmoniformes). Chromosoma **34**(4): 436–447. doi:10.1007/BF00326315.

Begun, D.J., Lindfors, H.A., Kern, A.D., and Jones, C.D. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. Genetics **176**(2): 1131–1137. doi:10.1534/genetics.106.069245.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic

Acids Res. **27**(2): 573–580. doi:10.1093/nar/27.2.573.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short,

S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., and Smith, A.J. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**(7218): 53–59. doi:10.1038/nature07517.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M., Quillet, E., Guyomard, R., Galiana, D., Bobe, J., Volff, J.-N., Genêt, C., Wincker, P., Jaillon, O., Crollius, H.R., and Guiguen, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. **5**: 3657. doi:10.1038/ncomms4657.

Betancur-R, R., Broughton, R.E., Wiley, E.O., Carpenter, K., Wiley, E.O., Carpenter, K., López, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton II, J.C., Zhang, F., Campbell, M.A., Ballesteros, J.A., Roa-Varon, A., Willis, S., Borden, W.C., Rowley, T., Reneau, P.C., Hough, D.J., Lu, G., Grande, T., Arratia, G., and Ortí, G. 2013. The tree of life and a new classification of bony fishes. PLoS Curr. Tree Life **732988**: Edition 1. doi:10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.Abstract.

Bianco, 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 [English Translation]. *In* Researches on Wildlife Conservation, Volume 2. *Edited by* G. de Filippo. IGF Publishing. pp. 1–13.

Blacket, M.J., Robin, C., Good, R.T., Lee, S.F., and Miller, A.D. 2012. Universal primers for fluorescent labelling of PCR fragments-an efficient and cost-effective approach to genotyping by fluorescence. Mol. Ecol. Resour. **12**(3): 456–463. doi:10.1111/j.1755-0998.2011.03104.x.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. **7**(5): R43. doi:10.1186/gb-2006-7-5-r43.

de Boer, J.G., Yazawa, R., Davidson, W.S., and Koop, B.F. 2007. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics **8**(1): 422. doi:10.1186/1471-2164-8-422.

Boetzer, M., and Pirovano, W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics **15**(1): 211. doi:10.1186/1471-2105-15-211.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**(6930): 433–438. doi:10.1038/nature01521.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A.M., Campbell, M.S., Barrell, D., Martin, K.J., Mulley, J.F., Ravi, V., Lee, A.P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F.E.G., Sun, Y., Hertel, J., Beam, M.J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J.H., Litman, G.W., Litman, R.T., Mikami, M., Ota, T., Saha, N.R., Williams, L., Stadler, P.F., Wang, H., Taylor, J.S., Fontenot, Q., Ferrara, A., Searle, S.M.J., Aken, B., Yandell, M., Schneider, I., Yoder, J.A., Volff, J.-N., Meyer, A., Amemiya, C.T., Venkatesh, B., Holland, P.W.H., Guiguen, Y., Bobe, J., Shubin, N.H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J.H. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. **48**(4): 427–437. doi:10.1038/ng.3526.

Brown, T.A. 2002. Chapter 5 Mapping Genomes. *In* Genomes, 2nd Editio. Oxford: Wiley-Liss. pp. 1–56.

Budd, A., Banh, Q., Domingos, J., and Jerry, D. 2015. Sex Control in Fish: Approaches, Challenges and Opportunities for Aquaculture. J. Mar. Sci. Eng. **3**(2): 329–355.

doi:10.3390/jmse3020329.

Butler, J.M., McCord, B.R., Jung, J.M., Wilson, M.R., Budowle, B., and Allen, R.O. 1994. Quantitation of polymerase chain reaction products by capillary electrophoresis using laser fluorescence. J. Chromatograpy B Biomed. Appl. **658**(2): 271–280. doi:10.1016/0378-4347(94)00238-X.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., and Vollmers, C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. **8**: 16027. doi:10.1038/ncomms16027.

Campbell, M.A., López, J.A., Sado, T., and Miya, M. 2013. Pike and salmon as sister taxa: Detailed intraclade resolution and divergence time estimation of Esociformes+Salmoniformes based on whole mitochondrial genome sequences. Gene **530**(1): 57–65. doi:10.1016/j.gene.2013.07.068.

Canadian Aquaculture Industry Alliance. 2009. Making a splash: Industry forecasts new aquaculture species will generate $880 million by 2020. Available from http://www.aquaculture.ca/files/archives.php.

Canapa, A., Barucca, M., Biscotti, M.A., Forconi, M., and Olmo, E. 2015. Transposons, genome size, and evolutionary insights in animals. Cytogenet. Genome Res. **147**(4): 217–239. doi:10.1159/000444429.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., and Yandell, M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**(1): 188–196. doi:10.1101/gr.6743907.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. 2011. Stacks : Building and genotyping loci de novo from short-read sequences. G3 Genes|Genomes|Genetics **1**(3): 171–182. doi:10.1534/g3.111.000240.

Chaisson, M.J.P., Wilson, R.K., and Eichler, E.E. 2015. Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. **16**(11): 627–640.

doi:10.1038/nrg3933.

Chakravarti, A. 1991. A graphical representation of genetic and physical maps: The Marey map. Genomics **11**(1): 219–222. doi:10.1016/0888-7543(91)90123-V.

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

Christoffels, A., Koh, E.G.L., Chia, J.M., Brenner, S., Aparicio, S., and Venkatesh, B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol. Biol. Evol. **21**(6): 1146–1151. doi:10.1093/molbev/msh114.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M.J.L. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11): 1422–1423. doi:10.1093/bioinformatics/btp163.

Cole, C.G., McCann, O.T., Collins, J.E., Oliver, K., Willey, D., Gribble, S.M., Yang, F., McLaren, K., Rogers, J., Ning, Z., Beare, D.M., and Dunham, I. 2008. Finishing the finished human chromosome 22 sequence. Genome Biol. **9**(5): R78. doi:10.1186/gb-2008-9-5-r78.

Conant, G.C., Birchler, J.A., Pires, J.C., O ’connor, S., and Brutnell, T. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. **19**: 91–98. doi:10.1016/j.pbi.2014.05.008.

Conant, G.C., and Wolfe, K.H. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. **9**(12): 938–950. doi:10.1038/nrg2482.

Consortium Molecular Ecology Resources Primer Development, Aggarwal, R.K., Allainguillaume, J., Barthwal, S., Bertolino, P., Chauhaun, P., Consuegra, S., Croxford, A., Dalton, D.L., Den Belder, E., Diaz-Ferguson, E., Douglas, M.R., Drees,

M., Elderson, J., Esselink, G.D., Fernandez-Manjarres, J.F., Frascaria-Lacoste, N., Gabler-Schwarz, S., De Leaniz, C.G., Ginwal, H.S., Goodisman, M.A.D., Guo, B., Hamilton, M.B., Hayes, P.K., Hong, Y.A.N., Kajita, T., Kalinowski, S.T., Keller, L., Koop, B.F., Kotze, A., Lalremruata, A., Leese, F., Li, C., Liew, W.Y., Martinelli, S., Matthews, E.A., Medlin, L.K., Messmer, A.M., Meyer, E.I., Monteiro, M., Moyer, G.R., Nelson, R.J., Nguyen, T.T.T., Omoto, C., Ono, J., Pavinato, V.A.C., Pearcy, M., Pinheiro, J.B., Power, L.D., Rawat, A., Reusch, T.B.H., Sanderson, D., Sannier, J., Sathe, S., Sheridan, C.K., Smulders, M.J.., Sukganah, A., Takayama, K., Tamura, M., Tateishi, Y., Vanhaecke, D., Vu, N. V., Wickneswari, R., Williams, A.S., Wimp, G.M., Witte, V., and Zucchi, M.I. 2011. Permanent Genetic Resources added to Molecular Ecology Resources Database 1 August 2010 – 30 September 2010. Mol. Ecol. Resour. **11**(1): 219–222. doi:10.1111/j.1755-0998.2010.02944.x.

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

Crossman, E.J., and Ráb, P. 2001. Chromosomal NOR Phenotype and C-Banded Karyotype of Olympic Mudminnow, *Novumbra hubbsi* (Euteleostei: Umbridae). Copeia **3**(3): 860–865. doi:10.1643/0045-8511(2001)001[0860:CNPACB]2.0.CO;2.

Curado, J., Iannone, C., Tilgner, H., Valcárcel, J., and Guigó, R. 2015. Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. Genome Biol. **16**(1): 236. doi:10.1186/s13059-015-0797-8.

Danzmann, R.G., Cairney, M., Davidson, W.S., Ferguson, M.M., Gharbi, K., Guyomard, R., Holm, L.-E., Leder, E., Okamoto, N., Ozaki, A., Rexroad III, C.E., Sakamoto, T., Taggart, J.B., and Woram, R.A. 2005. A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). Genome **48**(6): 1037–1051. doi:10.1139/g05-067.

Darwin, C. 1859. The Origin of Species by means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. *In* Nature. John Murray,

Albemarle Street, London.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. **12**(7): 499–510. doi:10.1038/nrg3012.

Davidson, W.S., Koop, B.F., Jones, S.J., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., and Omholt, S.W. 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol. **11**(9): 403. doi:10.1186/GB-2010-11-9-403.

Dehal, P., and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. **3**(10): e314. doi:10.1371/journal.pbio.0030314.

Dekker, J. 2002. Capturing Chromosome Conformation. Science **295**(5558): 1306–1311. doi:10.1126/science.1067799.

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. C. R. Biol. **337**(9): 521–534. doi:10.1016/j.crvi.2014.07.002.

Dittmar, K., and Liberles, D.A.. 2010. Evolution after Gene Duplication. Wiley. Available from http://books.google.ca/books?id=zociCymQCacC.

Donaldson, E.M. 1996. Manipulation of reproduction in farmed fish. Anim. Reprod. Sci. **42**(1–4): 381–392. doi:10.1016/0378-4320(96)01555-2.

Echave, K.B., Hanselman, D.H., Adkison, M.D., and Sigler, M.F. 2012. Interdecadal Change in Growth of Sablefish (*Anoplopoma fimbria*) in the Northeast Pacific Ocean. Fish. Bull. **110**(3): 361–374.

Edgar, R.C. 2004a. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R.C. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**: 113. doi:10.1186/1471-2105-5-113.

Edgar, R.C. 2010. Search and clutering orders of magnitude faster than BLAST. Bioinformatics **26**(19): 2460–2461. doi:10.1093/bioinformatics/btq461.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. 2009. Real-time DNA sequencing from single polymerase molecules. Science **323**(5910): 133–138. doi:10.1126/science.1162986.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J. a., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotyping by sequencing (GBS) approach for high diversity species. PLoS One **6**(5): e19379. doi:10.1371/journal.pone.0019379.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. 2012. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7**(11): e47768. doi:10.1371/journal.pone.0047768.

Eshel, O., Shirak, A., Dor, L., Band, M., Zak, T., Markovich-Gordon, M., Chalifa-Caspi, V., Feldmesser, E., Weller, J.I., Seroussi, E., Hulata, G., and Ron, M. 2014. Identification of male-specific amh duplication, sexually differentially expressed genes and microRNAs at early embryonic development of Nile tilapia (*Oreochromis niloticus*). BMC Genomics **15**(1): 774. doi:10.1186/1471-2164-15-774.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8**(3): 186–194. doi:10.1101/gr.8.3.186.

Fairgrieve, M.R., Shibata, Y., Smith, E.K., Hayman, E.S., and Luckenbach, J.A. 2016. Molecular characterization of the gonadal kisspeptin system: Cloning, tissue distribution, gene expression analysis and localization in sablefish (*Anoplopoma fimbria*). Gen. Comp. Endocrinol. **225**(Supplement C): 212–223. doi:10.1016/j.ygcen.2015.07.015.

FAO. 1988. Synopsis of biological Data on the Northern Pike. Food and Agriculture Organization of the United Nations. Available from https://books.google.ca/books?id=l9ZwEPc8y5cC.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature **260**(5551): 500–507. doi:10.1038/260500a0.

Fierst, J.L. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front. Genet. **6**: 220. Frontiers Media S.A. doi:10.3389/fgene.2015.00220.

Fisheries and Oceans Canada. 2013. Preliminary Summary Commercial Statistics 2012: Cumulative totals. Available from http://www.pac.dfo-mpo.gc.ca/stats/comm/summ-somm/index-eng.htm.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocyne, J.D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K. V., Fraser, C.M., Hamilton O. Smith, and Venter, J.C. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**(5223): 496–512. doi:10.1126/science.7542800.

Fletcher, G.L., Hew, C.L., and Davies, P.L. 2001. Antifreeze proteins of teleost fishes. Annu. Rev. Physiol. **63**(1): 359–390. doi:10.1146/annurev.physiol.63.1.359.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. 2011. Considering transposable element diversification in de novo annotation approaches. PLoS One **6**(1): e16526. doi:10.1371/journal.pone.0016526.

Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**(4): 1531–1545.

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**(2): 472–479. doi:10.1111/jfb.12712.

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**(10): 2165–2175. doi:10.1111/mec.13594.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D., and Kent, W.J. 2011. The UCSC genome browser database: Update 2011. Nucleic Acids Res. **39**(SUPPL. 1): D876–D882. doi:10.1093/nar/gkq963.

Gao, B., Shen, D., Xue, S., Chen, C., Cui, H., and Song, C. 2016a. The contribution of transposable elements to size variations between four teleost genomes. Mob. DNA **7**(1): 4. doi:10.1186/s13100-016-0059-7.

Gao, S., Bertrand, D., Chia, B.K.H., and Nagarajan, N. 2016b. OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome Biol. **17**(1): 102. doi:10.1186/s13059-016-0951-y.

Gates, M.A., Kim, L., Egan, E.S., Cardozo, T., Sirotkin, H.I., Dougan, S.T., Lashkari, D., Abagyan, R., Schier, A.F., and Talbot, W.S. 1999. A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. Genome Res. **9**(4): 334–347. doi:10.1101/gr.9.4.334.

Gautier, A., Le Gac, F., and Lareyre, J.J. 2011a. The *gsdf* gene locus harbors evolutionary conserved and clustered genes preferentially expressed in fish previtellogenic oocytes. Gene **472**(1–2): 7–17. doi:10.1016/j.gene.2010.10.014.

Gautier, A., Sohm, F., Joly, J.-S., Le Gac, F., and Lareyre, J.-J. 2011b. The proximal promoter region of the zebrafish *gsdf* gene is sufficient to mimic the spatio-temporal expression pattern of the endogenous gene in sertoli and granulosa cells. Biol. Reprod. **85**(6): 1240–1251. doi:10.1095/biolreprod.111.091892.

Gilbey, J., Verspoor, E., McLay, A., and Houlihan, D. 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim. Genet. **35**(2): 98–105. doi:10.1111/j.1365-2052.2004.01091.x.

Glover, N.M., Redestig, H., and Dessimoz, C. 2016. Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci. **21**(7): 609–621. doi:10.1016/j.tplants.2016.02.005.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., and Jaffe, D.B. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. **108**(4): 1513–1518. doi:10.1073/pnas.1017351108.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 Genes. Science **274**(5287): 546,563-567. doi:10.1126/science.274.5287.546.

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., and Houston, R.D. 2014. Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics **15**(1): 166. doi:10.1186/1471-2164-15-166.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., and McCombie, W.R. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. **25**(11): 1750–1756. doi:10.1101/gr.191395.115.

Gordon, D., Huddleston, J., Chaisson, M.J.P., Hill, C.M., Kronenberg, Z.N., Munson, K.M.,

Malig, M., Raja, A., Fiddes, I., Hillier, L.W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R.K., Haussler, D., Chin, C.-S., and Eichler, E.E. 2016. Long-read sequence assembly of the gorilla genome. Science **352**(6281): aae0344. doi:10.1126/science.aae0344.

Gotoh, R.O., Tamate, S., Yokoyama, J., Tamate, H.B., and Hanzawa, N. 2013. Characterization of comparative genome-derived simple sequence repeats for acanthopterygian fishes. Mol. Ecol. Resour. **13**(3): 461–472. doi:10.1111/1755-0998.12070.

Gout, J.F., and Lynch, M. 2015. Maintenance and loss of duplicated genes by dosage subfunctionalization. Mol. Biol. Evol. **32**(8): 2141–2148. doi:10.1093/molbev/msv095.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**(7): 644–652. doi:10.1038/nbt.1883.

Gregory, T.R. 2005. Animal Genome Size Database. doi:http://www.genomesize.com.

Grimholt, U., Hauge, H., Hauge, A.G., Leong, J., and Koop, B.F. 2015. Chemokine receptors in Atlantic salmon. Dev. Comp. Immunol. **49**(1): 79–95. doi:10.1016/j.dci.2014.11.009.

Guzmán, J.M., Luckenbach, J.A., Goetz, F.W., Fairgrieve, W.T., Middleton, M.A., and Swanson, P. 2015a. Reproductive Dysfunction in Cultured Sablefish (*Anoplopoma fimbria*). Bull. Fish. Res. Agency **40**(40): 111–119.

Guzmán, J.M., Luckenbach, J.A., Middleton, M.A., Massee, K.C., Jensen, C., Goetz, F.W., Jasonowicz, A.J., and Swanson, P. 2017. Reproductive life history of sablefish (*Anoplopoma fimbria*) from the U.S. Washington coast. PLoS One **12**(9): e0184413. doi:10.1371/journal.pone.0184413.

Guzmán, J.M., Luckenbach, J.A., da Silva, D.A.M., Ylitalo, G.M., and Swanson, P. 2015b. Development of approaches to induce puberty in cultured female sablefish (*Anoplopoma fimbria*). Gen. Comp. Endocrinol. **221**: 101–113. doi:10.1016/j.ygcen.2015.02.024.

Guzmán, J.M., Luckenbach, J.A., and Swanson, P. 2013. Molecular characterization and quantification of sablefish (*Anoplopoma fimbria*) gonadotropins and their receptors: Reproductive dysfunction in female captive broodstock. Gen. Comp. Endocrinol. **193**: 37–47. doi:10.1016/j.ygcen.2013.07.007.

Hardie, D.C., and Hebert, P.D. 2004. Genome-size evolution in fishes. Can. J. Fish. Aquat. Sci. **61**(9): 1636–1646. doi:10.1139/f04-106.

Hardie, D.C., and Hebert, P.D.N. 2003. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome **46**(4): 683–706. doi:10.1139/g03-040.

Harris, R.S. 2007. Improved pairwise alignment of genomic DNA. Pennsylvania State University.

Hattori, R.S., Murai, Y., Oura, M., Masuda, S., Majhi, S.K., Sakamoto, T., Fernandino, J.I., Somoza, G.M., Yokota, M., and Strussmann, C.A. 2012. A Y-linked anti-Mullerian hormone duplication takes over a critical role in sex determination. Proc. Natl. Acad. Sci. **109**(8): 2955–2959. doi:10.1073/pnas.1018392109.

He, L., Wang, Q., Jin, X., Wang, Y., Chen, L., Liu, L., and Wang, Y. 2012. Transcriptome profiling of testis during sexual maturation stages in *Eriocheir sinensis* using Illumina sequencing. PLoS One **7**(3): e33735. doi:10.1371/journal.pone.0033735.

Heule, C., Salzburger, W., and Böhne, A. 2014. Genetics of sexual development: An evolutionary playground for fish. Genetics **196**(3): 579–591. doi:10.1534/genetics.114.161158.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., Dodgson, J.B., Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., Fulton, R.S., Graves, T.A., Kremitzki, C., Layman, D., Magrini, V., McPherson, J.D., Miner, T.L., Minx, P., Nash, W.E.,

Nhan, M.N., Nelson, J.O., Oddy, L.G., Pohl, C.S., Randall-Maher, J., Smith, S.M., Wallis, J.W., Yang, S.-P., Romanov, M.N., Rondelli, C.M., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H.G., Robertson, L., Masabanda, J.S., Griffin, D.K., Vignal, A., Fillon, V., Jacobbson, L., Kerje, S., Andersson, L., Crooijmans, R.P.M., Aerts, J., van der Poel, J.J., Ellegren, H., Caldwell, R.B., Hubbard, S.J., Grafham, D. V., Kierzek, A.M., McLaren, S.R., Overton, I.M., Arakawa, H., Beattie, K.J., Bezzubov, Y., Boardman, P.E., Bonfield, J.K., Croning, M.D.R., Davies, R.M., Francis, M.D., Humphray, S.J., Scott, C.E., Taylor, R.G., Tickle, C., Brown, W.R.A., Rogers, J., Buerstedde, J.-M., Wilson, S.A., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M.M., Inoko, H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G.K.-S., Wang, J., Liu, B., Wang, J., Yu, J., Yang, H., Nefedov, M., Koriabine, M., deJong, P.J., Goodstadt, L., Webber, C., Dickens, N.J., Letunic, I., Suyama, M., Torrents, D., von Mering, C., Zdobnov, E.M., Makova, K., Nekrutenko, A., Elnitski, L., Eswara, P., King, D.C., Yang, S., Tyekucheva, S., Radakrishnan, A., Harris, R.S., Chiaromonte, F., Taylor, J., He, J., Rijnkels, M., Griffiths-Jones, S., Ureta-Vidal, A., Hoffman, M.M., Severin, J., Searle, S.M.J., Law, A.S., Speed, D., Waddington, D., Cheng, Z., Tuzun, E., Eichler, E., Bao, Z., Flicek, P., Shteynberg, D.D., Brent, M.R., Bye, J.M., Huckle, E.J., Chatterji, S., Dewey, C., Pachter, L., Kouranov, A., Mourelatos, Z., Hatzigeorgiou, A.G., Paterson, A.H., Ivarie, R., Brandstrom, M., Axelsson, E., Backstrom, N., Berlin, S., Webster, M.T., Pourquie, O., Reymond, A., Ucla, C., Antonarakis, S.E., Long, M., Emerson, J.J., Betrán, E., Dupanloup, I., Kaessmann, H., Hinrichs, A.S., Bejerano, G., Furey, T.S., Harte, R.A., Raney, B., Siepel, A., Kent, W.J., Haussler, D., Eyras, E., Castelo, R., Abril, J.F., Castellano, S., Camara, F., Parra, G., Guigo, R., Bourque, G., Tesler, G., Pevzner, P.A., Smit, A., Fulton, L.A., Mardis, E.R., and Wilson, R.K. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**(7018): 695–716. doi:10.1038/nature03154.

Hinegardner, R., and Rosen, D.E. 1972. Cellular DNA content and the evolution of

teleostean fishes. Am. Nat. **106**(951): 621–644. doi:10.1086/282801.

Hohenlohe, P.A., Amish, S.J., Catchen, J.M., Allendorf, F.W., and Luikart, G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Resour. **11**(SUPPL. 1): 117–122. doi:10.1111/j.1755-0998.2010.02967.x.

Holland, P.M., Abramson, R.D., Watson, R., and Gelfand, D.H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc. Natl. Acad. Sci. **88**(16): 7276–7280. doi:10.1073/pnas.88.16.7276.

Holland, P.W., Garcia-Fernàndez, J., Williams, N.A., and Sidow, A. 1994. Gene duplications and the origins of vertebrate development. Dev. Suppl. **1994**: 125–133.

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12**(1): 491. doi:10.1186/1471-2105-12-491.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.-J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assuncao, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.-H., Eyre, T., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Elliott, D., Threadgold, G., Harden, G., Ware, D., Begum, S., Mortimore, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Lloyd, C., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K.,

Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J.D., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C., Ersan-Ueruen, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlaender, M., Rudolph-Geiger, S., Teucke, M., Lanz, C., Raddatz, G., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Schuster, S.C., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M.J., Enright, A., Geisler, R., Plasterk, R.H.A., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nuesslein-Volhard, C., Hubbard, T.J.P., Crollius, H.R., Rogers, J., and Stemple, D.L. 2014. The zebrafish reference genome sequence and its relationship to the human genome. Nature **505**(7482): 248. doi:10.1038/nature12813.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. **4**(1): 44–57. doi:10.1038/nprot.2008.211.

Imamura, H., and Yabe, M. 2002. Demise of the Scorpaeniformes (Actinopterygii: Percomorpha) : An alternative phylogenetic hypothesis. Bull. Fish. Sci. Hokkaido Univ. **53**(3): 107–128.

Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. Proc. Natl. Acad. Sci. **112**(48): 14918–14923. doi:10.1073/pnas.1507669112.

Ishiguro, N.B., Miya, M., and Nishida, M. 2003. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the “Protacanthopterygii.”

Mol. Phylogenet. Evol. **27**(3): 476–488. doi:10.1016/S1055-7903(02)00418-9.

Jackson, T.R., Ferguson, M.M., Danzmann, R.G., Fishback, A.G., Ihssen, P.E., O'Connell, M., and Crease, T.J. 1998. Identification of two QTL influencing upper temperature tolerance in three rainbow trout (*Oncorhynchus mykiss*) half-sib families. Heredity (Edinb). **80**(October): 143–151. doi:10.1046/j.1365-2540.1998.00289.x.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., and Roest Crollius, H. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431**(7011): 946–957. doi:10.1038/nature03025.

Jain, M., Koren, S., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Miga, K.H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. 2017. Nanopore sequencing and assembly of a human genome with ultra-long reads. BioRxiv: 74. doi:10.1101/128835.

Jansen, H.J., Liem, M., Jong-Raadsen, S.A., Dufour, S., Weltzien, F.-A., Swinkels, W., Koelewijn, A., Palstra, A.P., Pelster, B., Spaink, H.P., Thillart, G.E. van den, Dirks, R.P., and Henkel, C. V. 2017. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. Sci. Rep. **7**(1): 7213. doi:10.1038/s41598-017-07650-6.

Jasonowicz, A.J., Goetz, F.W., Goetz, G.W., and Nichols, K.M. 2017. Love the one you're with: genomic evidence of panmixia in the sablefish (*Anoplopoma fimbria*). Can. J. Fish. Aquat. Sci. **74**(3): 377–387. doi:10.1139/cjfas-2016-0012.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antoniou, E., McCombie, W.R., Presting, G.G., McMullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., and Ware, D. 2017. Improved maize reference genome with single-molecule technologies. Nature: 22971. doi:10.1038/nature22971.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di Palma, F., Lindblad-Toh, K., and Kingsley, D.M. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**(7392): 55–61. doi:10.1038/nature10944.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**(1–4): 462–467. doi:10.1159/000084979.

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. 1996. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. **20**(1): 119–121. doi:10.1016/S0097-8485(96)80013-1.

Kai, W., Kikuchi, K., Fujita, M., Suetake, H., Fujiwara, A., Yoshiura, Y., Ototake, M., Venkatesh, B., Miyaki, K., and Suzuki, Y. 2005. A genetic linkage map for the tiger pufferfish, *Takifugu rubripes*. Genetics **171**(1): 227–238. doi:10.1534/genetics.105.042051.

Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., Masaoka, T., Ozaki, A., Kazeto, Y., Gen, K., Nagao, J., Tanaka, H., Kobayashi, T., and Ototake, M.

2014. A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. BMC Genomics **15**(1): 233. doi:10.1186/1471-2164-15-233.

Kalinowski, S.T. 2002. How many alleles per locus should be used to estimate genetic distances? Heredity (Edinb). **88**(1): 62–65. doi:10.1038/sj.hdy.6800009.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., Tohari, S., Brenner, S., Miyadai, T., Venkatesh, B., Suzuki, Y., and Kikuchi, K. 2012. A trans-species missense SNP in *Amhr2* is associated with sex determination in the tiger Pufferfish, *Takifugu rubripes* (Fugu). PLoS Genet. **8**(7): e1002798. doi:10.1371/journal.pgen.1002798.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., and Kohara, Y. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature **447**(7145): 714–719. doi:10.1038/nature05846.

Katju, V., and Bergthorsson, U. 2013. Copy-number changes in evolution: Rates, fitness effects and adaptive significance. Front. Genet. **4**(DEC): 273. doi:10.3389/fgene.2013.00273.

Kawahara, R., Miya, M., Mabuchi, K., Lavoué, S., Inoue, J.G., Satoh, T.P., Kawaguchi, A., and Nishida, M. 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): A new perspective based on whole mitogenome sequences from 75 higher teleosts. Mol. Phylogenet. Evol. **46**(1): 224–236. doi:10.1016/j.ympev.2007.07.009.

Kawahara, R., Miya, M., Mabuchi, K., Near, T.J., and Nishida, M. 2009. Stickleback

phylogenies resolved: Evidence from mitochondrial genomes and 11 nuclear genes. Mol. Phylogenet. Evol. **50**(2): 401–404. doi:10.1016/j.ympev.2008.10.014.

Kelley, J.M., Field, C.E., Craven, M.B., Bocskai, D., Kim, U.J., Rounsley, S.D., and Adams, M.D. 1999. High throughput direct end sequencing of BAC clones. Nucleic Acids Res. **27**(6): 1539–1546. doi:10.1093/nar/27.6.1539.

Kenny, N.J., Chan, K.W., Nong, W., Qu, Z., Maeso, I., Yip, H.Y., Chan, T.F., Kwan, H.S., Holland, P.W.H., Chu, K.H., and Hui, J.H.L. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity (Edinb). **116**(2): 190–199. doi:10.1038/hdy.2015.89.

Kent, W.J. 2002. BLAT -- the BLAST-like alignment tool. Genome Res. **12**: 656–664. doi:10.1101/gr.229202. Article published online before March 2002.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The Human Genome Browser at UCSC. Genome Res. **12**(6): 996–1006. doi:10.1101/gr.229102.

Kimura, D.K., Shimada, A.M., and Shaw, F.R. 1998. Stock structure and movement of tagged sablefish, *Anoplopoma fimbria*, in offshore northeast pacific waters and the effects of el nino-southern oscillation on migration and growth. Fish. Bull. **96**(3): 462–481.

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., Brahmbhatt, S., Rosner, J., Rexroad, C.E., McGowan, C.R., and Davidson, W.S. 2008. A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. BMC Genomics **9**(1): 545. doi:10.1186/1471-2164-9-545.

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

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. 2017.

Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. **27**(5): 722–736. doi:10.1101/gr.215087.116.

Kucuktas, H., Wang, S., Li, P., He, C., Xu, P., Sha, Z., Liu, H., Jiang, Y., Baoprasertkul, P., Somridhivej, B., Wang, Y., Abernathy, J., Guo, X., Liu, L., Muir, W., and Liu, Z. 2009. Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. Genetics **181**(4): 1649–1660. doi:10.1534/genetics.108.098855.

Kuraku, S., Meyer, A., and Kuratani, S. 2009. Timing of genome duplications relative to the origin of the vertebrates: Did cyclostomes diverge before or after? Mol. Biol. Evol. **26**(1): 47–59. doi:10.1093/molbev/msn222.

Lai, J., Ma, J., Swigoňová, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L., and Messing, J. 2004. Gene loss and movement in the maize genome. Genome Res. **14**(10 A): 1924–1931. doi:10.1101/gr.2701104.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S.,

Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M.J. 2001. Initial sequencing and analysis of the human genome. Nature **409**(6822): 860–921. doi:10.1038/35057062.

Launey, S., Krieg, F., Morin, J., and Laroche, J. 2003. Five new microsatellite markers for Northern pike (*Esox lucius*). Mol. Ecol. Notes **3**(3): 366–368. doi:10.1046/j.1471-8286.2003.00452.x.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. 2013. Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. **9**(8): e1003118. doi:10.1371/journal.pcbi.1003118.

Lee, B.-Y., and Kocher, T.D. 2007. Comparative Genomics and Positional Cloning. *In* Aquaculture Genome Technologies. Blackwell Publishing Ltd. pp. 323–335. doi:10.1002/9780470277560.ch19.

Lee, J.S.F., Cook, M.A., Luckenbach, J.A., Berejikian, B.A., Simchick, C.A., Oden, S.M., and Goetz, F.W. 2017. Investigation of long-term effects of larval rearing temperature on growth, deformities, flesh quality, and phenotypic sex of cultured sablefish (*Anoplopoma fimbria*). Aquaculture **479**: 91–99. doi:10.1016/j.aquaculture.2017.05.010.

Leffler, E.M., Gao, Z., Pfeifer, S., Segurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J.D., Sella, G., Donnelly, P., McVean, G., and Przeworski, M. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science **339**(6127): 1578–1582. doi:10.1126/science.1234070.

Lenormand, T., and Dutheil, J. 2005. Recombination difference between sexes: A role for haploid selection. PLoS Biol. **3**(3): 0396–0403. doi:10.1371/journal.pbio.0030063.

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., Davidson, W.S., and Koop, B.F. 2010. *Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. BMC Genomics **11**(1): 279. doi:10.1186/1471-2164-11-279.

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A., and Begun, D.J. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc. Natl. Acad. Sci. **103**(26): 9935–9939. doi:10.1073/pnas.0509809103.

Li, J., Xia, R., McDowall, R.M., López, J.A., Lei, G., and Fu, C. 2010a. Phylogenetic position of the enigmatic *Lepidogalaxias salamandroides* with comment on the orders of

lower euteleostean fishes. Mol. Phylogenet. Evol. **57**(2): 932–936. doi:10.1016/j.ympev.2010.07.016.

Li, M., Sun, Y., Zhao, J., Shi, H., Zeng, S., Ye, K., Jiang, D., Zhou, L., Sun, L., Tao, W., Nagahama, Y., Kocher, T.D., and Wang, D. 2015. A tandem duplicate of Anti-Müllerian Hormone with a missense SNP on the Y chromosome is essential for male sex determination in nile tilapia, *Oreochromis niloticus*. PLoS Genet. **11**(11): e1005678. doi:10.1371/journal.pgen.1005678.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20**(2): 265–272. doi:10.1101/gr.097261.109.

Lichten, M., and Goldman, A.S. 1995. Meiotic recombination hotspots. Annu. Rev. Genet. **29**(1): 423–444. doi:10.1146/annurev.ge.29.120195.002231.

Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt, S.W., and Kent, M.P. 2011. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics **12**(1): 615. doi:10.1186/1471-2164-12-615.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., and Davidson, W.S. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature **533**(7602): 200–205. doi:10.1038/nature17164.

Lipinski, K.J., Farslow, J.C., Fitzpatrick, K.A., Lynch, M., Katju, V., and Bergthorsson, U.

2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. Curr. Biol. **21**(4): 306–310. doi:10.1016/j.cub.2011.01.026.

Litt, M., and Luty, J.A. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44**(3): 397–401.

Liu, S., Rexroad, C.E., Couch, C.R., Cordes, J.F., Reece, K.S., and Sullivan, C. V. 2012. A microsatellite linkage map of striped bass (*Morone saxatilis*) reveals conserved synteny with the three-spined stickleback (*Gasterosteus aculeatus*). Mar. Biotechnol. **14**(2): 237–244. doi:10.1007/s10126-011-9407-2.

Liu, Z., and Cordes, J. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquacultrure **238**(1–4): 1–37. doi:10.1016/j.aquaculture.2004.05.027.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. 2012. RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. **40**(W1): W622–W627. doi:10.1093/nar/gks540.

Lu, J., Peatman, E., Tang, H., Lewis, J., and Liu, Z. 2012. Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. BMC Genomics **13**(1): 246. doi:10.1186/1471-2164-13-246.

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**(12): e25218. doi:10.1371/journal.pone.0025218.

Luckenbach, J.A., and Fairgrieve, W.T. 2016. Gonadal sex differentiation and effects of dietary methyltestosterone treatment in sablefish (*Anoplopoma fimbria*). Fish Physiol. Biochem. **42**(1): 233–248. doi:10.1007/s10695-015-0132-z.

Luckenbach, J.A., Fairgrieve, W.T., and Hayman, E.S. 2017. Establishment of monosex female production of sablefish (*Anoplopoma fimbria*) through direct and indirect

sex control. Aquaculture **479**: 285–296. doi:10.1016/j.aquaculture.2017.05.037.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. 2010. Regulation of Alternative Splicing by Histone Modifications. Science **327**(5968): 996–1000. doi:10.1126/science.1184208.

Lundin, L.G. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16**(1): 1–19. doi:10.1006/geno.1993.1133.

Lynch, M. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**(5494): 1151–1155. doi:10.1126/science.290.5494.1151.

Lynch, M., and Force, A. 2000. The probability of duplicate-gene preservation by subfunctionalization. Genetics **154**(1): 459–473. doi:10.1371/journal.pgen.0040029.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.-M., Hunt, T., Barnes, I.H.A., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B., and Tyler-Smith, C. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science **335**(6070): 823–828. doi:10.1126/science.1215040.

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 Biol. Sci. **281**(1778): 20132881–20132881. doi:10.1098/rspb.2013.2881.

Macqueen, D.J., Primmer, C.R., Houston, R.D., Nowak, B.F., Bernatchez, L., Bergseth, S., Davidson, W.S., Gallardo-Escárate, C., Goldammer, T., Guiguen, Y., Iturra, P., Kijas,

J.W., Koop, B.F., Lien, S., Maass, A., Martin, S.A.M., McGinnity, P., Montecino, M., Naish, K.A., Nichols, K.M., Ólafsson, K., Omholt, S.W., Palti, Y., Plastow, G.S., Rexroad, C.E., Rise, M.L., Ritchie, R.J., Sandve, S.R., Schulte, P.M., Tello, A., Vidal, R., Vik, J.O., Wargelius, A., and Yáñez, J.M. 2017. Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. BMC Genomics **18**(1): 484. doi:10.1186/s12864-017-3862-8.

Mank, J.E., and Avise, J.C. 2006. Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes. Genetica **127**(1–3): 321–327. doi:10.1007/s10709-005-5248-0.

Des Marais, D.L., and Rausher, M.D. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature **454**(7205): 762–765. doi:10.1038/nature07092.

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

Margarido, G.R.A., and Heckerman, D. 2015. ConPADE: Genome assembly ploidy estimation from next-generation sequencing data. PLOS Comput. Biol. **11**(4): e1004229. doi:10.1371/journal.pcbi.1004229.

Matsuda, M., Nagahama, Y., Shinomiya, A., Sato, T., Matsuda, C., Kobayashi, T., Morrey, C.E., Shibata, N., Asakawa, S., Shimizu, N., Hori, H., Hamaguchi, S., and Sakaizumi, M. 2002. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature **417**(6888): 559–563. doi:10.1038/nature751.

Matsuda, M., Sato, T., Toyazaki, Y., Nagahama, Y., Hamaguchi, S., and Sakaizumi, M. 2003. *Oryzias curvinotus* has DMY, a gene that is required for male development in the medaka, *O. latipes*. Zoolog. Sci. **20**(2): 159–161. doi:10.2108/zsj.20.159.

McCraney, W.T., Saski, C.A., and Guyon, J.R. 2012. Isolation and characterization of 12 microsatellites for the commercially important sablefish, *Anoplopoma fimbria*.

Conserv. Genet. Resour. **4**(2): 415–417. doi:10.1007/s12686-011-9563-8.

McIntosh, C.J., Lun, S., Lawrence, S., Western, A.H., McNatty, K.P., and Juengel, J.L. 2008. The proregion of mouse BMP15 regulates the cooperative interactions of BMP15 and GDF9. Biol. Reprod. **79**(5): 889–896. doi:10.1095/biolreprod.108.068163.

Messmer, A.M., Rondeau, E.B., Jantzen, S.G., Lubieniecki, K.P., Davidson, W.S., and Koop, B.F. 2011. Assessment of population structure in Pacific *Lepeophtheirus salmonis* (Kroyer) using single nucleotide polymorphism and microsatellite genetic markers. Aquaculture **320**(3–4): 183–192. doi:10.1016/j.aquaculture.2010.09.033.

Miller, J.R., Zhou, P., Mudge, J., Gurtowski, J., Lee, H., Ramaraj, T., Walenz, B.P., Liu, J., Stupar, R.M., Denny, R., Song, L., Singh, N., Maron, L.G., McCouch, S.R., McCombie, W.R., Schatz, M.C., Tiffin, P., Young, N.D., and Silverstein, K.A.T. 2017. Hybrid assembly with long and short reads improves discovery of gene family expansions. BMC Genomics **18**(1): 541. doi:10.1186/s12864-017-3927-8.

Miller, L.M., and Kapuscinski, A.R. 1996. Microsatellite DNA markers reveal new levels of genetic variation in northern pike. Trans. Am. Fish. Soc. **125**(6): 971–977. doi:10.1577/1548-8659(1996)125<0971:NMDMRN>2.3.CO;2.

Miller, L.M., and Senanan, W. 2003. A review of northern pike population genetics research and its implications for management. North Am. J. Fish. Manag. **23**(1): 297–306. doi:10.1577/1548-8675(2003)023<0297.

Miya, M., and Nishida, M. 1999. Organization of the mitochondrial genome of a deep-sea fish, *Gonostoma gracile* (Teleostei: Stomiiformes): First example of transfer RNA gene rearrangements in bony fishes. Mar. Biotechnol. **1**(5): 416–426. doi:10.1007/PL00011798.

Moen, T., Hayes, B., Baranski, M., Berg, P.R., Kjøglum, S., Koop, B.F., Davidson, W.S., Omholt, S.W., and Lien, S. 2008. A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics **9**(1): 223. doi:10.1186/1471-2164-9-223.

Morgan, T.H. 1911. The origin of five mutations in eye color in *Drosophila* and their modes of inheritance. Science **33**(849): 534–537. doi:10.1126/science.33.849.534-a.

Morgan, T.H. 1916. A critique of the theory of evolution. Princeton University Press, Princeton, NJ. doi:10.5962/bhl.title.1259.

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

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. 1986. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harb. Symp. Quant. Biol. **51**(1): 263–273. doi:10.1101/SQB.1986.051.01.032.

Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. 2017. Reconstructing the genome of the most recent common ancestor of flowering plants. Nat. Genet. **49**(4): 490–496. doi:10.1038/ng.3813.

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**(1): 163–170. doi:10.1534/genetics.111.137497.

Nadeau, J.H. 1989. Maps of linkage and synteny homologies between mouse and man. Trends Genet. **5**: 82–86. doi:10.1016/0168-9525(89)90031-0.

Neafsey, D.E., and Palumbi, S.R. 2003. Genome size evolution in pufferfish: A comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome Res. **13**(5): 821–830. doi:10.1101/gr.841703.

Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M., and Smith, W.L. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. **109**(34): 13698–13703. doi:10.1073/pnas.1206625109.

Nef, S., Schaad, O., Stallings, N.R., Cederroth, C.R., Pitetti, J.L., Schaer, G., Malki, S., Dubois-Dauphin, M., Boizet-Bonhoure, B., Descombes, P., Parker, K.L., and Vassalli, J.D. 2005. Gene expression during sex determination reveals a robust female genetic program at the onset of ovarian development. Dev. Biol. **287**(2): 361–377. doi:10.1016/j.ydbio.2005.09.008.

Nelson, J.S. 2006. Fishes of the World, 4th Edition. Wiley.

Nilsson, P.A., Skov, C., and Farrell, J.M. 2008. Current and future directions for pike ecology and management: a summary and synthesis. Hydrobiologia **601**(1): 137–141. doi:10.1007/s10750-007-9260-5.

NOAA Fisheries: Office of Science & Technology. 2012. Commercial Fisheries : Annual Landings. Available from http://www.st.nmfs.noaa.gov/st1/commercial/landings/annual_landings.html.

Nordin, R.N., and Pommen, L.W. 1985. Peace River Area Charlie Lake Sub-Basin: Water Quality Assessment and Objectives. Available from http://www.env.gov.bc.ca/wat/wq/objectives/charlie/Charlietech.pdf.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**(1): 205–217. doi:10.1006/jmbi.2000.4042.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. 1997. Evolution of genetic redundancy. Nature **388**(6638): 167–171. doi:10.1038/40618.

Ochman, H., Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784): 299–304. doi:10.1038/35012500.

Ohno, S. 1970. Evolution by Gene Duplication. Springer-Verlag. doi:10.1007/978-3-642-86659-3.

Ohno, S., Muramoto, J., Christian, L., and Atkin, N.B. 1967. Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae. Chromosoma **23**(1): 1–9. doi:10.1007/BF00293307.

Ohno, S., Wolf, U., and Atkin, N.B. 1968. Evolution from fish to mammals by gene

duplication. Hereditas **59**(1): 169–187. doi:10.1111/j.1601-5223.1968.tb02169.x.

Olasagasti, F., Lieberman, K.R., Benner, S., Cherf, G.M., Dahl, J.M., Deamer, D.W., and Akeson, M. 2010. Replication of individual DNA molecules under electronic control using a protein nanopore. Nat Nano **5**(11): 798–806. doi:10.1038/nnano.2010.177.

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. Conserv. Genet. Resour. **6**(2): 451–454. doi:10.1007/s12686-013-0123-2.

Palaiokostas, C., Bekaert, M., Davie, A., Cowan, M.E., Oral, M., Taggart, J.B., Gharbi, K., McAndrew, B.J., Penman, D.J., and Migaud, H. 2013a. Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. BMC Genomics **14**(1): 566. doi:10.1186/1471-2164-14-566.

Palaiokostas, C., Bekaert, M., Khan, M.G.Q., Taggart, J.B., Gharbi, K., McAndrew, B.J., and Penman, D.J. 2013b. Mapping and validation of the major sex-determining region in nile tilapia (*Oreochromis niloticus* L.) using RAD sequencing. PLoS One **8**(7): e68389. doi:10.1371/journal.pone.0068389.

Palti, Y., Genet, C., Gao, G., Hu, Y., You, F.M., Boussaha, M., Rexroad, C.E., and Luo, M.C. 2012. A Second Generation Integrated Map of the Rainbow Trout (*Oncorhynchus mykiss*) Genome: Analysis of Conserved Synteny with Model Fish Genomes. Mar. Biotechnol. **14**(3): 343–357. doi:10.1007/s10126-011-9418-z.

Pan, Q., Anderson, J., Bertho, S., Herpin, A., Wilson, C., Postlethwait, J.H., Schartl, M., and Guiguen, Y. 2016. Vertebrate sex-determining genes play musical chairs. C. R. Biol. **339**(7): 258–262. doi:10.1016/j.crvi.2016.05.010.

Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A.J., Herwig, R., Vingron, M., and Lehrach, H. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res. **13**: 1056–1066. doi:10.1101/gr.874803.

Park, D., Jung, J.W., Choi, B.-S., Jayakodi, M., Lee, J., Lim, J., Yu, Y., Choi, Y.-S., Lee, M.-L.,

Park, Y., Choi, I.-Y., Yang, T.-J., Edwards, O.R., Nah, G., and Kwon, H.W. 2015. Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. BMC Genomics **16**(1): 1. doi:10.1186/1471-2164-16-1.

Passarge, E., Horsthemke, B., and Farber, R.A. 1999. Incorrect use of the term synteny. Nat. Genet. **23**(4): 387. doi:10.1038/70486.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc. Natl. Acad. Sci. **101**(26): 9903–9908. doi:10.1073/pnas.0307901101.

Van de Peer, Y., Maere, S., and Meyer, A. 2009. The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10**(10): 725–732. doi:10.1038/nrg2600.

Peichel, C.L., Ross, J.A., Matson, C.K., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Mori, S., Schluter, D., and Kingsley, D.M. 2004. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Curr. Biol. **14**(16): 1416–1424. doi:10.1016/j.cub.2004.08.030.

Peichel, C.L., Sullivan, S.T., Liachko, I., and White, M.A. 2016. Improvement of the threespine stickleback (*Gasterosteus aculeatus*) genome using a Hi-C-based Proximity-Guided Assembly method. bioRxiv: 68528. doi:10.1101/068528.

Pereira, S.L. 2000. Mitochondrial genome organization and vertebrate phylogenetics. Genet. Mol. Biol. **23**(4): 745–752. doi:10.1590/S1415-47572000000400008.

Phillips, R.B., Faber-Hammond, J., and Luckenbach, J.A. 2013. The sablefish (*Anoplopoma fimbria*) karyotype including the location of 5S and 18S rDNA and information on cell culture conditions. Aquac. Res. **44**(11): 1801–1804. doi:10.1111/j.1365-2109.2012.03177.x.

Pie, M.R., Torres, R.A., and Brito, D.M.A. 2006. Evolution of genome size in fishes: a phylogenetic test of the Hinegardner and Rosen hypothesis. Genetica **131**(1): 51. doi:10.1007/s10709-006-9112-7.

Piferrer, F. 2001. Endocrine sex control strategies for the feminization of teleost fish.

Aquaculture **197**(1–4): 229–281. doi:10.1016/S0044-8486(01)00589-0.

Polani, P.E., Briggs, J.H., Ford, C.E., Clarke, C.M., and Berg, J.M. 1960. A Mongol girl with 46 chromosomes. Lancet **1**(7127): 721–4. doi:10.1016/S0140-6736(60)90614-0.

Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M. V., Oparina, N.Y., Polozov, R. V., Nechipurenko, Y.D., and Grokhovsky, S.L. 2015. Non-random DNA fragmentation in next-generation sequencing. Sci. Rep. **4**(1): 4532. doi:10.1038/srep04532.

Postlethwait, J.H., Yan, Y.L., Gates, M.A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E.S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., Abduljabbar, T.S., Yelick, P., Beier, D., Joly, J.S., Larhammar, D., Rosa, F., Westerfield, M., Zon, L.I., Johnson, S.L., and Talbot, W.S. 1998. Vertebrate genome evolution and the zebrafish gene map. Nat. Genet. **18**(4): 345–9. doi:10.1038/ng0498-345.

Prada, C.F., Delprat, A., and Ruiz, A. 2011. Testing chromosomal phylogenies and inversion breakpoint reuse in *Drosophila*. The martensis cluster revisited. Chromosom. Res. **19**(2): 251–265. doi:10.1007/s10577-011-9195-6.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J.J., Grigoriev, I. V, Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V. V, Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W.H., Satoh, N., and Rokhsar, D.S. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature **453**(7198): 1064–1071. doi:10.1038/nature06967.

Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J., Fields, A., Hartley, P.D., Sugnet, C.W., Haussler, D., Rokhsar, D.S., and Green, R.E. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range

linkage. Genome Res. **26**(3): 342–350. doi:10.1101/gr.193474.115.

Ráb, P., and Crossman, E.J. 1994. Chromosomal NOR phenotypes in North American pikes and pickerels, genus *Esox*, with notes on the Umbridae (Euteleostei: Esocae). Can. J. Zool. **72**(11): 1951–1956. doi:10.1139/z94-265.

R Core Team Computing. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, https://www.R-project.org/. doi:ISBN 3-900051-07-0.

Rebrikov, D. V., Desai, S.M., Siebert, P.D., and Lukyanov, S.A. 2004. Suppression Subtractive Hybridization. *In* Gene Expression Profiling. Humana Press. pp. 107–134. doi:10.1385/1-59259-751-3:107.

Reichwald, K., Petzold, A., Koch, P., Cellerino, A., Englert, C., and Platzer, M. 2015. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. Cell **163**(6): 1527–1538. doi:10.1016/j.cell.2015.10.071.

Reid, D.P., Smith, C.A., Rommens, M., Blanchard, B., Martin-Robichaud, D., and Reith, M. 2007. A genetic linkage map of Atlantic halibut (*Hippoglossus hippoglossus* L.). Genetics **177**(2): 1193–1205. doi:10.1534/genetics.107.075374.

Renwick, J.H. 1971. The mapping of human chromosomes. Annu. Rev. Genet. **5**: 81–120. doi:10.1146/annurev.ge.05.120171.000501.

Rexroad, C.E., Palti, Y., Gahr, S.A., and Vallejo, R.L. 2008. A second generation genetic map for rainbow trout (*Oncorhynchus mykiss*). BMC Genet. **9**(1): 74. doi:10.1186/1471-2156-9-74.

Rise, M.L., von Schalburg, K.R., Brown, G.D., Mawer, M.A., Devlin, R.H., Kuipers, N., Busby, M., Beets-Sargent, M., Alberto, R., Gibbs, A.R., Hunt, P., Shukin, R., Zeznik, J.A., Nelson, C., Jones, S.R.M., Smailus, D.E., Jones, S.J.M., Schein, J.E., Marra, M.A., Butterfield, Y.S.N., Stott, J.M., Ng, S.H.S., Davidson, W.S., and Koop, B.F. 2004. Development and application of a salmonid EST database and cDNA microarray: Data mining and interspecific hybridization characteristics. Genome Res. **14**(3): 478–490. doi:10.1101/gr.1687304.

Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Lien, S., Martin, S.A.M., Holland, P.W.H., Sandve, S.R., and Macqueen, D.J. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. **18**(1): 111. doi:10.1186/s13059-017-1241-z.

Rondeau, E.B., Laurie, C. V., Johnson, S.C., and Koop, B.F. 2016. A PCR assay detects a male-specific duplicated copy of Anti-Müllerian hormone (*amh*) in the lingcod (*Ophiodon elongatus*). BMC Res. Notes **9**(1): 230. BioMed Central, London. doi:10.1186/s13104-016-2030-6.

Rondeau, E.B., Messmer, A.M., Sanderson, D.S., Jantzen, S.G., von Schalburg, K.R., Minkley, D.R., Leong, J.S., Macdonald, G.M., Davidsen, A.E., Parker, W.A., Mazzola, R.S., Campbell, B., and Koop, B.F. 2013. Genomics of sablefish (*Anoplopoma fimbria*): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. BMC Genomics **14**(1): 452. doi:10.1186/1471-2164-14-452.

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**(7): e102089. doi:10.1371/journal.pone.0102089.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hickey, G., Hinrichs, A.S., Hubley, R., Karolchik, D., Learned, K., Lee, B.T., Li, C.H., Miga, K.H., Nguyen, N., Paten, B., Raney, B.J., Smit, A.F.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., and Kent, W.J. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. **43**(D1): D670–D681. doi:10.1093/nar/gku1177.

Ross, J.A., Urton, J.R., Boland, J., Shapiro, M.D., and Peichel, C.L. 2009. Turnover of sex chromosomes in the stickleback fishes (Gasterosteidae). PLoS Genet. **5**(2):

e1000391. doi:10.1371/journal.pgen.1000391.

Rozen, S., and Skaletsky, H. 2000. Primer3 on the WWW for General Users and for Biologist Programmers. Methods Mol. Biol. **132**: 365–386. doi:10.1385/1-59259-192-2:365.

Sakamoto, T., Danzmann, R.G., Gharbi, K., Howard, P., Ozaki, A., Khoo, S.K., Woram, R.A., Okamoto, N., Ferguson, M.M., Holm, L., Guyomard, R., and Hoyheim, B. 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. Genetics **155**(3): 1331–1345.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 2000. Molecular Cloning: A Laboratory Manual. *In* 3rd edition. Cold Spring Harbor Laboratory Press. Available from http://books.google.ca/books?id=Bosc5JVxNpkC.

Sandve, S.R., Rohlfs, R. V., and Hvidsten, T.R. 2017. Subfunctionalization versus neofunctionalization after whole-genome duplication. bioRxiv: 190090. doi:10.1101/190090.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., and Smith, M. 1977. Nucleotide sequence of bacteriophage ϕX174 DNA. Nature **265**(5596): 687–695. doi:10.1038/265687a0.

Sarropoulou, E., Nousdili, D., Magoulas, A., and Kotoulas, G. 2008. Linking the genomes of nonmodel teleosts through comparative genomics. Mar. Biotechnol. **10**(3): 227–233. doi:10.1007/s10126-007-9066-5.

Sawatari, E., Shikina, S., Takeuchi, T., and Yoshizaki, G. 2007. A novel transforming growth factor-β superfamily member expressed in gonadal somatic cells enhances primordial germ cell and spermatogonial proliferation in rainbow trout (*Oncorhynchus mykiss*). Dev. Biol. **301**(1): 266–275. doi:10.1016/j.ydbio.2006.10.001.

von Schalburg, K.R., Gowen, B.E., Leong, J.S., Rondeau, E.B., Davidson, W.S., and Koop, B.F. 2017. Subcellular localization and characterization of estrogenic pathway

regulators and mediators in Atlantic salmon spermatozoal cells. Histochem. Cell Biol.: 1–22. doi:10.1007/s00418-017-1611-3.

Schatz, M.C., Witkowski, J., and McCombie, W.R. 2012. Current challenges in *de novo* plant genome sequencing and assembly. Genome Biol. **13**(4): 243. doi:10.1186/gb4015.

Schlotterer, C., Amos, B., and Tautz, D. 1991. Conservation of plolymorphic simple sequence loci in cetacean species. Nature **354**: 63–65. doi:10.1038/354063a0.

Schwager, E.E., Sharma, P.P., Clarke, T., Leite, D.J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T., Buffry, A.D., Chao, H., Dinh, H., Doddapaneni, H., Dugan, S., Eibner, C., Extavour, C.G., Funch, P., Garb, J., Gonzalez, L.B., Gonzalez, V.L., Griffiths-Jones, S., Han, Y., Hayashi, C., Hilbrant, M., Hughes, D.S.T., Janssen, R., Lee, S.L., Maeso, I., Murali, S.C., Muzny, D.M., Nunes da Fonseca, R., Paese, C.L.B., Qu, J., Ronshaugen, M., Schomburg, C., Schönauer, A., Stollewerk, A., Torres-Oliva, M., Turetzek, N., Vanthournout, B., Werren, J.H., Wolff, C., Worley, K.C., Bucher, G., Gibbs, R.A., Coddington, J., Oda, H., Stanke, M., Ayoub, N.A., Prpic, N.-M., Flot, J.-F., Posnien, N., Richards, S., and McGregor, A.P. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol. **15**(1): 62. doi:10.1186/s12915-017-0399-x.

Schwarz, D., Robertson, H.M., Feder, J.L., Varala, K., Hudson, M.E., Ragland, G.J., Hahn, D.A., and Berlocher, S.H. 2009. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. BMC Genomics **10**(1): 633. doi:10.1186/1471-2164-10-633.

Seeb, J.E., Seeb, L.W., Oates, D.W., and Utter, F.M. 1987. Genetic variation and post-glacial dispersal of populations of northern pike (*Esox lucius*) in North America. Can. J. Fish. Aquat. Sci. **44**(3): 556–561. doi:10.1139/f87-068.

Sémon, M., and Wolfe, K.H. 2007. Rearrangement rate following the whole-genome duplication in teleosts. Mol. Biol. Evol. **24**(3): 860–867. doi:10.1093/molbev/msm003.

Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., van Heeringen, S.J., Quigley, I., Heinz, S., Ogino, H., Ochi, H., Hellsten, U., Lyons, J.B., Simakov, O., Putnam, N., Stites, J., Kuroki, Y., Tanaka, T., Michiue, T., Watanabe, M., Bogdanovic, O., Lister, R., Georgiou, G., Paranjpe, S.S., van Kruijsbergen, I., Shu, S., Carlson, J., Kinoshita, T., Ohta, Y., Mawaribuchi, S., Jenkins, J., Grimwood, J., Schmutz, J., Mitros, T., Mozaffari, S. V., Suzuki, Y., Haramoto, Y., Yamamoto, T.S., Takagi, C., Heald, R., Miller, K., Haudenschild, C., Kitzman, J., Nakayama, T., Izutsu, Y., Robert, J., Fortriede, J., Burns, K., Lotay, V., Karimi, K., Yasuoka, Y., Dichmann, D.S., Flajnik, M.F., Houston, D.W., Shendure, J., DuPasquier, L., Vize, P.D., Zorn, A.M., Ito, M., Marcotte, E.M., Wallingford, J.B., Ito, Y., Asashima, M., Ueno, N., Matsuda, Y., Veenstra, G.J.C., Fujiyama, A., Harland, R.M., Taira, M., and Rokhsar, D.S. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature **538**(7625): 336–343. doi:10.1038/nature19840.

Setiamarga, D.H.E., Miya, M., Yamanoue, Y., Azuma, Y., Inoue, J.G., Ishiguro, N.B., Mabuchi, K., and Nishida, M. 2009. Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. Biol. Lett. **5**(6): 812–816. doi:10.1098/rsbl.2009.0419.

Shen, Y., Chalopin, D., Garcia, T., Boswell, M., Boswell, W., Shiryev, S.A., Agarwala, R., Volff, J.-N., Postlethwait, J.H., Schartl, M., Minx, P., Warren, W.C., and Walter, R.B. 2016. *X. couchianus* and *X. hellerii* genome models provide genomic variation insight among *Xiphophorus* species. BMC Genomics **17**(1): 37. doi:10.1186/s12864-015-2361-z.

Shiina, T., Dijkstra, J.M., Shimizu, S., Watanabe, A., Yanagiya, K., Kiryu, I., Fujiwara, A., Nishida-Umehara, C., Kaba, Y., Hirono, I., Yoshiura, Y., Aoki, T., Inoko, H., Kulski, J.K., and Ototake, M. 2005. Interchromosomal duplication of major histocompatibility complex class I regions in rainbow trout (*Oncorhynchus mykiss*), a species with a presumably recent tetraploid ancestry. Immunogenetics **56**(12): 878–893.

doi:10.1007/s00251-004-0755-1.

Shinohara, G., and Imamura, H. 2007. Revisiting recent phylogenetic studies of "Scorpaeniformes." Ichthyol. Res. **54**(1): 92–99. doi:10.1007/s10228-006-0379-6.

Shtivelman, E., Lifshitz, B., Gale, R.P., and Canaani, E. 1985. Fused transcript of *abl* and *bcr* genes in chronic myelogenous leukaemia. Nature **315**(6020): 550–554. doi:10.1038/315550a0.

Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A., and Lukyanov, S.A. 1995. An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. **23**(6): 1087–1088. doi:10.1093/nar/23.6.1087.

Singer, A., Perlman, H., Yan, Y., Walker, C., Corley-Smith, G., Brandhorst, B., and Postlethwait, J. 2002. Sex-specific recombination rates in zebrafish (*Danio rerio*). Genetics **160**(2): 649–657.

Singh, P.P., Arora, J., and Isambert, H. 2015. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. PLoS Comput. Biol. **11**(7): e1004394. doi:10.1371/journal.pcbi.1004394.

Smit, A., Hubley, R., and Green, P. 2010. RepeatMasker Open-4.0. Available from http://www.repeatmasker.org.

Smith, C.A., Roeszler, K.N., Ohnesorg, T., Cummins, D.M., Farlie, P.G., Doran, T.J., and Sinclair, A.H. 2009. The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. Nature **461**(7261): 267–271. doi:10.1038/nature08298.

Smith, E.K., Guzmán, J.M., and Luckenbach, J.A. 2013. Molecular cloning, characterization, and sexually dimorphic expression of five major sex differentiation-related genes in a Scorpaeniform fish, sablefish (*Anoplopoma fimbria*). Comp. Biochem. Physiol. Part B **165**(2): 125–137. doi:10.1016/j.cbpb.2013.03.011.

Smith, W.L., and Wheeler, W.C. 2004. Polyphyly of the mail-cheeked fishes (Teleostei:

Scorpaeniformes): Evidence from mitochondrial and nuclear sequence data. Mol. Phylogenet. Evol. **32**(2): 627–646. doi:10.1016/j.ympev.2004.02.006.

Snel, B., Bork, P., and Huynen, M. 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet. **16**(1): 9–11. doi:10.1016/S0168-9525(99)01924-1.

Soderlund, C., Bomhoff, M., and Nelson, W.M. 2011. SyMAP v3.4: A turnkey synteny system with application to plant genomes. Nucleic Acids Res. **39**(10): e68. doi:10.1093/nar/gkr123.

Soderlund, C., Nelson, W., Shoemaker, A., and Paterson, A. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res. **16**(9): 1159–1168. doi:10.1101/gr.5396706.

Stanke, M., and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**(SUPPL. 2): ii215-ii225. doi:10.1093/bioinformatics/btg1080.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R.B., Tina, K.G., Espelund, M., Nepal, C., Previti, C., Karlsen, B.O., Moum, T., Skage, M., Berg, P.R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S.D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S.W., Stenseth, N.C., and Jakobsen, K.S. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature **477**(7363): 207–210. doi:10.1038/nature10342.

Stephens, S.G. 1951. Possible significances of duplication in evolution. Adv. Genet. **4**: 247–265. doi:10.1016/S0065-2660(08)60237-0.

Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G., and Bayley, H. 2009. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc. Natl. Acad. Sci. **106**(19): 7702–7707. doi:10.1073/pnas.0901054106.

Sturtevant, A.H. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. J. Exp. Zool. **14**(1): 43–59. doi:10.1002/jez.1400140104.

Sun, Y.-B., Xiong, Z.-J., Xiang, X.-Y., Liu, S.-P., Zhou, W.-W., Tu, X.-L., Zhong, L., Wang, L., Wu, D.-D., Zhang, B.-L., Zhu, C.-L., Yang, M.-M., Chen, H.-M., Li, F., Zhou, L., Feng, S.-H., Huang, C., Zhang, G.-J., Irwin, D., Hillis, D.M., Murphy, R.W., Yang, H.-M., Che, J., Wang, J., and Zhang, Y.-P. 2015. Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. Proc. Natl. Acad. Sci. **112**(11): E1257–E1262. doi:10.1073/pnas.1501764112.

Sutherland, B.J.G., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., and Bernatchez, L. 2016. Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. Genome Biol. Evol. **8**(12): evw262. doi:10.1093/gbe/evw262.

Takehana, Y., Matsuda, M., Myosho, T., Suster, M.L., Kawakami, K., Shin-I, T., Kohara, Y., Kuroki, Y., Toyoda, A., Fujiyama, A., Hamaguchi, S., Sakaizumi, M., and Naruse, K. 2014. Co-option of *Sox3* as the male-determining factor on the Y chromosome in the fish *Oryzias dancena*. Nat. Commun. **5**: 4157. doi:10.1038/ncomms5157.

Talevich, E., Invergo, B.M., Cock, P.J., and Chapman, B.A. 2012. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics **13**(1): 209. doi:10.1186/1471-2105-13-209.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17**(16): 6463–6471. doi:10.1093/nar/17.16.6463.

Tautz, D., and Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**(10): 4127–4138. doi:10.1093/nar/12.10.4127.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. 2003. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. Genome Res. **13**(3): 382–390. doi:10.1101/gr.640303.

Taylor, J.S., Van de Peer, Y., Braasch, I., and Meyer, A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos. Trans. R. Soc. B Biol. Sci. **356**(1414): 1661–1679. doi:10.1098/rstb.2001.0975.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**(6814): 796–815. doi:10.1038/35048692.

The C. elegans sequencing consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science **282**(5396): 2012–2018. doi:10.1126/science.282.5396.2012.

Thomas, J.W., Summers, T.J., Lee-Lin, S.-Q., Maduro, V.V.B., Idol, J.R., Mastrian, S.D., Ryan, J.F., Jamison, D.C., and Green, E.D. 2000. Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. Genome Res. **10**(5): 624–633. doi:10.1101/gr.10.5.624.

Thorgaard, G.H., Bailey, G.S., Williams, D., Buhler, D.R., Kaattari, S.L., Ristow, S.S., Hansen, J.D., Winton, J.R., Bartholomew, J.L., Nagler, J.J., Walsh, P.J., Vijayan, M.M., Devlin, R.H., Hardy, R.W., Overturf, K.E., Young, W.P., Robison, B.D., Rexroad, C., and Palti, Y. 2002. Status and opportunities for genomics research with rainbow trout. Comp. Biochem. Physiol. - B Biochem. Mol. Biol. **133**(4): 609–646. doi:10.1016/S1096-4959(02)00167-7.

Tseng, E., and Underwood, J.G. 2013. Full Length cDNA Sequencing on the PacBio RS. J. Biomecular Tech. **24**(May): S45–S45. Bethesda, MD.

Tuttle, E.M., Bergland, A.O., Korody, M.L., Brewer, M.S., Newhouse, D.J., Minx, P., Stager, M., Betuel, A., Cheviron, Z.A., Warren, W.C., Gonser, R.A., and Balakrishnan, C.N. 2016. Divergence and functional degradation of a sex chromosome-like supergene. Curr. Biol. **26**(3): 344–350. doi:10.1016/j.cub.2015.11.069.

Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer, Y. 2004. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc. Natl. Acad. Sci.

101(6): 1638–1643. doi:10.1073/pnas.0307968100.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. Di, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V, Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K.,

Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K. V, Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. 2001. The Sequence of the Human Genome. Science **291**(5507): 1304–1351. doi:10.1126/science.1058040.

Vinogradov, A.E. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. Cytometry **31**(2): 100–109. doi:10.1002/(SICI)1097-0320(19980201)31:2<100::AID-CYTO5>3.0.CO;2-Q.

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**(5499): 2114–2117. doi:10.1126/science.290.5499.2114.

Volff, J.N. 2006. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays **28**(9): 913–922. doi:10.1002/bies.20452.

Voorrips, R.E. 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. J. Hered. **93**(1): 77–78. doi:10.1093/jhered/93.1.77.

Walsh, P.S., Metzger, D.A., and Higuchi, R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques **10**(4): 506—513.

Wang, B., Ekblom, R., Bunikis, I., Siitari, H., and Höglund, J. 2014. Whole genome

sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics **15**(1): 180. doi:10.1186/1471-2164-15-180.

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., and Lander, E.S. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280**(5366): 1077–1082. doi:10.1126/science.280.5366.1077.

Wang, J.-T., Li, J.-T., Zhang, X.-F., and Sun, X.-W. 2012. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). BMC Genomics **13**(1): 96. doi:10.1186/1471-2164-13-96.

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.). Genet. Mol. Biol. **34**(1): 169–172. doi:10.1590/S1415-47572010005000114.

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

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., and Agarwal, P. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**(6915): 520–562. doi:10.1038/nature01262.

Weber, J.L., and May, P.E. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44**(3): 388–96.

Wegrzyn, J.L., Lin, B.Y., Zieve, J.J., Dougherty, W.M., Martínez-García, P.J., Koriabine, M., Holtz-Morris, A., DeJong, P., Crepeau, M., Langley, C.H., Puiu, D., Salzberg, S.L., Neale, D.B., and Stevens, K.A. 2013. Insights into the loblolly pine genome:

Characterization of BAC and fosmid sequences. PLoS One **8**(9): e72439. doi:10.1371/journal.pone.0072439.

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. 2016. The physiology and habitat of the last universal common ancestor. Nat. Microbiol. **1**(9): 16116. doi:10.1038/nmicrobiol.2016.116.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A.H. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**(12): 973–982. doi:10.1038/nrg2165.

Wolfe, K. 2000. Robustness - it's not where you think it is. Nat. Genet. **25**(1): 3–4. doi:10.1038/75560.

Wolfe, K.H., and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**(6634): 708–713. doi:10.1038/42711.

Woram, R.A., McGowan, C., Stout, J.A., Gharbi, K., Ferguson, M.M., Hoyheim, B., Davidson, E.A., Davidson, W.S., Rexroad, C., and Danzmann, R.G. 2004. A genetic linkage map for Arctic char (*Salvelinus alpinus*): evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome **47**(2): 304–315. doi:10.1139/g03-127.

Wu, T.D., and Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**(9): 1859–1875. doi:10.1093/bioinformatics/bti310.

Wybouw, N., Pauchet, Y., Heckel, D.G., and van Leeuwen, T. 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biol. Evol. **8**(6): 1785–1801. doi:10.1093/gbe/evw119.

Wyman, S.K., Jansen, R.K., and Boore, J.L. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**(17): 3252–3255. doi:10.1093/bioinformatics/bth352.

Xia, J., Liu, F., Zhu, Z., Fu, J., Feng, J., Li, J., and Yue, G. 2010. A consensus linkage map of

the grass carp (*Ctenopharyngodon idella*) based on microsatellites and SNPs. BMC Genomics **11**(1): 135. doi:10.1186/1471-2164-11-135.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, L., Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shao, C., Song, L., Sun, J., Ji, P., Wang, J., Li, Q., Xu, L., Sun, F., Feng, J., Wang, C., Wang, S., Wang, B., Li, Y., Zhu, Y., Xue, W., Zhao, L., Wang, J., Gu, Y., Lv, W., Wu, K., Xiao, J., Wu, J., Zhang, Z., Yu, J., and Sun, X. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat. Genet. **46**(11): 1212–1219. doi:10.1038/ng.3098.

Yan, Y.-L., Desvignes, T., Bremiller, R., Wilson, C., Dillon, D., High, S., Draper, B., Buck, C.L., and Postlethwait, J. 2017. The gonadal soma controls ovarian follicle proliferation through *Gsdf* in zebrafish. Dev. Dyn. **246**(11): 924–945. doi:10.1002/dvdy.24579.

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*. Curr. Biol. **22**(15): 1423–1428. doi:10.1016/j.cub.2012.05.045.

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**(3): 486–496. doi:10.1111/eva.12032.

Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I. 2017. ARCS: Assembly roundup by chromium scaffolding. Bioinformatics: btx675. doi:10.1093/bioinformatics/btx675.

Yoshimoto, S., Ikeda, N., Izutsu, Y., Shiba, T., Takamatsu, N., and Ito, M. 2010. Opposite roles of DMRT1 and its W-linked paralogue, DM-W, in sexual dimorphism of *Xenopus laevis*: implications of a ZZ/ZW-type sex-determining system. Development **137**(15): 2519–2526. doi:10.1242/dev.048751.

Zhang, J., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. **95**(7): 3708–3713. doi:10.1073/pnas.95.7.3708.

Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W.C., Strausberg, R.L., and Fraser, C.M. 2004. Human, mouse, and rat genome large-scale rearrangements: Stability versus speciation. Genome Res. **14**: 1851–1860. doi:10.1101/gr.2663304.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., Mudivarti, P.A., Wyatt, P.W., Bharadwaj, R., Makarewicz, A.J., Li, Y., Belgrader, P., Price, A.D., Lowe, A.J., Marks, P., Vurens, G.M., Hardenbol, P., Montesclaros, L., Luo, M., Greenfield, L., Wong, A., Birch, D.E., Short, S.W., Bjornson, K.P., Patel, P., Hopmans, E.S., Wood, C., Kaur, S., Lockwood, G.K., Stafford, D., Delaney, J.P., Wu, I., Ordonez, H.S., Grimes, S.M., Greer, S., Lee, J.Y., Belhocine, K., Giorda, K.M., Heaton, W.H., McDermott, G.P., Bent, Z.W., Meschi, F., Kondov, N.O., Wilson, R., Bernate, J.A., Gauby, S., Kindwall, A., Bermejo, C., Fehr, A.N., Chan, A., Saxonov, S., Ness, K.D., Hindson, B.J., and Ji, H.P. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. **34**(3): 303–311. doi:10.1038/nbt.3432.

Zheng, X., Kuang, Y., Zhang, X., Lu, C., Cao, D., Li, C., and Sun, X. 2011. A genetic linkage map and comparative genome analysis of common carp (*Cyprinus carpio* L.) using microsatellites and SNPs. Mol. Genet. Genomics **286**(3–4): 261–277. doi:10.1007/s00438-011-0644-x.

Zou, M., Guo, B., Tao, W., Arratia, G., and He, S. 2012. Integrating multi-origin expression data improves the resolution of deep phylogeny of ray-finned fish (Actinopterygii). Sci. Rep. **2**(1): 665. doi:10.1038/srep00665.

Appendix A

Publications and presentations during program

Peer-Reviewed

J Le Luyer, M Laporte, TD Beacham, KH Kaukinen, RE Withler, JS Leong, **EB Rondeau**, BF Koop, L Bernatchez. Parallel epigenetic modifications induced by hatchery rearing in a Pacific Salmon. PNAS 114(49): 12964-12969.

KR von Schalburg, BE Gowen, JS Leong, **EB Rondeau**, WS Davidson, BF Koop (2017). Subcellular localization and characterization of estrogenic pathway regulators and mediators in Atlantic salmon spermatozoal cells. Histochemistry and Cell Biology In Press: 1-22.

S Lien, BF Koop, SR Sandve, JR Miller, MP Kent, T Nome, TR Hvidsten, JS Leong, DR Minkley, A Zimin, F Grammes, H Grove, A Gjuvsland, B Walenz, RA Hermansen, KR von Schalburg**, EB Rondeau**, A Di Genova, JKA Samy, JO Vik, MD Vigeland, L Caler, U Grimholt, S Jentoft, DI Vage, P de Jong, T Moen, M Baranski, Y Palti, D Smith, JA Yorke, AJ Nederbragt, A Tooming-Klunderud, KS Jakobsen, X Jiang, D Fan, Y Hu, DA Liberles, R Vidal, P Iturra, SJM Jones, I Jonassen, A Maass, SW Omholt, WS Davidson (2016). The Atlantic salmon genome provides insights into rediploidization. Nature: 17164.

EB Rondeau, CV Laurie**,** SC Johnson, BF Koop (2016). A PCR assay detects a male-specific duplicated copy of Anti-Müllerian hormone (*amh*) in the lingcod (*Ophiodon elongatus*). *BMC Research Notes* **9**:230.

EB Rondeau, DR Minkley, JS Leong, AM Messmer, JR Jantzen, KR von Schalburg, C Lemon, NH Bird, BF Koop (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.

EB Rondeau, AM Messmer, DS Sanderson, SG Jantzen, KR von Schalburg, DR Minkley, JS Leong, GM Macdonald, AE Davidsen, WA Parker, RSA Mazzola, B Campbell, BF Koop (2013). Genomics of Sablefish (*Anoplopoma fimbria*): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. *BMC Genomics* **14**:452.

RJ Nelson, C Bouchard, M Madsen, K Praebel, **E Rondeau**, K von Schalburg, JS Leong, S Jantzen, Z Sandwith, S Puckett, A Messmer, SE Fevolden, BF Koop (2013). Microsatellite loci for genetic analysis of the arctic gadids *Boreogadus saida* and *Arctogadus glacialis*. *Conservation Genetics Resources* **5**: 445-448

KR von Schalburg, BE Gowen, **EB Rondeau**, NW Johnson, DR Minkley, JS Leong, WS Davidson, BF Koop (2013). Sex-specific expression, synthesis and localization of aromatase regulators in one-year-old Atlantic salmon ovaries and testes. *CBP Part B* **164**: 236-246.

AM Messmer, **EB Rondeau,** SG Jantzen, KP Lubieniecki, WS Davidson, BF Koop (2011). Assessment of population structure in Pacific *Lepeophtheirus salmonis* (Krøyer) using single nucleotide polymorphism and microsatellite genetic markers. *Aquaculture* **320**: 183-192.

KR von Schalburg, M Yasuike, R Yazawa, JG de Boer, L Reid, S So, A Robb, **EB Rondeau**, RB Phillips, WS Davidson, BF Koop (2011). Regulation and expression of sexual differentiation factors in embryonic and extragonadal tissues of Atlantic salmon. *BMC Genomics* **12**:31.

<u>Submitted</u>

AM Messmer*, JS Leong*, **EB Rondeau***, A Mueller, CA Despins, DR Minkley, B Boyce, D Morrison, MD Fast, RG Danzmann, BF Koop. A 200K SNP chip reveals a novel Pacific salmon louse genotype linked to differential efficacy of Emamectin benzoate. In revision, Marine Genomics

EB Rondeau, JS Leong, DR Minkley, BF Koop. The second version of the northern pike *(Esox lucius*) genome helps trace the evolution of the salmonids post whole-genome duplication. To be submitted, BMC Genomics

Presentations

The second version of the northern pike (*Esox lucius*) genome helps trace the evolution of the salmonids post whole-genome duplication. Coastwide salmonid genetics 2016 in Astoria, OR.

The second version of the northern pike (*Esox lucius*) genome helps trace the evolution of the salmonids post whole-genome duplication. **Invited speaker** at the 3rd International Conference on Integrative Salmonid Biology, April 2016 in Puerto Varas, Chile.

The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. **Invited speaker** at the 2nd International Conference on Integrative Salmonid Biology, June 2014 in Vancouver, BC.

An EST/RNAseq Library and First-generation Linkage Map for Sablefish (*Anoplopoma fimbria*). Oral and Poster presentation at Plant and Animal Genomics (PAG), January 2012 in San Diego, CA.

Appendix B

Supplementary Files, Chapter 2

Website: https://doi.org/10.1186/1471-2164-14-452

Links to access original published supplementary files are found below.

Additional File B1 (Figure S1) **- Sablefish mitochondrial genome**. A graphical representation of the sablefish mitochondrial genome, including the relative placement of the 13 genes, 22 tRNA and 2 rRNAs: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM1_ESM.pdf

Additional File B2 (Table S1) - **Sablefish Microsatellite Primers**. Details for all primers designed that successfully targeted polymorphic microsatellites. Primer sequences are detailed, as well as design sequence, BLAST and BLAT (stickleback) ID and repeat motif are given: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM2_ESM.xls

Additional File B3 (Table S2) - **Sablefish SNP Primers**. Details for all primers designed that successfully targeted one or more polymorphic SNPs. Primer sequences are detailed, as well as design sequence, BLAST and BLAT (stickleback) ID, and total number of SNPs identified between the two families: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM3_ESM.xls

Additional File B4 (Table S3) - **Sablefish SNPs**. List of all identified SNPs and their position within the surrounding sequence: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM4_ESM.xls

Additional File B5 (Figure S2) **- Sablefish linkage Map: Male, female and merged maps**. All 24 linkage groups are presented. Each triplicate displays the merged linkage group in the middle, with the male-specific and female-specific linkage map to the left and right respectively: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM5_ESM.pdf

Additional File B6 (Table S4) **- SNP markers identified through Genotyping-by-sequencing**. List of SNPs identified through Genotyping-by-sequencing and parameters used in Stacks: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM6_ESM.xls

Additional File B7 (Table S5) - ***Gsdf* and upstream promoter region**. Sequences for Genbank:KC623942 and Genbank:KC623943 masked to show major features including sex-specific sequences, *gsdf*, sex-specific sequences and repeat elements: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM7_ESM.doc

Additional File B8 (Table S6) - **Mitochondrial genomes used in the phylogenetic analysis of the Scorpaeniformes and closely related Gasterosteiformes and Perciformes**. All mitochondrial genome accession numbers used in the phylogenetic analysis. Accession numbers not included in Kawahara et al. (Kawahara et al. 2008) are marked in bold: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM8_ESM.xls

Additional File B9 (Supplementary Methods) - **Primers and relevant information for Sex-specific amplifications**. All primer sequences used to amplify and sequence the region containing the sex-specific sequences are given, as well as relevant information such as annealing temperatures and multiplexing strategies: https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-14-452/MediaObjects/12864_2013_5158_MOESM9_ESM.doc

Appendix C

Supplementary Files, Chapter 3

Website: https://doi.org/10.1371/journal.pone.0102089

Links to access original published supplementary files are found below.

Additional File C1 (Figure S1) - **Individual male and sex-specific linkage maps for the northern pike.** Sex-specific maps and how markers are merged into the final, merged linkage map are presented. Individual linkage maps produced from each of the two fathers are also presented, as well as positions for which a distance was estimated below an LOD of 3.0 for efficient merging. Star symbol denotes markers designed from the same scaffold that appear disrupted by one or more additional scaffolds in the merged linkage map, due to lack of shared informative markers in merging sex-specific maps: https://doi.org/10.1371/journal.pone.0102089.s001

Additional File C2 (Figure S2) - **Conservation of gene order within the largest 50 northern pike genomic scaffolds relative to *Gasterosteus aculeatus*, *Oryzias latipes*, and *Danio rerio*.** Reciprocal best hit results (≥50% of total length, ≤1e-5) and mapped scaffold position plotted. Average position ((Start position + End position)/2) taken to represent a point position for each mapped transcript: https://doi.org/10.1371/journal.pone.0102089.s002

Additional File C3 (Table S1) - **Contribution of major Transposable Element categories to the *Esox lucius* genome.** Produced by parsing results from Repeat masker output; variation between the direct Repeat masker output (18.1% of genome) and the sum of the annotated categories is likely due to overlap in the masked elements: https://doi.org/10.1371/journal.pone.0102089.s003

Additional File C4 (Table S2) - **Trinity-assembly FPKM results for 13 *Esox lucius* tissues.** All values submitted to NCBI GEO profile database. https://doi.org/10.1371/journal.pone.0102089.s004

Additional File C5 (Table S3) - **MAKER2-assembly FPKM results for 13 *Esox lucius* tissues:** https://doi.org/10.1371/journal.pone.0102089.s005

Additional File C6 (Table S4) - **Top 10 expressed specialized transcripts from each tissue in *Esox lucius*.** Sorted by highest FPKM for transcripts >3 standard deviations above average across all tissues; based on results from 38K Trinity-assembled transcriptome: https://doi.org/10.1371/journal.pone.0102089.s006

Additional File C7 (Table S5) - **Pathways Identified in duplicated transcripts.** Significant DAVID results (sorted by p-value) for duplicated transcripts between 77–89% similarity in northern pike. Gene_Ontology terms presented from the molecular function (GOTERM_MF_FAT) and biological process (GOTERM_BP_FAT) categories: https://doi.org/10.1371/journal.pone.0102089.s007

Additional File C8 (Table S6) - **Comparative synteny between *Esox lucius* and three model organisms.** Reciprocal blast results (≥50% of total length, ≤1e-5) between scaffold-linked *E. lucius* transcripts and published genome assemblies of A) *Gasterosteus aculeatus* (gasAcu1 – Feb 2006), B) *Oryzias latipes*(oryLat2 – Oct 2005) and C) *Danio rerio* (ZV9 - Jul 2010). The table provides numerical results used to produce Figure 2.6: https://doi.org/10.1371/journal.pone.0102089.s008

Additional File C9 (Table S7) - **Microsatellite marker information and primer sequences used in construction of first-generation linkage map.** Includes relevant information

regarding the location of the microsatellite marker within the scaffolds, the primers used to amplify the repeats, and the results from tandem repeats finder representing information on the repeats, based on the EST/scaffold sequence used to design the primer pairs targeting the repeats:
https://doi.org/10.1371/journal.pone.0102089.s009

Additional File C10 (File S1) - **19K MAKER2-assembled transcriptome in FASTA format:**
https://doi.org/10.1371/journal.pone.0102089.s010

Appendix D

Supplementary Files, Chapter 4

Additional File D1 (Figure S1) - ***Salmo salar* and *Esox lucius* LastZ comparisons**

The LastZ dotplots between the repeat masked Atlantic salmon and northern pike chromosomes for major syntenic blocks. Gene content was plotted as number of genes (annotation release 101 NCBI) per 100,000bp while Repeat content was plotted as percent-masked bases in non-overlapping windows of 100,000bp. Tick marks on the x and y-axis of the LastZ plots refer to boundaries of scaffolds. Circle on y-axis refers to approximate centromere position, while horizontal lines refer to the boundaries between chromosome arms, with thick dashed lines referencing boundary between p and q arms and light dashed lines between a and b pieces of the q-arm.