# An evaluation of the role of adaptation in salmon evolution using genome based approaches

Marine S.O. Brieuc

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Kerry-Ann Naish, Chair

Steven Roberts

James Seeb

Program Authorized to Offer Degree:

School of Aquatic and Fishery Sciences

University of Washington

**Abstract**

# An evaluation of the role of adaptation in salmon evolution using genome based approaches

Marine S.O. Brieuc

Chairperson of the Supervisory Committee:
Kerry-Ann Naish
School of Aquatic and Fishery Sciences

Studying the results of selection may provide insights into the extent of adaptation, processes affecting population divergence, and gene diversity. Here, the role of adaptation in salmon evolution was evaluated at different taxonomic levels using genome based approaches. The first part of this thesis was aimed at developing a bioinformatic methodology to detect genes under selection on a large scale in non-model species. In such species, coding sequences can be incomplete because of limited genomic resources. However, these sequences are information rich, and can be used to estimate neutral versus non neutral divergence across species. Incomplete DNA sequences can complicate estimates of non-neutral divergence based on comparisons between synonymous ($d_S$) and non-synonymous ($d_N$) nucleotide substitutions, commonly used to study selection between species. The first chapter describes a series of steps that can be used to examine positive selection on a large scale between non model species using partial sequences. The

methodology is described for six species of salmonids, where approaches are complicated by the fact that a whole duplication event occurred in the lineage leading to these species. Therefore, challenges associated with duplicated genomes, specifically the separation of orthologs from paralogs, were also addressed. We found that multi-way BLAST optimized the number of alignments between partial coding sequences. We recommend that reading frames should be manually detected after alignment with sequences in Genbank using the BLASTX program. Finally, phylogenetic approaches were determined to be suitable to separate orthologs from paralogs in duplicated genomes.

The second part of the thesis was aimed at conducting a genome-wide assessment of the role of adaptive evolution of Chinook salmon in the Columbia River in the Pacific Northwest of the USA. The first step involved the construction of a dense linkage map for Chinook salmon, thus providing the necessary resources for a genome-wide analysis in wild populations (Chapter 2). We mapped 7146 RAD loci on the 34 chromosomes of Chinook salmon, spanning 4163cM. All the chromosome arms were identified through centromere mapping. Placement of 799 duplicated loci revealed that they were preferentially distributed on distal regions of eight pairs of chromosome arms. This result suggests that homeologs diverged at different rates following whole genome duplication. Our results supported near complete interference during recombination for Chinook salmon, and confirmed previously identified homologies between Chinook salmon and rainbow trout. In the third chapter, we aimed to determine the role of adaptive divergence in the evolution of Chinook salmon in the Columbia River Basin. A population survey of divergence was conducted using 14105 RAD markers in eleven populations in the Columbia River Basin,

representative of the three main lineages identified in previous studies. Our results supported the hypothesis of colonization of the Columbia River Basin from two main refugia following the last glaciation event. We identified 301 outlier loci that did not conform to neutral evolution, consistent with adaptive divergence. Of these, 148 and 153 were associated with the pre- and post- glaciation divergence respectively. Using the linkage map created in the second chapter, we identified chromosomal regions of high divergence, most of which were located in distal regions from the centromere. Although some regions of elevated divergence were observed in common between lineages, many appeared to be specific to pre- or post-glaciation divergence. Finally we investigated whether we could find molecular evidence supporting observations of parallel evolution in a phenotypic trait across populations, adult return timing. Random forest analyses, a regression-based approach, detected some loci that predicted run timing, specifically Spring and Fall return timing, two of which mapped to the same position on the linkage map. In this chapter, beyond improving our understanding of Chinook salmon evolution, we have demonstrated the usefulness of dense linkage maps in identifying regions of the genome that may have been involved in adaptive evolution.

The research presented in this thesis will facilitate the study of adaptive divergence between non-model species. Novel and extensive genomic resources for Chinook salmon have also been developed. These resources have provided insights into chromosome evolution following whole genome duplication, and have greatly contributed to the understanding of adaptive evolution of populations of Chinook salmon in the Columbia River Basin.

## Contents

## List of figures

## List of tables

## Preface

Maintaining the long-term adaptability of threatened or endangered species is one of the main goals of conservation approaches in rapidly changing environments. Increasing levels of animal and plant extinctions have been reported in recent years (Myers & Knoll 2001). Predictions for the future diversity across major taxonomic groupings are alarming; 33-39% of the species assessed by the IUCN Red List (at least 18,351 species) are threatened with extinction (IUCN 2010). Many factors may be responsible for this threat, including habitat destruction, over-exploitation, invasive alien species, pollution, disease or climate change. Conservation actions, such as supportive breeding, translocation or design of reserves, have potential to increase population abundance, but these measures they can also pose a risk for natural populations. For example, supportive breeding may lead to domestication (Ford 2002a), a decrease of effective size (Ryman *et al.* 1995), or deleterious genetic effects (Waples 1991). There is therefore an urgent need to improve the effectiveness of conservation measures (IUCN 2004). These conservation measures should aim at maintaining genetic diversity (Laikre *et al.* 2008; Laikre *et al.* 2010), especially at genomic regions implicated in fitness traits, and thus maintain the potential for populations to evolve (Frankel 1974). Therefore, it is important to identify the genetic diversity underlying such regions.

Selection is the only evolutionary force that directly allows adaptation and is therefore a key component of population longevity and capacity of populations to respond to change (reviewed in Ford 2002b; Yang & Bielawski 2000; Nielsen 2005; Allendorf *et al.* 2010). However, genetic drift, mutation and migration are involved in adaptation as they contribute to the genetic variation on which selection will be able to act (Barton & Partridge 2000; Rieseberg & Burke 2001; Morjan & Rieseberg 2004; Nosil *et al.* 2009). In this study we are interested in characterizing the relative importance of adaptation in shaping the genetic variation and species and population divergence. Signatures of natural selection can be detected on various time scales (Hohenlohe *et al.* 2010) using a variety of markers and models (reviewed in Ford 2002b; Nielsen 2005; Hancock & Di Rienzo 2008; Kelley & Swanson 2008). Detecting selection between and within populations or species

could therefore provide key insights about past evolution and adaptation as well as capacity to adapt to future environmental challenges (Naish & Hard 2008).

Detection of positive selection has usually been limited to a few candidate genes or loci in non-model organisms (Beaumont & Balding 2004; O'Malley *et al.* 2007; Boulding *et al.* 2008). However, genome-wide studies have provided insight into the adaptive evolution within and between species, and have identified specific genes involved in these processes in model species (*e.g.* Clark *et al.* 2003; Clark & Swanson 2005; Bakewell *et al.* 2007). It is now possible to conduct similar studies in non-model organisms. The rapid expansion of next-generation sequencing technologies (Hudson 2008) has resulted in significant potential to initiate such studies (Shendure & Ji 2008). Several approaches produce millions of sequences, even for species with little or no genetic information available (Davey *et al.* 2011).

The study of evolutionary processes in salmon species is of considerable interest in conservation and management. Salmonids are of great economic (FAO 2011) and cultural value (Committee on Protection and Management of Pacific Northwest Anadromous Salmonids 1996). Unfortunately many populations of salmonids have faced significant declines in the last few decades and many salmon species have populations listed under the Endangered Species Act (NOAA 2013). This decline has motivated much research aimed at gaining insights into the significance of salmonid population diversity. Interest in detecting selection in salmonids is increasing (*e.g.* Consuegra & Johnston 2008; Tonteri *et al.* 2010) and partial sequences for these genera have become widely available on public databases (*e.g.* NCBI website, www.ncbi.nlm.nih.gov; cGrasp database, http://web.uvic.ca/grasp). Interestingly, salmonids are descended from a single ancestor that underwent a duplication event 25 to 100 million years ago (Allendorf & Thorgaard 1984); salmon species are residual tetraploids. Duplicated genes provide genetic material on which selection can act: previous studies have shown evidence for selection on paralogous genes (*e.g.* Emes & Yang 2008; Han *et al.* 2009). Moreover incorrect conclusions may be reached if genes are aligned and analyzed together as orthologous genes when they are in fact paralogous genes. The

duplicated genome of salmonids permits an exploration of selection acting on paralogous genes as well as approaches for separating paralogous and orthologous genes.

Chinook salmon (*Oncorhynchus tshawytscha*) is of considerable economic, cultural and conservation interest. This species exhibit a wide variety of life history trait as a result of local adaptation (Quinn 2005). Interestingly populations in the Columbia River Basin are hypothesized to be derived from two main refugia following the last glaciation event (Beacham *et al.* 2006; Waples *et al.* 2008). Moreover there has been evidence for parallel evolution in Chinook salmon where run-timing has evolved independently but in a similar way in several different lineages (Waples et al. 2004). We are therefore poised to investigate the effects of selection over different evolutionary scales, as well as evidence for parallel evolution at the molecular level.

The overall aim of this study is to identify genetic variation that has played an important role in adaptation of salmon species. Adaptive variation has rarely been characterized in salmon and understanding its role relative to other evolutionary forces, such as migration and drift, can provide key insights about salmon evolution and their capacity to adapt to future challenges. Here, selection will be studied at different time scales. First, attempts will be made to study adaptive evolution that has played a role in the divergence of salmonids. A bioinformatic approach to studying selection between non-model species is underdeveloped, but is necessary for advancing this field. Using salmon species as an example, we will develop tools to facilitate the study of selection in non-model species. Second, adaptation at the species level, specifically Chinook salmon (*Oncorhynchus tshawytscha*) in the Columbia River basin, USA, will be examined using genome wide approaches that will be developed and described here. The products of this research will facilitate study of adaptive divergence in other non-model species and provide insights about evolution of Chinook salmon.

## References


Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. B.J. T). Plenum Press, New York.

Bakewell MA, Shi P, Zhang JZ (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7489-7494.

Barton N, Partridge L (2000) Limits to natural selection. *Bioessays* **22**, 1075-1084.

Beacham TD, Jonsen KL, Supernault J*, et al.* (2006) Pacific Rim population structure of Chinook salmon as determined from microsatellite analysis. *Transactions of the American Fisheries Society* **135**, 1604-1621.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Boulding EG, Culling M, Glebe B, *et al.* (2008) Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity* **101**, 381-391.

Clark AG, Glanowski S, Nielsen R, *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960-1963.

Clark NL, Swanson WJ (2005) Pervasive adaptive evolution in primate seminal proteins. *Plos Genetics* **1**, 335-342.

Committee on Protection and Management of Pacific Northwest A, Salmonids (1996) Upstream: salmon and society in the Pacific Northwest. *Upstream: salmon and society in the Pacific Northwest*, i.

Consuegra S, Johnston IA (2008) Effect of natural selection on the duplicated lysyl oxidase gene in Atlantic salmon. *Genetica* **134**, 325-334.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution* **26**, 298-306.

Emes RD, Yang Z (2008) Duplicated Paralogous Genes Subject to Positive Selection in the Genome of Trypanosoma brucei. *Plos One* **3**.

FAO (2011) Fisheries and Aquaculture Information and Statistics Service.

Ford MJ (2002a) Applications of selective neutrality tests to molecular ecology. Molecular Ecology 11, 1245-1262.

Ford MJ (2002b) Selection in captivity during supportive breeding may reduce fitness in the wild. *Conservation Biology* **16**, 815-825.

Frankel OH (1974) Genetic Conservation – Our evolutionary responsibility. *Genetics* **78**, 53-65.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Research* **19**, 859-867.

Hancock AM, Di Rienzo A (2008) Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. *Annual Review of Anthropology* **37**, 197-217.

Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences* **171**, 1059-1071.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3-17.

IUCN (2004) IUCN Red list of threatened species. A global species assessment. IUCN species survival commission, Gland, Switzerland, and Cambridge, UK.

Kelley JL, Swanson WJ (2008) Positive selection in the human genome: From genome scans to biological significance. *Annual Review of Genomics and Human Genetics* **9**, 143-160.

Laikre L, Allendorf FW, Aroner LC, *et al.* (2010) Neglect of Genetic Diversity in Implementation of the Convention on Biological Diversity. *Conservation Biology* **24**, 86-88.

Laikre L, Larsson LC, Palme A, *et al.* (2008) Potentials for monitoring gene level biodiversity: using Sweden as an example. *Biodiversity and Conservation* **17**, 893-910.

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology* **13**, 1341-1356.

Myers N, Knoll AH (2001) The biotic crisis and the future of evolution. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5389-5392.

Naish KA, Hard JJ (2008) Bridging the gap between the genotype and the phenotype: linking genetic variation, selection and adaptation in fishes. *Fish and Fisheries* **9**, 396-422.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

NOAA (2013) Salmon & Steelhead Listings - http://www.nwr.noaa.gov/protected_species/salmon_steelhead/salmon_and_steelhead_listings/salmon_and_steelhead_listings.html - accessed June 26, 2013

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375-402.

O'Malley KG, Camara MD, Banks MA (2007) Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **16**, 4930-4941.

Quinn TP (2005) *The behavior and ecology of Pacific salmon and trout* American Fisheries Society, Bethesda, Md.

Quinn TP, Unwin MJ, Kinnison MT (2000) Evolution of temporal isolation in the wild: Genetic divergence in timing of migration and breeding by introduced chinook salmon populations. *Evolution* **54**, 1372-1385.

Rieseberg LH, Burke JM (2001) The biological reality of species: gene flow, selection, and collective evolution. *Taxon* **50**, 47-67.

Ryman N, Jorde PE, Laikre L (1995) Supportive breeding and variance effective population size. *Conservation Biology* **9**, 1619-1628.

Shendure J, Ji HL (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145.

Tonteri A, Vasemagi A, Lumme J, Primmer CR (2010) Beyond MHC: signals of elevated selection pressure on Atlantic salmon (*Salmo salar*) immune-relevant loci. *Molecular Ecology* **19**, 1273-1282.

Waples RS (1991) Genetic interactions between hatchery and wild salmonids – Lessons from the Pacific-Northwest. *Canadian Journal of Fisheries and Aquatic Sciences* **48**, 124-133.

Waples RS, Teel DJ, Myers JM, Marshall AR (2004) Life-history divergence in Chinook salmon: Historic contingency and parallel evolution. *Evolution* **58**, 386-403.

Waples RS, Pess GR, Beechie T (2008) Evolutionary history of Pacific salmon in dynamic environments. *Evolutionary Applications* **1**, 189-206.

Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation: when evolution repeats itself. *Genetica* **123**, 157-170.

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**, 496-503.

# Chapter 1 – Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources[1]

Marine S.O. Brieuc and Kerry A. Naish

*University of Washington, School of Aquatic and Fishery Sciences, Seattle, WA 98195 USA*

## Abstract

Studying the actions of selection provides insight into adaptation, population divergence, and gene function. Next-generation sequencing produces large amounts of partial sequences, potentially facilitating efforts to detect signatures of selection based on comparisons between synonymous ($d_S$) and non-synonymous ($d_N$) substitutions, and Single Nucleotide Polymorphism assays placed in selected genes would improve the ability to study adaptation in population surveys. However, sequences generated by these technologies are typically short. In non-model organisms that are a focus of evolutionary studies, the lack of a reference genome that facilitates assembly of short sequences has limited surveys of positive selection in large numbers of genes. Here, we describe a series of steps to facilitate these surveys. We provide PERL scripts to assist data analysis, and describe the use of commonly available programs. We demonstrate these approaches in six salmon species, which have partially duplicated genomes. We recommend using multi-way BLAST to optimize the number of alignments between partial coding sequences. Reading frames should be manually detected after alignment with sequences in Genbank using the BLASTX program. We encourage the use of a phylogenetic approach to separate orthologs from paralogs in duplicated genomes. Simple simulations on a gene known to have

[1] This chapter has been published as Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources, Brieuc MSO and Naish KA, Molecular Ecology Resources **11** (Suppl. 1), 172–183, Copyright © 2011 Blackwell Publishing Ltd

undergone selection in salmon species, transferrin, showed that the ability to detect selection in short sequences (less than 600bp) depended on the proportion of codons under selection (1 to 2%) within that sequence. This relationship was less relevant in longer sequences. In this exploratory study, we detected 11 genes showing evidence of positive selection.

## Introduction

Selection is an evolutionary process that can provide key insights about adaptation, population divergence, and gene function (Nielsen 2005). Understanding the genetic basis underlying evolution in species-specific traits or phenotypes is of great interest and has been a recent major focus in human evolutionary biology (Kelley & Swanson 2008). The extent to which natural selection has shaped molecular evolution is still unknown for most species, but there is evidence that it has played a key role in recent molecular evolution across a number of species (Bustamante *et al.* 2005; Shen *et al.* 2010). Identifying genes targeted by natural selection can greatly improve our knowledge of the role of adaptation in species evolution.

The signatures of natural selection can be detected using a variety of markers and models (reviewed in Ford 2002; Nielsen 2005; Hancock & Di Rienzo 2008; Kelley & Swanson 2008). Approaches to detect signatures of selection can be divided in two main categories: polymorphism-based methods and divergence-based methods (Kelley & Swanson 2008). Polymorphism-based approaches rely on the detection of selective sweeps (Kelley & Swanson 2008). Recent selection events will result in an increase in frequency of the beneficial allele, as well as a reduced heterozygosity around the beneficial allele (hitchhiking). If selection does not act equally on all populations within a species, differentiation at the selected locus and the flanking haplotype will increase between the populations. A signature of positive selection can be interpreted if divergence at the haplotype is greater than divergence expected under neutral expectations (Akey *et al.* 2002). The ability to detect a selective sweep will depend on the number of generations after the selection event occurred, the effective size of the population, and the distance

between a marker and the selected site (Nielsen 2005). As new mutations and recombination occur, the flanking haplotype unique to the locus under selection will be shorter (Nielsen 2005). Polymorphism-based approaches can be applied using various markers such as microsatellites, Single Nucleotide Polymorphisms (SNPs) or haplotypes. Divergence-based approaches rely on comparisons of protein-coding sequences (Nielsen 2005). Nucleotide substitution profiles can be used to infer selection by comparing the rate of non-synonymous substitutions ($d_N$) to the rate of synonymous substitutions ($d_S$). If selection has not acted on a protein coding region, $d_N$ and $d_S$ will be the same, and $d_N/d_S$=1. Purifying (negative) selection will act against non synonymous sites and $d_N/d_S$< 1, while diversifying (positive) selection will result in $d_N/d_S$ >1. Divergence-based approaches can therefore directly identify the genes under selection. Demographic factors, recombination rates and mutation rates do not have a major impact on the power to detect genes under selection using divergence-base approaches (Nielsen *et al.* 2009). However, this approach relies on higher degrees of divergence between organisms compared to polymorphism based methods (Anisimova *et al.* 2001).

The use of divergence-based approaches was originally limited to the detection of positive selection for a few candidate genes or groups of genes (Li 1997; Swanson *et al.* 2003; Yang 1998). However, large-scale or genome-wide studies seeking evidence for positive selection have become common with the advances in sequencing techniques, gene discovery and data processing (Endo *et al.* 1996), particularly within humans and model organisms and between closely related model species (Anisimova *et al.* 2007; Clark *et al.* 2003; Ge *et al.* 2008; Kitano *et al.* 2004). Larger scale studies have provided insight into the adaptive evolution within and between species, and have identified specific genes involved in these processes (e.g. Clark *et al.* 2003; Clark & Swanson 2005).

Large-scale detection of genes that have responded to selection is still very limited in non-model species, where DNA sequence coverage is incomplete and gene identity is often unknown. However this information is needed to better understand evolution and adaptation in non-model organisms (Dalziel et al. 2009; Naish & Hard 2008). The rapid expansion of next-generation sequencing technologies (Hudson 2008) has resulted in

significant amounts of data, which are often suitable for comparative genomic studies of non-model organisms (e.g. Kunstner *et al.* 2010). However, read lengths obtained with these sequencing techniques are usually shorter than traditional sequencing (Wheat 2010), and require *de-novo* assemblies or assemblies against a reference genome. The rapid expansion in the number of partial coding regions generated in evolutionary and ecological studies provide an exciting opportunity for selection studies in non-model species.

Detecting evidence for selection using divergence-based approaches relies on sequence alignment and identification of coding regions. There are several limitations associated with analyzing large numbers of partial sequences for evidence of selection in non-model species. First, alignments of partial sequences are usually overlapping but incomplete. Researchers therefore need to view each alignment and trim sequences to matching portions in all compared species. This task can rapidly become onerous. Second, the codons from partial sequences are not readily identifiable because the majority of non-model organisms lack a reference genome. It is unknown how well the algorithms implemented in open reading frame (ORF) detectors that are available in many software packages perform with partial coding sequences. Third, a fraction of genes may be duplicated; incorrect conclusions may be reached if paralogous and orthologous genes are aligned and analyzed together. Therefore, it is necessary to use analytical approaches to separate duplicated genes.

Single Nucleotide Polymorphisms (SNPs) are the most abundant form of DNA polymorphism in the genome (Collins *et al.* 1998). In non-model species SNPs have been showed to be suitable for population structure (*e.g.* Freamo *et al.* 2011, Miller *et al.* 2010), genetic stock identification and mixed stock analysis (Hess *et al.* 2011, Templin *et al.* 2011), as well as parentage analysis (Anderson & Garza 2006, Hauser *et al.* 2011) among others. The high abundance of SNPs also allows detection of natural selection and adaptive divergence using genome-wide scans (Akey *et al.* 2002; Willing *et al.* 2010) or smaller surveys of individual loci (Narum *et al.* 2008). Such studies rely on polymorphism or outlier approaches to detect evidence of selection, where a SNP under positive selection will be characterized by a higher $F_{ST}$ than expected under the neutral model (Beaumont & Balding

2004). Low polymorphism of individual SNP loci is a main limitation of SNP loci in these types of studies (Morin *et al.* 2009; Nielsen 2005). SNPs are for the main part biallelic markers, therefore a greater number of SNPs are required to obtain the resolution reached with more polymorphic markers (Morin *et al.* 2009). A way to increase the power of SNPs in population genetics generally would be to use linked loci to infer haplotypes (Morin *et al.* 2009). If these haplotypes span genes under selection, then the power of SNP markers for studying selection and adaptive divergence would increase. Here, we suggest surveying partial sequences of transcriptomes or exons that are typically generated in SNP discovery projects for evidence of directional selection. Such sequences can then be targeted for SNP development, either as single loci or as haplotype markers. Developing SNP markers in these regions will substantially improve the ability to study the role of selection in the evolution of non-model species.

In this study, we describe a series of steps to detect genes under selection on a large scale in non-model species, using partial sequence data generated using next generation sequencing. We provide PERL scripts to assist with the analysis of the data, and describe analytical approaches available in commonly available programs.

We demonstrate the steps using sequence data from six salmon species (Atlantic salmon in the genus *Salmo*; Pacific salmon in the genus *Oncorhynchus* and charr, in the genus *Salvelinus*). These species were selected because partial sequences for these genera have become widely available on public databases (e.g. NCBI website, www.ncbi.nlm.nih.gov; cGrasp database, http://web.uvic.ca/grasp) and interest in detecting selection in salmonids is increasing (e.g Consuegra & Johnston 2008; Lai *et al.* submitted; Tonteri *et al.* 2010). Moreover, salmonids are descended from a single ancestor that underwent a duplication event 25 to 100 million years ago (Allendorf & Thorgaard 1984); salmon species are residual tetraploids. Their duplicated genome permits an exploration of approaches for separating paralogous and orthologous genes. Finally, we examine the relationship between DNA sequence length and ability to detect selection by performing a simulation using a gene that is known to have undergone selection in salmon species (Ford 2001; Ford *et al.* 1999).

## Material and Methods

### Sequence data collection and assembly

We collected Expressed Sequence Tags (ESTs) and cDNA sequences for six salmon species. Sequence data for *Oncorhynchus keta* were produced and assembled in 54422 contigs using 99% identity (Seeb *et al.*, in press). Publically available sequence data for *Salmo salar*, *O. mykiss*, *O. tshawytscha* and *O. nerka*, assembled in 119912, 69358, 9185 and 6598 contigs respectively (Koop *et al.* 2008) using PHRAP (Green 1996) (minimum overlap of 100 base pairs and 99% identity), was downloaded from the cGRASP database (http://web.uvic.ca/grasp). *Salvelinus fontinalis* sequence data was obtained by combining the 10051 ESTs from cGRASP (http://web.uvic.ca/grasp) with the 10030 ESTs available on the NCBI website (16 April 2008); 4273 contigs were assembled using PHRAP (minimum overlap of 100 base pairs and 99% identity, followed by minimum overlap of 300 and 96% identity

### Alignment of partial sequences (coding regions) common to all species

We used the BLAST (Basic Local Alignment Search Tool; Altschul *et al.* 1990) program to detect sequences common to all six species using an e-value cut-off of $10^{-30}$ for a positive match. A commonly implemented approach with complete coding regions is the reciprocal best hit method (e.g. Bakewell *et al.* 2007). However, incomplete coverage will limit the power of this approach. The common sequences detected with BLAST will differ and will depend on the species that is used in the query (Figure 1.1). To identify the optimal BLAST search strategy, we executed a one-way blast of the *S. fontinalis* sequences against the other five species, and compared this approach to a multi-way BLAST (blast of sequences from each species against every other species). We then compared the length and the number of contigs we obtained with each strategy.

We automated the BLAST search using a PERL script (Goetz G., University of Milwaukee). Sequences were then aligned using CLUSTALW version 2.0 (Larkin *et al.* 2007). Alignments were automated using a PERL script (modified from Bakewel M., University of Michigan).

We developed a PERL script to trim aligned sequences in order to retain the regions common to all six species. All scripts are available as supplementary material, with permission from the original coders [S1.1-S1.3].

## Designation of codons in partial sequences

The coding region of a gene begins with a start codon and ends with a stop codon. Open Reading Frame (ORF) detectors available in sequence analysis software packages use algorithms that will recognize putative ORFs after identification of either a start or stop codon, or both. In partial sequences where the start and stop codons might be missing, it is difficult to detect the codons. This difficulty is exacerbated when the sequenced species has no reference genome or a coding region database. Using software that reliably detects reading frames for partial coding sequences could greatly reduce data processing. To test the ability of algorithms implemented in ORF detectors to identify reading frames in partial coding sequences, we compared the results obtained with CLC Genomics Workbench 3 software (CLC bio, Denmark) with those obtained using BLASTX (BLAST of nucleotide sequence against the non-redundant protein sequences (nr) database; Altschul *et al.* 1997). BLASTX is a reliable tool for identifying coding regions by comparing similarities between a nucleotide sequence and a known protein sequence (Gish & States 1993). We manually inspected all the BLASTX hits. A positive hit had an e-value<$10^{-10}$ and a percentage of identity greater than 85% over the coding region surveyed.

## Detecting positive selection in partial coding sequences

Evidence for positive selection at each coding sequence was detected using the maximum-likelihood method (Nielsen & Yang 1998; Yang *et al.* 2000) implemented in the program CODEML of PAML (version 4.0; Yang 1997, 2007). Models that estimate an average $d_N/d_S$ ratio across the whole gene have been shown to be too stringent, since only a few sites will be under selection in the majority of cases (Yang *et al.* 2000). Implementation of whole gene approaches will also have less power in partial coding sequences. CODEML implements statistical models that allow heterogeneous ω ($d_N/d_S$ ratios) among sites (codons), which is particularly appropriate and more powerful for detecting evidence for selection in partial

sequences. The maximum-likelihood method is based on the comparison of likelihood for neutral and non-neutral models with different ω classes and ω distributions among sites (Models 7, 8 and 8a in PAML) using a likelihood ratio test. We compared the likelihood of Model 7 (M7; neutral) vs Model 8 (M8; non neutral), and Model 8a (M8a; neutral) vs Model 8 (non neutral) (Yang *et al.* 2000). Model M7 permits the ω for each codon to be derived from 10 classes that are drawn from the beta distribution, which in turn describes the distribution of ω among codons. All of the classes have to be smaller or equal to 1 (Neutral) Yang *et al.* (2000). In model M8, ω can belong to the same 10 classes, but a proportion $p_1$ of sites can now have $\omega_1 > 1$ where $\omega_1$ is constant (non neutral). Model M8a, implemented by Swanson *et al.* (2003), differs from M8 in that the $11^{th}$ class of codon can take the unique value ω, where ω is always equal to 1 (Neutral). The latter test has more power, since it reduces the degree of freedom from 2 to 1 (Findlay *et al.* 2008). Models M7 and M8 were replicated with different starting omega values to ensure convergence of the likelihood. If the likelihood ratio test was not significant for the comparison M7 versus M8, but was significant for M8a versus M8, comparisons with the highest likelihood for the neutral model (M7 or M8a) were favored. PAML requires a phylogeny for the analysis. We used the Salmoninae phylogeny described in Crespi and Fulton (2004).

Performing codon-specific tests for selection generates a large number of statistical tests that result in Type-1 error. Previous studies have shown that genes with whole-gene $\omega > 0.5$ or $d_N > 0.05$ are more likely to have specific sites under selection (Clark & Swanson 2005). It is possible, therefore, to restrict analyses to a subset of genes with complete coding regions in order to minimize the number of tests (Findlay *et al.* 2008). To test whether we could restrict the analysis to a subset of genes using conditions based on whole-region ω or $d_N$, we estimated $d_N$, $d_S$ and the $d_N/d_S$ ratio for each available coding region, most of which were partial coding regions, using Model 0 in PAML developed by Goldman and Yang (1994). We then compared those values for genes showing sign of positive selection in order to determine whether we could restrict our analysis to a subset of sequences (determine whether there was an ω or $d_N$ value above which the probability of a locus under selection was higher).

To test the relationship between sequence length and ability to detect selection within our data, we performed simulations using the sequence data from transferrin, which has been shown to be under positive selection in salmonids (Ford, 2001). The alignment comprised five species (*S. salar, S. fontinalis, O. mykiss, O. tshawytscha and O. nerka*) and was 2016 base pairs long. Individual sites under selection were initially identified using the Bayes Empirical Bayes (BEB) analysis (Yang *et al.* 2005) implemented in CODEML. To test the ability to detect selection in partial coding sequences using the PAML procedures implemented here, we randomly extracted 10 regions of 50, 100, 150, 200, 300, 400 and 500 codons in the aligned transferrin sequences and applied the tests of selection. We also extracted two additional aligned sequences of 50 codons in regions where there was no site under selection, to test for false discovery rate.

**Duplicated loci: separating orthologs and paralogs in partial alignments**

Salmon are residual tetraploids, and therefore we expected a significant number of duplicated genes. Higher divergence between paralogs, as well as possible presence of positive selection between paralogs, could affect our ability to detect selection between orthologs and lead to incorrect interpretations. We therefore used a *post hoc* phylogenetic approach to verify whether genes showing evidence of positive selection were true orthologs. We used BLAST to identify matching sequences of all six salmon species and unrelated fish species that could serve as outgroups (Evalue ≤ $10^{-30}$ ; Medaka (*Oryzias latipes*) or Northern Pike (*Esox lucius*) were favored when available). We then derived a bootstrapped Neighbor Joining phylogenetic tree of the sequences using the program MEGA 4 (Tamura *et al.* 2007). We tested for evidence of selection between the true orthologs as well as between paralogs. This step allowed us to determine whether the initial detection of selection was due to selection between paralogs or true orthologs, and whether selection between orthologs was masked by the presence of a paralog.

**Testing the reliability of sequence assembly in separating orthologs and paralogs**

A number of studies use different thresholds for sequence assembly (0.97 similarity (Renaut *et al.* 2010), 0.95 (Barbazuk *et al.* 2007), 0.90 (Meyer *et al.* 2009)). To test whether

these thresholds are appropriate to separate orthologs and paralogs within salmon species, we derived bootstrapped Neighbor Joining trees for randomly selected genes that appeared neutral (no sign of selection in tests using PAML) and examined the percentage of similarity between orthologs and between paralogs.

### Detecting genes under lineage-specific selection

We attempted to detect lineage-specific selection in genes that showed evidence for positive selection (likelihood was at a maximum in Model 8) using CODEML (Yang 1997). A likelihood ratio test was performed by comparing M8 to M0, where ω was now constant across a single lineage for M0, but each lineage was permitted to have a different ω value (Goldman & Yang 1994) (model=1, NSsite=0). Evidence for positive selection in a specific lineage was assumed if ω>1 for that lineage, and if the likelihood was higher for M0 than M8.

## Results

### Alignments

We found that it was important to conduct multi-way BLAST searches to attain the maximum number of alignments in partial sequences. We identified 654 orthologous sequences in all six species using the one-way BLAST, 287 of which had an overlapping region conserved between all six species and an overall identity greater than 70%. In contrast, we identified 18,523 common sequences using the multi-way BLAST. Approximately 5090 of these contigs had an overlapping region that was conserved in all six species, with an overall identity greater than 70%. This threshold was the same as the percent similarity implemented by Koop *et al.* (2008) for the reconstruction of the phylogeny of Salmonidae. The contig lengths ranged from 27 base pairs to 1421 base pairs for the one-way BLAST and 22 base pairs to 1617 base pairs for the multi-way BLAST, with an average of 513 (± 269 SD) and 489 (± 239 SD) base pairs respectively (Figure 1.2). The two means were not significantly different (Welch t-test, $p = 0.15$).

## Detection of coding regions

We found that the best way to detect reading frames in partial sequences was to rely on alignments with coding regions of Salmoninae or other species in Genbank using BLASTX. Detecting the coding regions was labor intensive. A subset of 434 contigs randomly chosen from the multi-way BLAST was therefore used to develop and test our methodology.
Of the 434 contigs surveyed, reading frames were detected for 273 contigs using CLC Genomic Workbench, and for 222 contigs using BLASTX. The BLASTX search identified 29 complete coding regions within the 222 contigs. The reading frames for 35 of the contigs identified in CLC did not match the BLASTX results. In addition, CLC identified 31 longer reading frames than BLASTX; that is, CLC identified non coding regions as coding. Some coding regions that were not represented in the NCBI protein database but potentially correctly identified by CLC might be ignored when relying on BLASTX to identify coding regions. However, the approach we used here was conservative, and might be expected to reduce the risk for false positive results.

## Detection of positive selection in the partial coding region of the transferrin gene

The between-species alignment across tranferrin (2016 base pairs from Ford 2001) was almost complete. The tests of selection were significant (pvalue = 2.05E-11 for M7-M8 and 2.76E-12 for M8a-M8). Twenty sites were identified as being under positive selection (P>95%).

In the entire simulation there were four fragments with no site under selection. Three were 50 codons long and one was 100 codons in length. The likelihood ratio test showed that all of these contigs were neutral; in other words, there were no false positive results. Not surprisingly, the power to detect selection increased both as the length and the proportion of codons under selection increased (Table 3). For shorter fragments (less than 200 codons or 600 bp), selection was successfully detected when more than 1 to 2% of the sites were under selection.

## Detection of positive selection

Of the 222 remaining contigs in which protein sequences were identified by BLASTX, 20 contigs were redundant, and 23 contigs had a stop codon within the sequence. Therefore 179 contigs, representing 152 genes, were surveyed for evidence of selection.
The likelihood ratio tests were significant ($p<0.05$) for 10 contigs for the comparisons M7-M8 and 20 for M8a-M8. Ten contigs were in common between the two comparisons. Of the remaining 11 contigs detected by the M8a-M8 comparison, five had a higher likelihood for M8a than M7. It was therefore possible to infer positive selection for 15 contigs in total, representing 14 different genes (Table 1). Only two tests for the comparisons M7-M8 and 3 tests for the M8a-M8 were significant after the strict Bonferroni correction.

We learned that we could not limit the number of genes tested to reduce computation time, using the whole-region ω or $d_N$ criteria. The whole-region $d_N$ and $d_S$ for all 179 contigs are plotted on Figure 1.3. One gene (fibroleukine) only had non synonymous substitutions. Therefore $d_N$ / $d_S$ was undefined and was not plotted in Figure 1.3. Four of the 15 contigs showing significant selection had a $\omega < 0.5$ and/or a $d_N < 0.05$.

## Importance of assembly parameters in partial sequences

We learned that assembly criteria for partial sequences must be very stringent to separate duplicated genes in salmon species. For example, derivation of a phylogenetic tree for a neutral gene (14-3-3 protein beta/alpha-2; Figure 1.4) revealed 99.8% and 100% pairwise similarity respectively within the two Atlantic salmon orthologs, compared to 97.8% pairwise similarity between the paralogs. The similarity between paralogs was higher than has been used for the *S. fontinalis* assembly, where a 96% similarity criterion had been used; therefore, we repeated the assembly with PHRAP (minscore: 100, repeat stringency: 99) and revisited the alignments. We identified 10 contigs in which the *S. fontinalis* sequences assembled differently from the original assembly. None of these contigs demonstrated evidence of selection.

## Separation of orthologs and paralogs using a phylogenetic approach

A phylogenetic approach can successfully separate orthologs and paralogs. Phylogenetic trees were constructed for all 15 contigs that showed evidence of selection. Five comparisons were between true orthologs, 7 comparisons were between paralogs and 3 comparisons remain unresolved using the phylogenetic approach.

Some trees also highlighted evidence of an ancient duplication that preceded the event in salmon (60S ribosomal protein L5). A few genes appeared to have only one copy; there was no sign of duplicated gene in any of the six species for Transaldolase or Nucleoside diphosphate kinase A.

## Positive selection on orthologs and lineage-specific selection

We re-applied the likelihood ratio test to 16 genes (true ortholog comparisons) and 7 comparisons that included orthologs and paralogs when sequences for at least four species were available for each ortholog. A total of 23 tests were executed, 12 of which were significant (Table 2). Two of those tests remained significant after a strict Bonferroni correction. However it is important to note that for 23 tests, we only expected about 2 false negatives with a critical pvalue of 0.10, rather than the 10 identified by the test. The strict Bonferroni correction therefore appears to be too conservative.

Ten ortholog and two paralog comparisons demonstrated significant evidence for selection (Table 2). For four comparisons where paralogs and orthologs were available, only one ortholog was under selection, but there was no sign of selection between paralogs. In all cases, the presence of a paralogous sequence for one or two species in the initial alignment did not affect the power to detect selection between the orthologs. For one gene, myeloid leukemia differentiation protein homologue, the presence of a paralogous sequence in the initial test resulted in a false-positive detection of selection. After further observation 60S ribosomal protein L5 had been incorrectly aligned and the correct alignment did result in a non significant test.

Selection was lineage-specific for five genes. Lineages under selection are described in Figure 1.5.

## Discussion

Here, we have demonstrated approaches that may be taken to detect evidence of positive selection using divergence-based approaches in large datasets comprising partial sequences, and have adapted and developed PERL scripts to automate components of the data analyses. The remaining approaches are dependent on publically-available bioinformatic resources.

We found that identification of aligned sequences was optimized when using a multi-way BLAST compared to one-way BLAST. The increase in the number of contigs obtained was due to two main reasons. First, in larger databases such as the Atlantic salmon database, there were several allelic versions for some genes. Second, the presence of partial coding sequences will lead to discontinuity within a single gene (Figure 1.1). Sequences are of various lengths across species; a single sequence can comprise the complete gene for one species, or two discontinuous sequences might represent that same gene in another species. The direction of the BLAST search affects the length of the overlapping regions detected between the two species. Performing a multi-way BLAST will maximize this length. The best reciprocal hit method is commonly implemented to reduce the effect of overrepresentation of certain genes and to minimize false ortholog alignments. The approach favors pairs of genes that result from a reciprocal match through a bi-directional BLAST. However, this method is not suitable here because partial sequences and discontinuity in overlapping regions complicate the analyses.

Inference of selection for non-model species across large numbers of partial sequences can be very time consuming, because codon alignments for most genes are unknown. We were able to partly overcome this limitation by automating the BLAST search and extracting the alignments using PERL scripts. However, we were not able to limit the study to a subset of genes using criteria dependant on overall $\omega$ or $d_N$ for partial coding sequences.

We recommend relying on BLASTX searches for the detection of reading frames, since commonly implemented algorithms for detecting open reading frames were only correct in a small fraction of cases (about 25% in our study). The power of BLASTX to detect reading frames is limited to the proteins available in the NCBI protein databases. However as advances are made and proteins are discovered, the protein databases will be more complete. The manual extraction of reading frames is one of the limitations of the method we describe here. Programs such as OCPAT (Liu *et al.* 2007) are a first step in the automated extraction of the information about the reading frames. However, this program is limited to the multiple alignments of sequences of gene coding regions from humans and 13 other vertebrate tetrapods only.

We recommend a *post hoc* phylogenetic analysis of sequences because aligned sequences may comprise paralogs as well as orthologs. Observations of a high degree of similarity between paralogs (97.8% for 14-3-3 protein beta/alpha-2) were consistent with findings from Andreassen *et al.* (2009) who find up to 98.7% identity between putative paralogs in Atlantic salmon (*Salmo salar*). Therefore, the thresholds for separating duplicated genes need to be carefully studied in a species of interest; in salmon species, the assembly criteria need to be very stringent. We found that paralog identification cannot be achieved using a simple threshold for alignment. We recommend a *post hoc* analysis, because constructing phylogenies on all sequences will be very time consuming. The inclusion of paralogs in the initial tests for selection does not seem to affect the power to detect selection between orthologs, but might result in false positive results over the whole phylogeny (orthologs and paralogs included). The use of phylogenetic approaches can reveal ancient duplications, and therefore we recommend using the approach even for diploid species.

In this study we were able to demonstrate positive selection for 11 genes, including the transferrin gene, out of 158 genes represented in our database (7%). However, the number of genes tested includes both orthologs and paralogs. Since we performed a *post hoc* test to separate orthologs and paralogs, it is not possible to provide a clear estimate of the two classes in the original data set. The Human-Chimp-Mouse comparison (Clark *et al.* 2003) detected about 20% of genes that exhibited signs of positive selection. Our study is broadly

comparable to the study on five *Streptococcus* species (Anisimova *et al.* 2007), where about 8% of the genes were identified as being under selection. Selective constraints on proteins have been shown to be affected by effective population size and population structure (Wright & Andolfatto 2008). In addition, the power to detect selection has been shown to be affected by sequence length and divergence rate (Anisimova *et al.* 2001). In this study 83.8% of the sequences were partial coding sequences. Failure to detect selection for a partial coding sequence is not necessarily a sign that the gene is not under selection, but rather that the specific sites under selection might not have been sequenced. Use of full-length cDNA inserts or preference for next-generation sequencing that yield longer reads should therefore be favored in order to optimize the length of sequences available in coding regions. The longer sequence will also reduce the assembly errors due to the high degree of similarity between paralogs.

Here, we developed the approach using between-species divergence, paving the way for similar studies within species. The broad applicability of the approach to within-species comparisons is likely controlled by three factors. First, given the short read lengths, it is challenging to generate homologous sequences consistently across individuals. Sequencing the transcriptome might exacerbate this problem, since gene expression can be expected to vary substantially between individuals. Next generation approaches that consistently sequence the same fragment, such as exon capture (Hodges *et al.* 2007) may produce more reliable results. Second, the average distribution of SNP polymorphisms is estimated to be 500bp throughout the entire genome *within* most species (Brumfield *et al.* 2003; Morin *et al.* 2004). Our exploratory study was based on between-species comparisons, where divergence can be expected to be higher than within species.

Our simple simulation performed on transferrin indicated that percentage of sites under selection included in the sequence was an important predictor of statistical power. Ability to detect signatures of selection in shorter sequences (less than 200 codons or 600 bp) was limited to those that had greater than 1 to 2% of codons under selection (1 codon in 150 bp to 3 codons in 600 bp). Sequence length on its own was not a sufficient predictor of power. However, at longer sequence lengths (greater than 600 pb in our example), the percentage

of sites under selection as low as 0.2% could be detected with PAML. In our example, the average contig length was 489 bp, which means that it was possible that genes under selection were not detected. This method therefore has more power when divergence is greater between individual short sequences, which is more likely between populations and species rather than within populations (Kryazhimskiy & Plotkin 2008). However, increasing the number of sequences when sequence divergence is low greatly increases the power of detecting positively selected genes (Anisimova *et al.* 2001).

The methods we describe here would be a key first step for studies addressing adaptive divergence between populations that exhibit different life history traits and do not have *a priori* information of the genes involved in this divergence. Developing single SNP or haplotype markers in selected would allow targeted genotyping, therefore limiting unnecessary resequencing.

## Acknowledgments

This work was funded by a grant from the Washington Sea Grant Program, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA07OAR4170007, Project No. R/F-51 awarded to KAN. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA, or any of their sub-agencies. We would like to thank Margaret Bakewell, University of Michigan, and Giles Goetz, University of Wisconsin, for sharing their scripts, Dave Morris for his extensive help with writing the sequence trimming script. We would also like to thank Jim Seeb, University of Washington, for giving us access to the *O. keta* sequences before publication, Willie Swanson, University of Washington, for advice on data analyses, two anonymous referees whose comments greatly improved this manuscript, and Gary Carvalho for very useful editorial advice.

## Tables

Table 1.1- Genes showing evidence for positive selection using PAML; $d_N$ (M0), $d_S$ (M0) and $d_N/d_S$ (M0) were estimated with the model M0; $\omega_1$ M8 was estimated with model M8 and the p-value associated with the likelihood ratio test when comparing model M8a and M8 is in parenthesis; $p_1$ is the proportion of sites in each gene with an estimated $d_N/d_S$ ratio of $\omega_1$. All the substitutions for Fibroleukine were non synonymous, therefore $d_S = 0$ and $d_N/d_S$ is undefined and represented as *999* in the table.

| **Gene** | **length (# codons)** | **$d_N$ (M0)** | **$d_S$ (M0)** | **$d_N/d_S$ (M0)** | **$\omega_1$ M8 (p-value)** | **$p_1$** |
|---|---|---|---|---|---|---|
| Tubulin alpha chain | 61 | 0.088 | 6.280 | 0.014 | 1 (0.038) | 0.02867 |
| Myeloid leukemia differentiation protein homologue | 163 | 0.224 | 0.460 | 0.486 | 1.770 (0.015) | 0.33329 |
| Voltage-gated hydrogen channel 1 | 228 | 0.166 | 0.285 | 0.581 | 1.827 (0.032) | 0.39045 |
| Nucleoside diphosphate kinase A | 157 | 0.215 | 0.711 | 0.302 | 2.367 (0.016) | 0.32455 |
| Cathepsin S precursor | 224 | 0.078 | 0.137 | 0.566 | 2.453 (0.036) | 0.27168 |
| Cathepsin S precursor | 320 | 0.101 | 0.274 | 0.367 | 4.432 (0.011) | 0.06211 |
| Cytosolic non-specific dipeptidase | 118 | 0.026 | 0.317 | 0.082 | 4.493 (0.014) | 0.0299 |
| LGN-like [*Oncorhynchus mykiss*] | 184 | 0.146 | 0.161 | 0.904 | 5.623 (3.7E-06) | 0.20498 |
| Elongation factor 1-gamma | 176 | 0.037 | 0.204 | 0.183 | 6.381 (0.015) | 0.03077 |
| Gelsolin precursor | 60 | 0.066 | 0.500 | 0.133 | 6.545 (0.017) | 0.08716 |
| Hemoglobin subunit beta-4 | 140 | 0.043 | 0.089 | 0.477 | 12.807 (0.002) | 0.06451 |
| Transaldolase | 52 | 0.148 | 0.011 | 13.254 | 13.255 (0.002) | 1 |
| Protein disulfide isomerase associated 3 | 83 | 0.479 | 0.105 | 4.555 | 13.545 (3.8E-11) | 0.51304 |
| 60S ribosomal protein L5 | 66 | 0.035 | 0.077 | 0.453 | 42.898 (0.018) | 0.0226 |
| Fibroleukine | 48 | | | *999* | *999* (0.014) | 0.43705 |

Table 1.2 - Genes showing evidence for positive selection with PAML for true ortholog and paralog comparisons (identified in the *post hoc* phylogenetic analysis); ortholog comparisons are indicated by [O-1] and [O-2], O-1 and O-2 correspond to the two orthologs available; [P] represents paralog comparisons; dN (M0), dS (M0) and dN/dS (M0) were estimated with the model M0; $\omega_1$ M8 was estimated with model M8 and the level of significance of the likelihood ratio test when comparing model M8a and M8 are characterized as follows: *: 0.1; **:0.05; ***:0.005; N.S.: Non significant; LS: lineage specific selection; $p_1$ is the proportion of sites in each gene with an estimated $d_N/d_S$ ratio of $\omega_1$.

| Gene | length (# codons) | $d_N$ (M0) | $d_S$ (M0) | $d_N/d_S$ (M0) | $\omega_1$ M8 | $p_1$ |
|---|---|---|---|---|---|---|
| Gelsolin precursor [O-1] | 81 | 0.010 | 0.102 | 0.099 | 34.55* | 0.03 |
| Gelsolin precursor [O-2] | 112 | 0.008 | 0.118 | 0.070 | N.S. | N.S. |
| Gelsolin precursor [P] | 81 | 0.054 | 0.358 | 0.151 | N.S. | N.S. |
| Cytosolic non-specific dipeptidase [O-1] | 72 | 3E-05 | 0.305 | 1.E-04 | N.S. | N.S. |
| Cytosolic non-specific dipeptidase [O-2] | 142 | 0.011 | 0.078 | 0.143 | 28.59* | 0.02 |
| Cytosolic non-specific dipeptidase [P] | 72 | 0.011 | 0.418 | 0.026 | N.S. | N.S. |
| Voltage-gated hydrogen channel 1 [O-1] | 215 | 0.065 | 0.095 | 0.686 | 24.99** LS | 0.01 |
| Voltage-gated hydrogen channel 1 [P] | 135 | 0.131 | 0.342 | 0.383 | N.S. | N.S. |
| Cathepsin S precursor [O-1] | 224 | 0.078 | 0.137 | 0.566 | 2.45** LS | 0.27 |
| Cathepsin S precursor [O-2] | 165 | 0.051 | 0.122 | 0.414 | 5.31*** | 0.06 |
| Cathepsin S precursor [P] | 282 | 0.140 | 0.337 | 0.415 | 2.94** | 0.09 |
| Myeloid leukemia differentiation protein homologue [O-1] | 76 | 0.045 | 0.096 | 0.476 | N.S. | N.S. |
| Myeloid leukemia differentiation protein homologue [O-2] | 162 | 0.095 | 0.117 | 0.808 | N.S. | N.S. |
| Myeloid leukemia differentiation protein homologue [P] | 76 | 0.216 | 0.422 | 0.513 | N.S. | N.S. |
| LGN-like [O-1] | 176 | 0.065 | 0.082 | 0.797 | 16.72** LS | 0.02 |
| LGN-like [O-2] | 65 | 0.032 | 0.058 | 0.556 | N.S. | N.S. |

| | | | | | | |
|---|---|---|---|---|---|---|
| LGN-like [P] | 65 | 0.134 | 0.165 | 0.811 | 2.06* | 0.43 |
| Protein disulfide isomerase associated 3 [O-1] | 152 | 0.077 | 0.276 | 0.280 | 26.08** LS | 0.02 |
| Protein disulfide isomerase associated 3 [P] | 184 | 0.144 | 0.723 | 0.199 | N.S. | N.S. |
| Transaldolase | 52 | 0.148 | 0.011 | 13.254 | 13.26*** LS | 1.00 |
| 60S ribosomal protein L5 | 63 | 0.014 | 0.085 | 0.161 | N.S | N.S. |
| Nucleoside diphosphate kinase A | 157 | 0.215 | 0.711 | 0.302 | 2.367 ** | 0.32 |
| Fibroleukine | 48 | | | *999* | *999* ** | 0.44 |

Table 1.3 – Relationship between sequence length, percentage of codons under selection and ability to detect selection in the partial coding sequences of the transferrin gene in five salmon species.

| Length (codons) | Proportion of fragments showing signs of selection | Known percentage of codons under selection | Proportion of fragments showing no sign of selection | Known percentage of codons under selection |
|---|---|---|---|---|
| **50** | 0.67 | 2.0% - 8.0% | 0.33 | 2.0% |
| **100** | 0.89 | 2.0% - 5.0% | 0.11 | 1.0% |
| **150** | 0.80 | 2.0% - 5.3% | 0.20 | 1.3% |
| **200** | 0.90 | 2.0% - 3.5% | 0.10 | 1.5% |
| **300** | 1 | 2.0% - 3.7% | 0 | NA |
| **400** | 1 | 2.3% - 3.3% | 0 | NA |
| **500** | 1 | 0.2% - 3.2% | 0 | NA |

## Figures

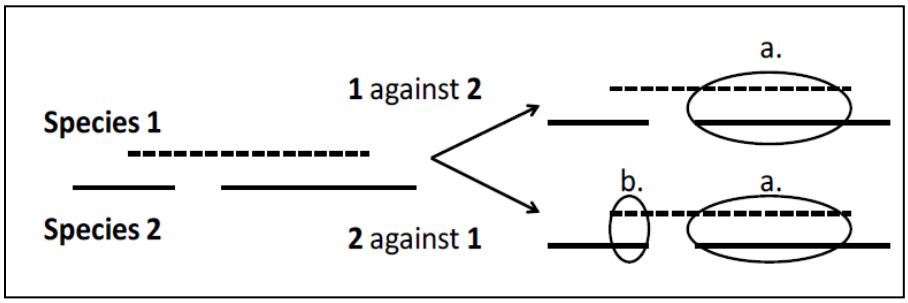


Figure 1.1 – Multi-way BLAST searches in partial sequences, illustrating the likelihood of detecting the sequences in common; the available sequence for species 1 is longer than for species 2, and there are two overlapping regions between the species. A BLAST of species 1 will only detect one overlapping region; a BLAST of species 2 against species 1 will detect both regions. Multi-way BLASTs are therefore necessary to systematically detect discontinuous overlapping regions.

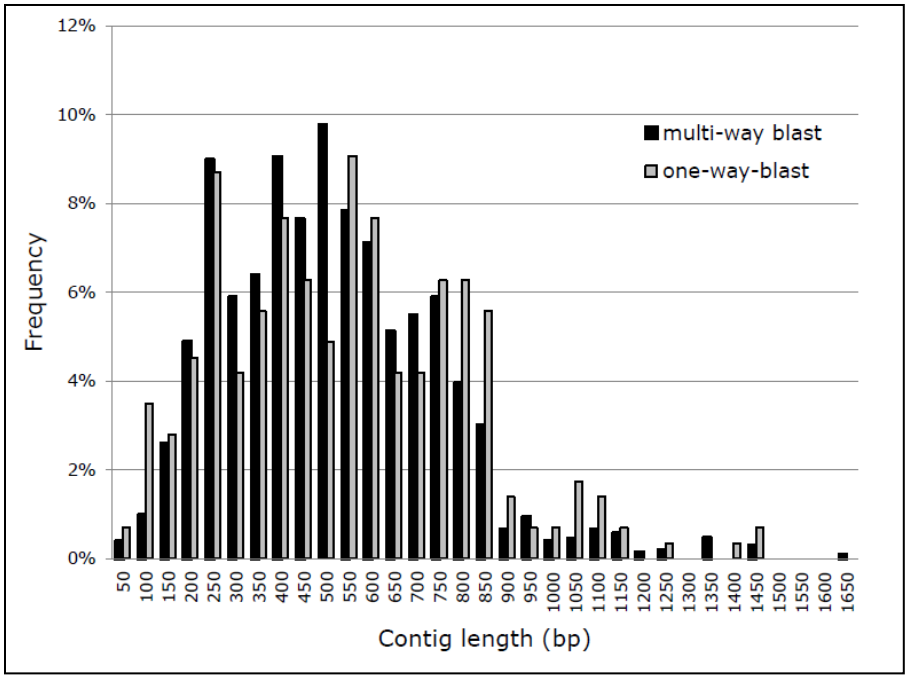


Figure 1.2 – Distribution of the length of aligned sequence regions between six salmon species (*Salmo salar*, *Salvelinus fontinalis, Oncorhynchus mykiss*, *Oncorhynchus tshawytscha, Oncorhynchus nerka and Oncorhynchus keta)* using multi-way BLAST and one-way BLAST. *Lengths are categorized in 50 base pairs categories: for example category 450 contains contigs with a length between 401 and 450 base pairs*

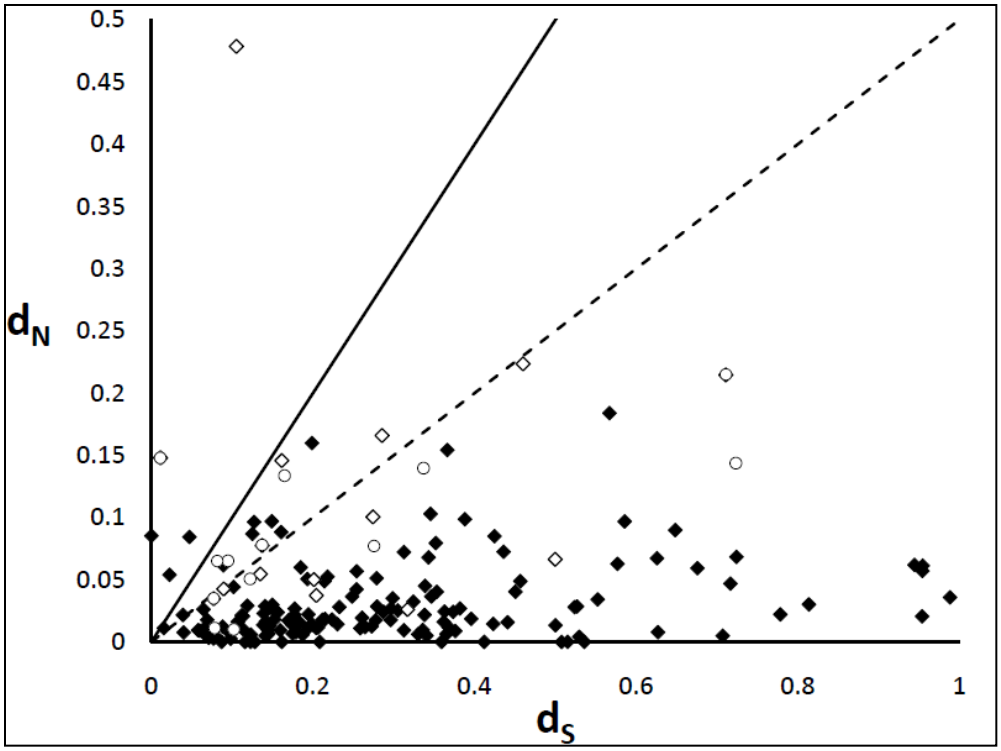


Figure 1.3 – Whole-gene $d_N$ and $d_S$ estimates across partial gene sequences in six salmon species. *The dashed and solid lines correspond to ω=0.5 and ω=1 respectively. Open squares represent genes showing evidence of positive selection but with possible paralogs present. Open circles represent true orthologs comparisons showing evidence of selection based on site-specific tests.*

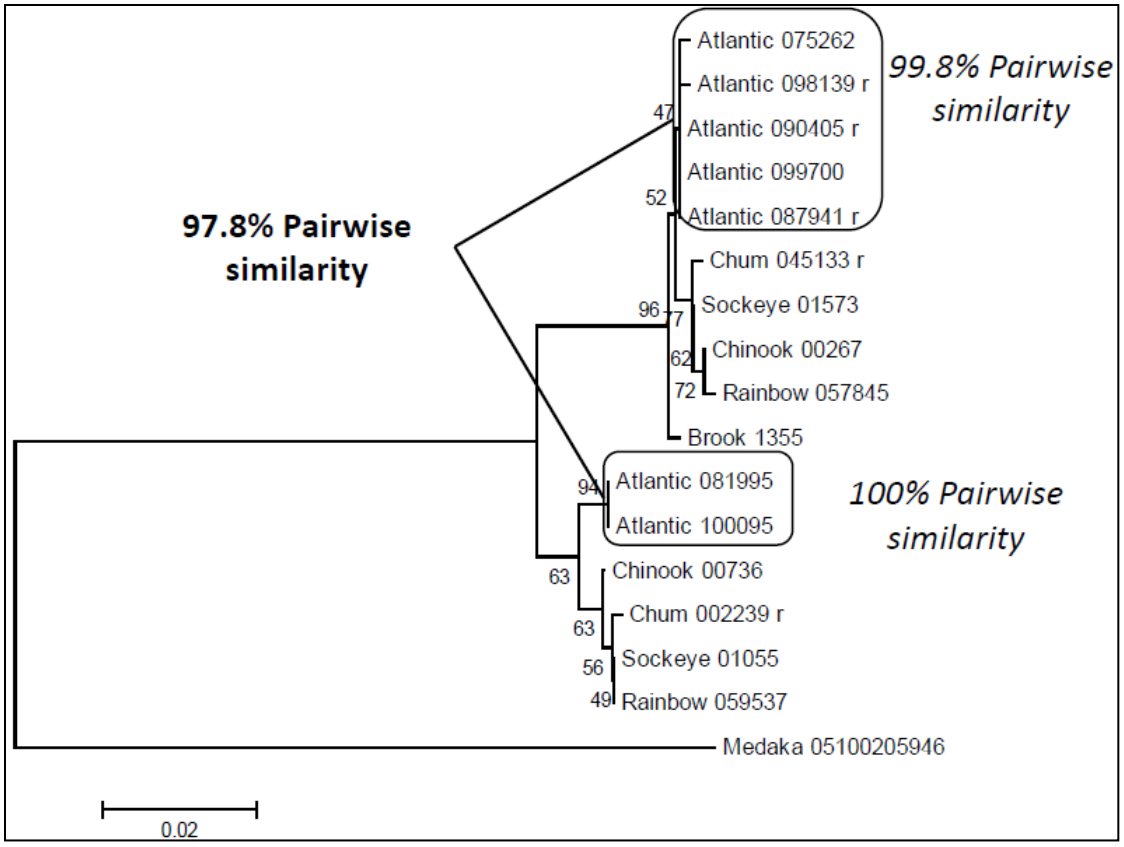


Figure 1.4 – Neighbor-Joining tree of 14-3-3 protein beta/alpha-2 in salmon species. *Pairwise similarities between orthologs (in boxes) are shown in italics, and between paralogs (denoted by arrows) in bold.*

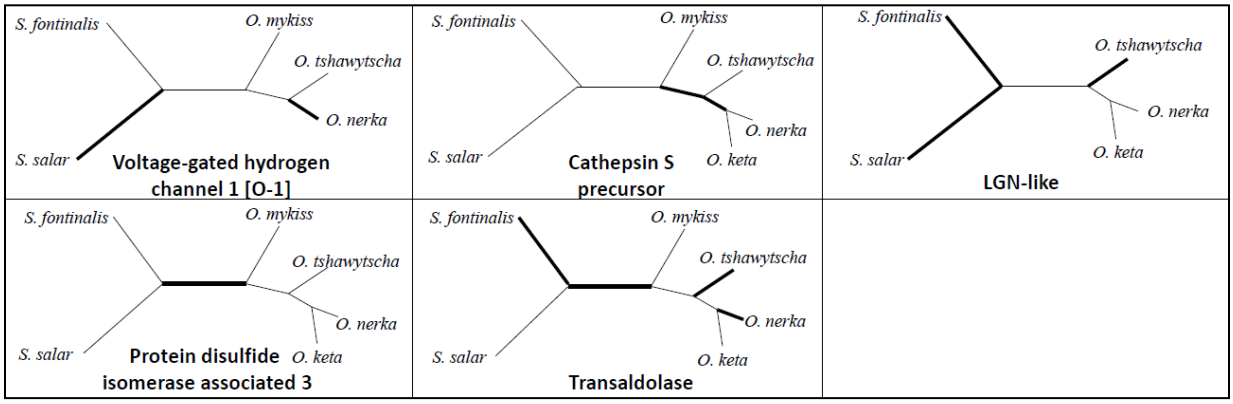


Figure 1.5 – Lineage specific selection. *Lineages under positive selection are represented in bold*

## Supplementary Material

**S1.1-blast_it** - Script for automated BLAST with instructions

**S1.2-clustal_align** - Script for automated CLUSTAL alignment

**S1.3-sequence_trim** - Script for automated sequence trimming; uses the output file from clustal_align.pl

## References


Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805-1814.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. Turner EBJ). Plenum Press, New York.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics* **172**, 2567-2582.

Andreassen R, Lunner S, Hoyheim B (2009) Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (Salmo salar). *BMC Genomics* **10**.

Anisimova M, Bielawski J, Dunn K, Yang Z (2007) Phylogenomic analysis of natural selection pressure in Streptococcus genomes. *BMC Evolutionary Biology* **7**.

Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution* **18**, 1585-1592.

Bakewell MA, Shi P, Zhang JZ (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proceedings *of the National Academy of Sciences of the United States of America* **104**, 7489-7494.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**, 910-918.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**, 249-256.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M,

Clark AG (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng XG, White TJ, Sninsky JJ, Adams MD, Cargill M, (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960-1963.

Clark NL, Swanson WJ (2005) Pervasive adaptive evolution in primate seminal proteins. *Plos Genetics* **1**, 335-342.

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research* **8**, 1229-1231.

Consuegra S, Johnston IA (2008) Effect of natural selection on the duplicated lysyl oxidase gene in Atlantic salmon. *Genetica* **134**, 325-334.

Crespi BJ, Fulton MJ (2004) Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny. *Molecular Phylogenetics and Evolution* **31**, 658-679.

Dalziel AC, Rogers SM, Schulte PM (2009) Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology* **18**, 4997-5017.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution* **13**, 685-690.

Findlay GD, Yi XH, MacCoss MJ, Swanson WJ (2008) Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. *Plos Biology* **6**, 1417-1426.

Ford MJ, Thornton PJ, Park LK (1999) Natural selection promotes divergence of transferrin among salmonid species. *Molecular Ecology* **8**, 1055-1061.

Ford MJ (2001) Molecular evolution of transferrin: Evidence for positive selection in salmonids. Molecular *Biology and Evolution* **18**, 639-647.

Ford MJ (2002) Applications of selectively neutral tests to molecular ecology. *Molecular Ecology* **11**, 1245-1262.

Freamo H, O'Reilly P, Berg PR, Lien S, Boulding E (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources* **11** (Suppl. 1), 243–256.

Ge GT, Cowen L, Feng XC, Widmer G (2008) Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Comparative and Functional Genomics*, **6**.

Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. *Nature Genetics* **3**, 266-272.

Goldman N, Yang ZH (1994) Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences. *Molecular Biology and Evolution* **11**, 725-736.

Green P (1996) Documentation for Phrap.

Hancock AM, Di Rienzo A (2008) Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. *Annual Review of Anthropology* **37**, 197-217.

Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources* **11** (Suppl. 1), 150–161.

Hess JE, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Molecular Ecology Resources* **11** (Suppl. 1), 137–149.

Hodges E, Xuan Z, Balija V, et al. (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics* **39**, 1522-1527.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3-17.

Kelley JL, Swanson WJ (2008) Positive selection in the human genome: From genome scans to biological significance. *Annual Review of Genomics and Human Genetics* **9**, 143-160.

Kitano T, Liu YH, Ueda S, Saitou N (2004) Human-specific amino acid changes found in 103 protein-coding genes. *Molecular Biology and Evolution* **21**, 936-944.

Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmbhatt S, Rosner J, Rexroad CE, McGowan CR and Davidson WS (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* **9**.

Kryazhimskiy S, Plotkin JB (2008) The Population Genetics of dN/dS. *Plos Genetics* **4**.

Kunstner A, Wolf JBW, Backstrom N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, Jarvis ED, Warren WC and Ellegren H (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology* **19**, 266-276.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG, (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Li W-H (1997) Molecular evolution. Sinauer, Sunderland, Mass.

Liu G, Uddin M, Islam M, Goodman M, Grossman LI, Romero RW, Derek E (2007) OCPAT: an online codon-preserved alignment tool for evolutionary genomic analysis of protein coding sequences. *Source Code for Biology and Medicine* **2**.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009) Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**.

Miller JM, Poissant J, Kijas JW, Coltman DW and the International Sheep Genomics Consortium (2010) A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources*. no. doi: 10.1111/j.1755-0998.2010.02918.x

Morin PA, Luikart G, Wayne RK, Grp SNPW (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources* **9**, 66-73.

Naish KA, Hard JJ (2008) Bridging the gap between the genotype and the phenotype: linking genetic variation, selection and adaptation in fishes. *Fish and Fisheries* **9**, 396-422.

Narum SR, Banks M, Beacham TD, Bellinger MR, Campbell MR, Dekoning J, Elz A, Guthrie CM, Kozfkay C, Miller KM, Moran P, Phillips R, Seeb LW, Smith CT, Warheit K, Young SF, Garza JC (2008) Differentiating salmon populations at broad and fine

geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology* **17**, 3464-3477.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Research* **19**, 838-849.

Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology* **19**, 115-131.

Seeb J, Pascal C, Seeb L, Grau E, Templin W, Roberts S, Harkins T (in press) Next generation transcriptome sequencing and high-resolution melt analyses enhance SNP discovery in chum salmon, *Oncorhynchus keta*. *Molecular Ecology Resources*

Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP (2010) Adaptive evolution of energy metabolism genes and the origin of flight in bats. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8666-8671.

Swanson WJ, Nielsen R, Yang QF (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution* **20**, 18-20.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

Templin WA, Seeb JE, Jasper JR, Barclay AW, Seeb LW (2011) Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. *Molecular Ecology Resources* **11** (Suppl. 1), 215–235.

Tonteri A, Vasemagi A, Lumme J, Primmer CR (2010) Beyond MHC: signals of elevated selection pressure on Atlantic salmon (Salmo salar) immune-relevant loci. *Molecular Ecology* **19**, 1273-1282.

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* **138**, 433-451.

Willing EM, Bentzen P, van Oosterhout C, Hoffman M, Cable J, Bredent F, Weigel D, Dreyer C (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology* **19**, 968-984.

Wright SI, Andolfatto P (2008) The Impact of Natural Selection on the Genome: Emerging Patterns in Drosophila and Arabidopsis. *Annual Review of Ecology Evolution and Systematics* **39**, 193-213.

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555-556.

Yang ZH (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568-573.

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591.

Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449.

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107-1118.

# Chapter 2 – A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) with extensive chromosomal coverage, and a reference database comprising sequenced Restriction-Associated DNA loci

Marine S.O. Brieuc, Charles D. Waters, James E. Seeb, Kerry A. Naish
School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA

## Abstract

Comparisons between salmon species reveal that they underwent extensive chromosomal rearrangements following the whole genome duplication event that occurred in their common ancestor 25 to 100 million years ago. Salmonids are now residual tetraploid and occasional pairing between homeologous chromosomes exists in males. This phenomenon may result in reduced divergence between some pairs of homeologs; this can be investigated by mapping duplicated loci. Very dense linkage maps can be constructed with sequences derived from reduced representation libraries, such as by RAD (restriction-site associated DNA) sequencing. However, identification of loci *de novo* may be challenging with short sequences, especially for species with a duplicated genome. Development of a reference database of loci in these species would facilitate locus identification and data sharing. Linkage maps are also a valuable tool for genome-wide applications (genome-wide association studies, quantitative trait loci mapping or genome scans). In this study we have developed a reference database of RAD loci for Chinook salmon comprising 48528 non-duplicated loci and 6409 known duplicated loci. We have mapped 7146 of these RAD loci on a very dense linkage map with predicted complete coverage of all 34 chromosomes of the species, and for which all arms have been identified through centromere mapping. After examining the position of 799 duplicated loci we determined that homeologs have diverged at different rates following whole genome duplication, and that the rate of divergence was variable along the chromosome arms. We also identified two new homeologies with high support for Chinook salmon. Comparison with rainbow trout confirmed previously identified homologies between the species. As chromosome arms are highly conserved

across species, the major resources developed for Chinook salmon in this study are also relevant for other related species.

## Introduction

Salmonids are descended from a single ancestor that underwent a whole genome duplication event 25 to 100 million years ago (Allendorf & Thorgaard 1984). Partial diploidy was restored following the event through different mechanisms: within the subfamily Thymallinae, chromosomes have evolved by inversions, whereas chromosome structure within the subfamilies Salmoninae and Coregoninae evolved by Robertsonian rearrangements (Phillips & Rab 2001). Comparisons between species show that the chromosome arm number (NF) has been conserved (around 100) within the Salmoninae subfamily, specifically *Salmo, Salvelinus and Oncorhynchus,* with the exception of Atlantic salmon (*Salmo salar*) which has NF=54-58 (Allendorf & Thorgaard 1984; Phillips & Rab 2001). However, the number of chromosomes varies extensively between species, from 2n=52-54 in Pink salmon (*Oncorhynchus gorbutscha*) to 2n=84-86 in the Japanese char (*Salvelinus pluvius*). This variation is the result of extensive Robertsonian rearrangement (Ohno 1970; Phillips & Rab 2001).

Linkage maps can be used to study the differences of chromosomal rearrangements between salmonid species that have occurred since the whole genome duplication event. Previous studies have demonstrated that most rearrangements resulted from fusions of two acrocentric chromosomes into one metacentric chromosome, fissions of metacentrics into two acrocentrics, or a combination of both, therefore maintaining a relatively constant number of chromosome arms (e.g. Danzmann *et al.* 2005; Gharbi *et al.* 2006, Naish *et al.* submitted). Extensive genome mapping efforts have shown that most chromosome arms are syntenic between species (Danzmann *et al.* 2005; Gharbi *et al.* 2006; Guyomard *et al.* 2012; Phillips & Rab 2001). For example, comparisons between rainbow trout and Atlantic salmon revealed that 12 large acrocentric chromosomes in the latter were the result of tandem fusions of regions corresponding to full chromosome arms in rainbow trout (Phillips *et al.* 2009). Similarly, comparisons between Chinook salmon and rainbow trout

identified conservation of 13 whole chromosomes between the two species, while other chromosomes appear to be the result of independent Robertsonian arrangements (Naish *et al.* submitted). Identification of chromosomal rearrangements between all salmonids therefore provides insights into chromosome evolution and will allow genomic resources from one species to be rapidly applied to other salmon species with fewer available resources.

Linkage maps can also be used to examine the remnant effects of whole genome duplication on chromosomal divergence and recombination frequency. Evidence of tetrasomic inheritance in Salmoninae supports the fact that restoration of diploidy is incomplete (Allendorf & Danzmann 1997; Allendorf & Thorgaard 1984; Wright *et al.* 1983). A model of secondary tetrasomy has been proposed, in which homologous chromosomes pair first in regions proximal to the centromere, followed by homeologous pairing and recombination in the distal regions (Allendorf & Thorgaard 1984; Wright *et al.* 1983). Homeologous pairing can result in residual tetrasomic inheritance and pseudolinkage (Allendorf & Thorgaard 1984; Wright *et al.* 1983). The latter is characterized by an excess of non-parental progeny types in crosses and can be identified by the observation of linkage disequilibrium between physically unlinked loci using two-point linkage analysis (e.g. Lien *et al.* 2011). Residual tetrasomic inheritance or pseudolinkage has been observed in many salmonids, such as Brown trout (*Salmo trutta*; Gharbi *et al.* 2006), rainbow trout (*Oncorhynchus mykiss*; Allendorf & Danzmann 1997; Danzmann *et al.* 2005; Sakamoto *et al.* 2000) and Atlantic salmon (S*almo salar*; Lien *et al.* 2011). In salmon species, homeologous pairing is thought to be limited to males (Allendorf & Thorgaard 1984; Wright *et al.* 1983). Only few instances of pseudolinkage in females have been inferred through two-point linkage analysis (Danzmann *et al.* 2005; Lien *et al.* 2011). Homeologous pairing likely prevents distal regions of chromosome arms from diverging (Allendorf & Thorgaard 1984; Wright *et al.* 1983) and may result in a higher retention of duplicated loci on the chromosome arms involved in ongoing recombination. In a recent study in Atlantic salmon based on single nucleotide polymorphisms (SNPs), Lien *et al.* (2011) determined that duplicated loci were not randomly distributed among all chromosomes within this species. The relative position of the duplicated loci along each homeologous chromosomes pair has

yet to be investigated; such information will reveal the regions of the homeologous chromosomes with reduced divergence and likely involved in pairing between such chromosomes.

Linkage maps also enable QTL mapping (e.g. Collard *et al.* 2005; Nichols *et al.* 2008), genome-wide association analyses (e.g. Cichon *et al.* 2009; Magwire *et al.* 2012), and identification of regions of the genome that have played a significant role in adaptive evolution (e.g. Bradbury *et al.* 2013; Tsumura *et al.* 2012). However, these applications require linkage maps with high marker density. The recent and rapid improvement in sequencing technologies (Hudson 2008; Shendure & Ji 2008) has facilitated the characterization of thousands of variable markers, even for species with little or no available genetic information (Davey *et al.* 2011). Several of these approaches take advantage of the large amount of information afforded by sequencing a reduced portion of the genome. Restriction site associated DNA (RAD) sequencing (Baird *et al.* 2008; *c.f.* Miller *et al.* 2007) has the advantage of targeting a consistent portion of the genome across individuals. Constructing linkage maps with such markers could therefore facilitate genome-wide studies relying on genome position, such as QTL mapping or genome scans, for non-model species.

One of the challenges associated with the use of sequences derived from reduced representation libraries is the consistent assignment of reads to corresponding loci across individuals, especially when reads are short (60 to100 nucleotides, typical of RAD sequencing). Identification of loci *de novo* for non-model species has been facilitated with the use of newly developed bioinformatic approaches, (for example, approaches implemented in STACKS (Catchen *et al.* 2011) or RAINBOW (Chong *et al.* 2012)), but inconsistencies are still apparent for species with a duplicated genome. This issue could be resolved with the creation of a reference database of RAD loci for the species of interest, where duplicated loci would be identified. This database would rapidly facilitate alignment of newly sequenced individuals in related studies, and promote data sharing across research groups.

Interest for RAD sequencing has recently increased for salmon research (*e.g.* Amish *et al.* 2012; Hecht *et al.* 2012; Houston *et al.* 2012). Among salmonids, Chinook salmon (*Oncorhynchus tshawytscha*) is of economic, cultural and conservation interest. Its distribution ranges from Japan to California. On the West coast of the United States 17 evolutionary significant units (ESUs, Waples 1995) have been identified, nine of which are now threatened or endangered (Good *et al.* 2005). Conservation efforts would benefit from the development of a genetic map, which can be used in a wide variety of ecological, evolutionary and management applications. Although extensively studied, there is only one linkage map available for Chinook salmon (Naish *et al.* submitted). The existing map comprises 361 microsatellite markers, many of which amplify both in Chinook salmon and rainbow trout (*Oncorhynchus mykiss*). The linkage groups described by this map have been aligned to the 34 chromosomes of Chinook salmon (Phillips *et al.* submitted). However the centromeres have not been identified on the microsatellite map. Most Robertsonian fusions occur at the centromere (Slijepcevic 1998). Determining the position of the centromere is therefore necessary to reliably identify chromosome arms. Centromere-mapping has successfully been implemented in other salmonids (Allendorf *et al.* 1986; Guyomard *et al.* 2006; Lindner *et al.* 2000). Gynogenetic diploid crosses can be used for gene-centromere (or half-tetrad) mapping (reviewed in Danzmann & Gharbi 2001; Johnson *et al.* 1996). The second polar body is retained during creation of gynogenetic diploid progeny. Therefore a progeny will be heterozygous at a locus if a crossover event occurred somewhere between a given marker and the centromere during meiosis I. The percentage of heterozygous offspring at a locus is expected to be 0% at the centromere, increasing to 100% in the telomeric region because salmonids exhibit complete to near complete interference and typically have one crossover event per chromosome arm (Thorgaard *et al.* 1983). Construction of a linkage map with a high density of markers and location of the centromeres would therefore be a valuable resource for Chinook salmon genetic studies. A dense map derived from RAD markers exists for rainbow trout (Miller *et al.* 2012), facilitating additional comparisons between the two species.

Our overall aims are to develop genomic resources for Chinook salmon, to study chromosomal evolution in this species, and to compare this evolution to that of other

salmonids. Our specific objectives are first, to construct a reference database of RAD sequences that can be used for alignments of sequences generated in future projects. Second, we aim to improve the genomic map for Chinook salmon by populating the existing genome map with thousands of RAD markers from the reference database and identifying the location of the centromeres on the map. Third, we explicitly aim to improve our understanding of divergence of homeologous chromosome arms in the salmonids by mapping duplicated loci. Finally, we will improve our understanding of chromosome arm rearrangement between Chinook salmon and rainbow trout using comparative analyses of marker-dense maps for the two species.

## Methods

### Sample collection

We used 159 individuals from a total of 10 populations from the Columbia River basin, Pacific Northwest, USA to identify the preliminary list of loci for Chinook salmon. Additionally, three haploid crosses comprising 46, 48 and 72 individuals per family were used to identify the duplicated loci in the database and construct the linkage map. All haploid individuals should be homozygous at every locus. Therefore if an individual appears heterozygous at a locus, that locus has to be duplicated. The haploid crosses were created at the Cle Elum Supplementation and Research Facility (CESRF) by fertilizing eggs with UV irradiated milt following the protocol of Thorgaard *et al.* (1983) and sampled before hatching. Whole embryos were collected and stored in ethanol. We also sampled 44 F2 progeny from the diploid cross of Naish *et al.* (submitted) to verify linkage group representation in the haploid families using the microsatellite markers mapped in Naish *et al.* (submitted) as the 34 Chinook linkage groups corresponding to the 34 chromosomes have been identified (Phillips *et al.* submitted). Recombination rates vary between the sexes in salmonids (Lien *et al.* 2011; McClelland & Naish 2008; Moen et al. 2004; Moen *et al.* 2008); we therefore mapped the female meiosis for the diploid cross to obtain accurate marker order. Finally we used three gynogenetic diploid families, created at the University of Washington hatchery facility, to map the centromere on each linkage group. Eggs were

fertilized with UV irradiated milt and subsequently heat shocked to retain the second polar body (Thorgaard *et al.* 1983). Fish were harvested as parr and stored in 100% ethanol. We sampled the dam and 84, 90 and 93 progeny from each gynogenetic diploid family.

Genomic DNA was extracted using the DNeasy extraction kit (QIAGEN, Valencia, CA, USA) following the manufacturer's procedures. Each sample was prepared for RAD sequencing, using *Sbf*I as a restriction enzyme and 6-nucleotide individual-specific barcodes, as described in Baird *et al.* (2008). Twelve to forty-eight individuals were then sequenced per lane on an Illumina platform (GAII or HiSeq) using 100-nucleotide single-read sequencing. Reads were sorted per individual and barcodes were removed using PROCESS_RADTAGS implemented in STACKS (Catchen *et al.* 2011). The last 20 nucleotides were subsequently trimmed because the last 20 base pairs of the sequence had a consistently lower quality. For the purpose of this study, we defined a locus as a 74-nucleotide RAD sequence.

**Reference database of RAD loci**

Creation of the reference database of RAD loci was carried out using three steps: the generation of a preliminary database of loci for Chinook salmon, the screening of the preliminary database for repetitive loci and loci aligning with repeat sequences, and the identification of duplicated loci.

Reads for all diploid individuals were sorted into polymorphic and monomorphic loci *de novo* using STACKS (Catchen *et al.* 2011). We retained a consensus sequence for every locus that had been sequenced with a depth greater than 5X in more than 135 individuals (85%) as a temporary database of loci: these loci were used for further screening.

The screening steps were aimed at identifying repetitive loci and loci with repeat units. We used two alignment-based strategies. First, loci within the temporary database were aligned against themselves using BOWTIE (Langmead *et al.* 2009), allowing up to three nucleotide mismatches. A locus that aligned to several loci, including itself, was excluded from the database. Second, a BLAST search (Basic Local Alignment Search Tool; Altschul *et al.* 1990) of the temporary database was conducted against itself. Loci within the database

that did not return a match or where the best match for a given locus was not itself were discarded from the temporary database. The latter typically occurred when there was a repeat unit in the sequence.

Finally, duplicated loci were identified using the three haploid families. Reads for all the haploid individuals were sorted into loci by alignment to the temporary database using BOWTIE. Individual reads from the haploids that aligned to more than one locus in the database could not be confidently relied upon in further analyses, and so were removed from the database. Loci with a depth of less than 10 reads for an individual were discarded for that individual. Genotypes for each individual were obtained using STACKS, which uses a maximum likelihood approach to identify polymorphisms (Catchen *et al.* 2011). Within each haploid family, the presence of a single heterozygous genotype at a locus was considered the result of a sequencing error, and the locus was retained in the database. However, if more than one haploid individual was heterozygous at a locus, this locus was identified as being duplicated. This final step provided the final database of RAD loci for Chinook salmon, against which all further alignments were made.

## Linkage mapping

### *Genotyping*

Genotypes at every non duplicated polymorphic locus in the haploid crosses were identified during the creation of the reference database. Duplicated markers identified in the haploids during database development were used for mapping when one of the paralogs was polymorphic (OPP – one paralog polymorphic, parental genotype aa & ab, or aa & bc) or when both paralogs were polymorphic for different alleles (BPP, parental genotype ab & ac and ab & cd). We also observed ab & ab but did not map these genotypes because heterozygous individuals were uninformative. The types of duplicated loci we observed are described in Table 2.1. We aligned all the reads for the diploid cross and gynogenetic diploid crosses to the reference database using BOWTIE and identified the polymorphic loci using STACKS. STACKS uses a maximum likelihood approach to determine whether a polymorphism in an individual is true, or whether it is due to a sequencing error

(Hohenlohe *et al.* 2010). This approach can be biased against the designation of heterozygous genotypes for individuals that differ in sequence depth between the two alleles. To correct this bias, we developed a Python script (available upon request) that called a heterozygote if both alleles had a depth of more than two and the total read depth at the locus was 10X or greater; this was the minimum depth we designated previously. Parental haplotypes for the diploid cross were determined using linkage relationships with the previously mapped microsatellite markers.

Additionally, we used 5'-nuclease genotyping as in Seeb et al. (2011) to screen and attempt to map 384 SNPs that originated from other labs (Smith et al. 2005a; Smith et al. 2005b; Campbell & Narum 2008; Abadia-Cardoso et al. 2011) in two haploid families. Many of these loci are polymorphic ESTs that used in conservation and management applications for Chinook salmon across Pacific North America (e.g., Smith et al. 2005c; Hess et al. 2011; Templin et al. 2011; Matala et al. 2012).

*Linkage groups and chromosome identification*

We used ONEMAP (Margarido *et al.* 2007) for genome mapping in the haploid crosses and the F2 diploid cross. The Chinook salmon karyotype comprises 34 pairs of chromosomes (Phillips & Rab 2001). We therefore predicted 34 linkage groups per mapping cross. Linkage groups were identified independently for each haploid and diploid family using ONEMAP with a maximum recombination fraction of 0.25 and a starting LOD of 3.0. This LOD was subsequently increased by increments of 1.0 until the number of linkage groups identified was 34 or higher. We then used the microsatellite markers previously mapped and the RAD loci polymorphic in the diploid cross and the haploid crosses to identify each chromosome. Markers on each linkage group were subsequently ordered using ONEMAP for each haploid family. Individual haploid maps were merged using MERGEMAP (Wu *et al.* 2011) to create a consensus map.

*Centromere mapping*

We estimated the proportion of heterozygous progeny in each gynogenetic diploid family at every non-duplicated marker mapped on the haploid map and polymorphic in the

gynogenetic diploid crosses. This information was used to identify the centromere and the chromosome type (acrocentric or metacentric) for each haploid family. The diploid map was then used to characterize the short (p) arm and long (q) arm for each chromosome arm as defined in Phillips *et al.* (submitted).

## Analysis of the properties of the Chinook salmon linkage groups

### *Frequency of recombination*

Recombination is usually reduced around the centromere in most species (Nachman 2002) and in the telomeric regions in the female in salmonids (Lien *et al.* 2011). Reduced recombination will result in a high number of loci mapping to the same position. Here we examined the distribution of the markers along the linkage groups relative to the center of the centromere to determine recombination frequency.

### *Crossover frequency and interference*

Salmonids are thought to exhibit complete to near complete interference (Thorgaard *et al.* 1983). We estimated the number of crossover events per chromosome arm using LINKMFEX (Danzmann 2005). Metacentric linkage groups were divided in two chromosome arms. For each chromosome arm we counted the number of progeny with 0, 1 or more crossovers. Absence of double crossovers on all chromosome arms for every progeny would confirm the hypothesis of complete interference.

### *Distribution of duplicated markers across the genome*

Two types of duplicated markers were used in this study. Duplicated loci with both paralogs polymorphic (BPP) were used to infer homeologies, since both paralogs could be mapped. Occasional homeologous chromosome pairing in salmon may result in reduced divergence between the arms involved. We examined the position of the duplicated loci on the consensus haploid map to determine whether there was a bias in distribution of these loci. We reasoned that this analysis would identify chromosomal regions of reduced divergence between homeologs, indicating possible map positions where homeologs have a tendency to pair. Here we aimed at determining the relative proportion of duplicated loci

along the linkage groups. Because map positions are not uniformly distributed along the chromosomes, we used a kernel smoothing sliding window approach with a bandwidth of 2cM to determine the relative proportion of duplicated loci along the linkage groups.

### Comparative mapping with rainbow trout

We aligned the 40649 RAD loci identified in Miller *et al.* (2012) with the reference dataset of loci created for Chinook salmon using BOWTIE, allowing a maximum of three nucleotide mismatches per locus. Mapped loci in common between the two species were used to identify homologies between rainbow trout and Chinook salmon and confirm alignment with previous studies (Naish *et al.* submitted, Phillips *et al.* submitted).

## Results

### Reference database of RAD loci for Chinook salmon

A total of 62249 putative loci were sequenced in at least 135 individuals from the Columbia River with a minimum depth of five per individual: these sequences formed the temporary database of RAD loci. Of these, 2713 were removed because they did not align uniquely to themselves. After conducting a BLAST search of the temporary database against itself, 1451 loci did not have a BLAST result or the best hit was not itself, mostly due to the presence of repetitive units in the sequence (data not shown). Alignments of all reads for the haploid individuals against the updated temporary database were not unique for 3148 loci and these were therefore removed from the database. Finally, 6409 duplicated loci were identified as heterozygous in more than two progeny in at least one haploid family and were identified as such. The final reference database comprised 48528 putative non-duplicated loci and 6409 duplicated loci .

## Linkage mapping

*Haploid and Diploid Linkage Maps*

The three haploid families (here, family A, B and C) had 3528, 3325 and 3403 biallelic polymorphic RAD loci respectively, representing 7146 unique RAD loci. Two families were genotyped using the 384 5'-nuclease panel (family A and B); each had 98 and 92 polymorphic SNPs respectively, 149 of which were unique. We used 2674 informative biallelic RAD loci scored in the diploid cross to develop sex-specific linkage groups. A subset of 1189 loci was polymorphic in the female parent and linked to previously mapped microsatellite markers (Naish *et al.* submitted). We identified thirty four linkage groups corresponding to the chromosomes for each haploid cross using 578 RAD loci that were in common between the diploid and haploid families.

We mapped 3485, 3291 and 3273 non duplicated and duplicated markers within each of the three haploid families (Table 2.2). The map lengths ranged between 2834.9 cM and 3099.6 cM (Table 2.2). A total of 2319 loci were polymorphic in more than one family and were used to merge the haploid maps. The consensus haploid map comprised 7304 markers and measured 4163.9 cM (Figure 2.1, Supplementary material S2.1).

The diploid map comprised 1101 non duplicated RAD markers and 242 microsatellite loci (Supplementary material S2.1). All 34 chromosomes were identified, but five chromosomes were represented by 2 linkage groups each (Ots08, Ots15, Ots19, Ots26 and Ots29). The number of individuals scored per locus was variable due to lower DNA quality. As a result, the marker order on the female diploid map was not consistently reliable. Therefore, the diploid map was not merged with the haploid maps. However, the microsatellite markers proved reliable in assigning linkage group arms to chromosomes.

*Centromere mapping*

Of the 6348 non duplicated RAD markers placed on the haploid map, 3021 were polymorphic in at least one of the three gynogenetic diploid crosses and were used to identify the centromeres. Placement of the centromere permitted identification of 16

metacentric linkage groups (Ots01 - Ots16) and 18 acrocentric linkage groups (Ots17 - Ots34), corresponding to the known Chinook salmon karyotype (Figure 2.1; Supplementary material S2.2). The small (p) chromosome arm of acrocentric chromosomes (Phillips & Rab 2001) is usually uncharacterized in mapping studies because there are often insufficient markers describing this region. In this study, we identified the small arm for three acrocentric chromosomes (Ots19, Ots20 and Ots33) (Figure 2.2). It is interesting to note that the linkage map sizes did not correlate with the sizes of the chromosomes, but the metacentric linkage groups (Ots01 to Ots16) were longer than the acrocentric linkage groups (Ots17 to Ots34). Ots19, Ots20 and Ots33 were the longest acrocentric linkage groups.

## Analysis of the properties of the Chinook salmon linkage groups

### *Frequency of recombination*

Examination of the distribution of markers across all chromosomes (Figure 2.3) showed a bias in marker placement. The greatest numbers of mapped loci were placed at the centromeres and towards the telomeres; the number of markers increased with increasing distance from the centromere regardless of the type of chromosome (Figure 2.3). This over-representation of markers at distal positions suggests that there is reduced recombination in the telomeres in the female.

### *Crossover frequency and interference*

We used one haploid family (Family A) with 46 progeny to examine the number of crossovers in 50 chromosome arms (2300 chromosome arms). We only observed 60 instances (2.6%) of double crossovers. The occurrences of double crossovers were not randomly distributed between chromosomes. The chromosomes with the highest frequency of double crossovers were acrocentric. Double crossovers occurred in Ots19, Ots20 and Ots33, for 10, 6 and 6 progeny respectively. However, the second crossover always occurred on the short arm of these three chromosomes. The remaining double crossovers occurred on 21 chromosome arms (Supplementary material S2.3). Finally the

frequency of double crossovers was not correlated to the number of duplicated loci on the linkage groups (t-test p-value: 0.42).

In gynogenetic diploid progeny, the maximum proportion of heterozygotes (MPH) at a locus in the telomeric region should be 0.67 if there is no interference and 1.00 if there is complete interference (Thorgaard *et al.* 1983). Here the average MPH for each chromosome arm was 0.90. The MPH ranged from 0.75 to 0.99, except for Ots11p where the MPH was 0.49. Here, it was only possible to genotype the non-duplicated loci in the gynogenetic diploids. Given that the distal regions from the centromere of 16 chromosome arms mainly comprised duplicated loci, we did not have full coverage of these arms in the gynogenetic diploid crosses. Indeed, we observed that all the arms with the lowest MPH (lower than 0.85) had a higher proportion of duplicated loci. Therefore, we concluded that the lower MPH observed for those arms was due to a lack of coverage with the gynogenetic diploid rather than absence of interference.

*Duplicated loci and homeologies*

A total of 799 duplicated loci detected by RAD sequencing were placed on the linkage map. The duplicated loci were not distributed uniformly between the chromosomes. We observed two categories of chromosome arms: those with very few duplicated loci (1 to 7, corresponding to 0.5% to 5.6% of all duplicated markers), and those with many duplicated loci (17 to 62 markers, corresponding to 15% to 61% of all markers). A total of 89.7% of the duplicated loci were located on 16 chromosome arms (Figure 2.4). Homeologies were inferred between 8 pairs of chromosome arms using 98 paralogs that were polymorphic at both loci (Table 2.3). Six of the homeologies had been identified by Naish *et al.* (submitted), but two were novel (Ots01q/06q and Ots07p/14p). Ots01q and Ots06q had the lowest number of duplicated markers: 85% and 76% of the markers mapping to these linkage groups respectively were not duplicated (Table 2.3). All other chromosome arms had between 35% and 51% duplicated loci, except for Ots04q that had the highest percentage of duplicated loci (61%). Finally the duplicated loci were not evenly distributed along the 16 chromosome arms that had a higher number of these loci. The regions distal from the centromere almost exclusively comprised duplicated loci (Figure 2.5).

**Comparison with rainbow trout**

A total of 40649 RAD loci have been described in rainbow trout (Miller *et al.* 2012). Over 50% of these loci (20,436) aligned uniquely to the non-duplicated markers in the Chinook salmon reference database. A total of 317 RAD loci mapped in both species, allowing us to confirm previously described homologies between the two species (Naish *et al.*, submitted) (Table 2.4). We confirmed the speculation in Naish *et al.* (submitted) that Ck04 is homologous to Omy11p and Omy09q, which is in agreement with the observations in Phillips *et al.* (submitted). These earlier studies showed that Ots16p and Ots16q are homologous to a portion of Omy11p and Omy9q respectively. Here we observed one marker from Omy11 on Ots16q (Figure 2.6). We were not able to compare the order of the RAD loci between the rainbow trout map and the Chinook salmon map because most of the markers polymorphic in both species on a linkage group mapped to a single position on the rainbow trout map.

## Discussion

Here, we aimed to improve genomic resources in Chinook salmon and study the chromosomal evolution of the species relative to that of other salmonids. We have developed a reference database of RAD loci for Chinook salmon comprising 48528 non-duplicated loci and 6409 known duplicated loci. We identified 7146 polymorphic RAD loci in three haploid families that were used to create a consensus map with a length of 4163cM. The map comprised 34 linkage groups, which were anchored to all Chinook salmon chromosome arms using microsatellite loci that have been physically mapped in previous studies. The placement of 799 duplicated loci on the linkage map revealed an uneven distribution of these loci across chromosomes, suggesting that homeologs diverged at different rates following whole genome duplication. Crossover frequency measured in one haploid family confirmed near complete interference across chromosome arms. Finally, the genome map supports previously published homologies among rainbow trout and Chinook salmon chromosome arms, but these homologies are supported using more

extensive data and centromere placement. Additionally, we used 5'-nuclease genotyping to map 149 SNP loci that are currently used in conservation and management studies.

The reference database of RAD loci in Chinook salmon is extensive and provides a resource against which future RAD sequences generated using *Sbf*I as a restriction enzyme can be aligned. Markers that were polymorphic in the mapping families have been annotated by chromosome arm. We attempted as far as possible to identify loci that had repeat units or were located in repeat regions using a series of screening tests based on self-alignment. However, the use of three haploid families might not have identified paralogs in the database that were not polymorphic. Therefore we recommend aligning initial sequences generated in future studies to the reference database, and treating the loci that are not uniquely lined to individual reads in this database as suspicious. It was not possible to systematically assign alleles to one paralog versus the other across the mapping families. Population genetic studies based on RAD loci in duplicated regions might be limited, and so we recommend using mapped ESTs or microsatellites to target these regions.

The Chinook salmon linkage map has 7304 markers that span all 34 chromosomes; this coverage is comparable to published maps in the salmonids. For example, the map for Atlantic salmon (*Salmo salar*) comprises 5650 SNPs (Lien *et al.* 2011), for rainbow trout 4563 RAD markers (Miller *et al.* 2012) and for sockeye salmon 3430 RAD markers (Everett *et al.* 2012). The present map had a size of 4163.9 cM, which is significantly larger than the first generation map available for Chinook salmon (2206.2 cM for the sex average map, Naish *et al*. submitted). The size of the current map is larger than other maps available for salmon species. The Atlantic salmon map (Lien *et al.* 2011) has a coverage of 2402.3 cM for females and 1746.2 cm for males. The most recently published female rainbow trout map size has a total length of 3600 cM (Guyomard *et al.* 2012). Genotyping errors (Hackett & Broadfoot 2003) and nonrandom missing values (Jorgenson *et al.* 2005) are two factors that can inflate map distances. Here, missing values were not randomly distributed across individuals ($\chi^2$ test for uniform distribution across individuals: pvalue $\sim$ 0 for each family). In addition, we previously highlighted the fact that SNP calls using STACKS might be biased against the detection of heterozygotes. While this constraint is not likely to be a concern for

non-duplicated loci in haploid families, it is likely to increase the genotyping error for the duplicated loci in these families. Therefore we suspect that map distances might be especially inflated in the regions with many duplicated loci, compared to the regions with few of these loci.

We were able to characterize each chromosome type and locate the centromere for 18 acrocentric chromosomes and 16 metacentric chromosomes, using over 3000 loci for three gynogenetic diploid lines. We believe that we were able to detect the p arms for some acrocentric chromosomes because of the high density of markers. Loci from the p arms of some acrocentric chromosomes may have been mapped in other dense salmon maps, but were not identified as such because the centromeres were not characterized. The centromeric regions were sometimes large. The percentage of heterozygote offspring was constrained by the number of progeny in each cross (84 to 93). Increasing the number of crosses, as well as the number of progeny, would facilitate the narrower placement of the centromere relative to the mapped markers. The location of the centromere allowed us to conclusively support previous findings on chromosome arm arrangement in Chinook salmon (Phillips & Rab 2001; Phillips *et al.* submitted). Additionally, we were able to confirm that Ots25 (Ck06) was acrocentric and Ots16 (Ck04) was metacentric as speculated in Naish *et al.* (submitted) and Phillips *et al.* (submitted). Six of the homeologies detected in this study had been previously identified in Chinook salmon, two were novel (Ots01q/06q and Ots07p/14p) and highly supported, and three previously identified homeologies were not observed here (Table 2.2). Eleven homeologies have therefore been identified to date for Chinook salmon.

Homeologous pairing can reduce divergence between such chromosomes through recombination (Allendorf & Thorgaard 1984; Wright *et al.* 1983). We observed two distinct categories of linkage group arms in Chinook salmon: those with almost no duplicated loci, and those with a high density of duplicated loci primarily located in distal regions from the centromere. This result suggests that the eight pairs of homeologs we identified are likely involved in homeologous pairing in Chinook salmon, all of which involved at least one metacentric chromosome. As we stated earlier, the results also suggest that divergence

rates of homeologs following whole genome duplication have not been uniform. Comparative mapping shows that the homeologous pairings we identified in Chinook salmon have also been shown in other salmonids (Danzmann *et al.* 2005; Danzmann *et al.* 2008; Gharbi *et al.* 2006; Lien *et al.* 2011; Sakamoto *et al.* 2000). One other study, that of Lien *et al.* (2011), also surveyed the proportion of a large number of duplicated markers across chromosomes and showed that this proportion also varied across linkage groups. Both studies identified eight pairs of chromosomes that had homeologies supported by duplicated loci. Three homeologous pairings were strongly supported by large numbers of markers in both studies (Ots03p/23 and Ssa02p/05q; Ots15q/17 and Ssa07q/17qa; Ots09q/27 and Ssa03q/06q). Two pairings had strong support in Chinook salmon but not in Atlantic salmon (Ots02p/32 and Ssa12qa/2q; Ots04q/12q and Ssa26/11qb), and one pairing had the opposite pattern (Ots11p/34 and Ssa04p/08q). Interestingly, two homeologous pairings in Chinook salmon (Ots07p/14p and Ssa17qa/16qb, Ots06q/Ots01q and Ssa01qa/18qa) were not supported by duplicated loci in Atlantic salmon in Lien *et al.* (2011). However, these were supported using sequence alignments within Atlantic salmon and with the threespine stickleback (*Gasterosteus aculeatus*). Two homeologies in Atlantic salmon (Ssa13qa/15qb and Ssa19qb/29) were not observed in the other species. There are three possible explanations for these last observations. The first is that the rates of divergence between homeologs might differ across salmonids. The second might simply be due to marker density – these reported differences might diminish with extensive sequencing. The third is methodological; the duplicated loci in Atlantic salmon were mapped using SNP markers with two alleles, whereas the present study mapped paralogs that had up to four alleles. Loci in Chinook salmon were considered duplicated if the paralogs had a maximum of three substitutions. Relaxing the alignment parameters, or using SNPs with more alleles in Atlantic salmon, might permit identification of duplicated loci that have higher divergences, and by extension, a greater number of homeologs.

Our data supports the hypothesis of near complete interference in Chinook salmon, where we observed very few occurrences of double crossovers and a maximum proportion of heterozygotes close to one for all chromosomes in the gynogenetic diploid lines. This result agrees with previous studies (e.g. Guyomard *et al.* 2006) but is supported by a much higher

number of markers and recombination events observed. We also observed that the frequency of recombination was reduced in the telomeric regions in females, as suggested in Moen *et al.* (2004), Danzmann *et al.* (2008) or Lien *et al.* (2011). The higher proportion of markers mapping in the telomeric regions suggest that the map created in this study covers the entire genome, but that the order of the markers in the telomeres is likely not fully resolved, but could be by mapping male meiosis. Indeed, the male-based map based on RAD markers in rainbow trout (Miller *et al.* 2012) showed that most recombination events occurred at the telomeres.

Comparisons of the Chinook salmon linkage map and the rainbow trout linkage map (Miller *et al.* 2012) confirmed all rearrangements and homologies previously identified (Naish *et al.* submitted, Phillips *et al.* submitted). Our data also supports the fact that Ots16 (Ck04) comprises a fusion between a fragment of one chromosome arm from a metacentric chromosome, Omy11p, and another, Omy9q. However, the higher resolution on the current map shows that markers from Omy11p are found on both arms of Ots16, suggesting that there may have been a centromeric inversion on Ots16. The number of RAD loci shared between Chinook salmon and rainbow trout suggests that determining chromosome evolution across salmonids is increasingly feasible as more species are mapped using RAD loci. The identification of the centromeres permitted the accurate characterization of all the chromosome arms, compared to the earlier studies on Chinook salmon. Since chromosome arms are mainly conserved across species, this map can also be used for genome-wide studies in other salmon species.

Here, we developed two major genomic resources for Chinook salmon: a reference database of RAD loci and a very dense linkage map anchored to the chromosomes, where arms have been identified by placement of the centromeres. These resources will facilitate genome-wide studies in Chinook salmon, such as genome scans or QTL mapping, as well as studies in related species.

## Acknowledgements

This work was funded by a grant from the Washington Sea Grant Program, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award awarded to KAN and JES. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of their sub-agencies. We thank Linda Park, Robin Waples, Paul Moran, David Teel and Michael Ford, for giving us access to the samples from the Columbia River used for the database. We thank the Cle Elum Supplementation and Research Facility (CESRF) for allowing us to create and raise the haploid crosses in their facilities. We thank Todd Seamons, Ken Warheit, Sewall Young from the Washington Department of Fish and Wildlife, for helping with haploid crosses. We thank James Meyers and Jeff Hard for the gynogenetic diploid and diploid crosses. We thank Ruth Phillips for her invaluable advice throughout. We thank Isadora Jimenez-Hidalgo, Melissa Baird, Carita Pascal, Daniel Drinan, Miyako Kodama and Kotaro Ono for their help with lab work, data analysis and paper review.

## Tables

| Parental genotype | | Segregation ratio expected for each paralog in a haploid cross | | Segregation ratio expected in the offspring for a haploid | Marker(s) mapped in this study |
|---|---|---|---|---|---|
| **paralog 1** | **paralog 2** | **Paralog 1** | **Paralog 2** | | |
| aa | bb | all a | all b | all ab | none |
| ab | ab | 0.5 a; 0.5 b | 0.5 a; 0.5 b | 0.25 aa; 0.5 ab; 0.25 bb | none |
| aa | ab | all a | 0.5 a; 0.5 b | 0.5 aa; 0.5 ab | Paralog 2 |
| aa | bc | all a | 0.5 b; 0.5 c | 0.5 ab; 0.5ac | Paralog 2 |
| ab | ac | 0.5 a; 0.5 b | 0.5 a; 0.5 c | 0.25 aa; 0.25 ac; 0.25 ab; 0.25 bc | Paralogs 1 and 2 |
| ab | cd | 0.5 a; 0.5 b | 0.5 c; 0.5 d | 0.25 ac; 0.25 ad; 0.25 bc; 0.25 bd | Paralogs 1 and 2 |

Table 2.1 – Types of duplicated loci encountered in this study, expected segregation ratio per paralog and expected segregation ratio when both paralogs are analyzed as a single locus, which is the case in this study. The type of duplicated markers is inferred from the observed segregation ratio and the alleles observed in the offspring generation.

| | Non duplicated RAD loci | OPP | BPP | SNP | Total | Map length (cM) |
|---|---|---|---|---|---|---|
| **Family A** | 3001 | 324 | 62 | 98 | **3485** | **2834.9** |
| **Family B** | 2922 | 245 | 32 | 92 | **3291** | **3099.6** |
| **Family C** | 3011 | 230 | 32 | - | **3273** | **2991.8** |
| **Consensus Map** | **6352** | **603** | **196** | **153** | **7304** | **4163.9** |

Table 2.2 – Number of Loci mapped and map length for each haploid family and for the consensus map. Four types of markers were used: non duplicated RAD loci, duplicated RAD loci for which only one paralog was polymorphic (OPP) or both paralogs were polymorphic (BPP), and SNP loci.

| **Chinook salmon homeologs** | **Number of marker pairs supporting homeology in this study** | **Homeology in Atlantic salmon** | **Number of markers and type of support for the homeology in Lien *et al.* (2011)** |
|---|---|---|---|
| Ots01q - Ots06q | 11 | Ssa01qa-Ssa18qa | BLAST |
| Ots02p - Ots32 | 15 | Ssa02q - Ssa12qa | 1 MSV5 |
| Ots03p - Ots23 | 11 | Ssa02p - Ssa05q | 39 MSV5 |
| Ots04q - Ots12q | 19 | Ssa26 - Ssa11qb | 2 MSV5 |
| Ots07p - Ots14p | 17 | Ssa17qb - Ssa16qb | BLAST |
| Ots09q - Ots27 | 9 | Ssa03q - Ssa06q | 7 MSV5 |
| Ots11p - Ots34 | 3 | Ssa04p - Ssa08q | 14 MSV5 |
| Ots15q - Ots17 | 13 | Ssa07q - Ssa17qa | 33 MSV5 |

Table 2.3 – Homeologous chromosome pairs identified for Chinook salmon, number of pairs of markers supporting the homeologies, corresponding homeologs in Atlantic salmon from Phillips *et al.* (2009) and Phillips *et al.* (submitted) and type of support for the homeologies in Lien *et al.* (2011). Support for the homeologies in Lien *et al.* (2011): duplicated SNP loci with both paralog polymorphic (MSV5) or alignment-based using BLAST within Atlantic salmon or with stickleback. Minor differences between the homeologies identified in Phillips *et al.* (2009) and Lien *et al.* (2011) are in brackets.

| **Chinook chromosome** (current study, Phillips *et al.* submitted) | **Chinook linkage group** (Naish *et al.* submitted) | **Rainbow trout Chromosome** (Phillips *et al.* 2006) | **Rainbow trout linkage group** (Miller *et al.* 2012) | **Number of markers supporting homology** |
|---|---|---|---|---|
| Ots01p | Ck13 | Omy04p | WS01 | 7 |
| Ots01q | | Omy23 | WS27 | 6 |
| Ots02p,q | Ck12 | Omy17p,q | WS23 | 13 |
| Ots03p,q | Ck05 | Omy03p,q | WS06 | 7 |
| Ots04p,q | Ck08 | Omy06p,q | WS13 | 15 |
| Ots05p | Ck11 | Omy08p | WS05 | 5 |
| Ots05q | | Omy05q | WS03 | 18 |
| Ots06p,q | Ck17 | Omy01p,q | WS20 | 17 |
| Ots07p,q | Ck16 | Omy07p,q | WS25 | 5 |
| Ots08p,q | Ck14 | Omy25 | WS24 | 19 |
| Ots09p,q | Ck02 | Omy12p,q | WS21 | 13 |
| Ots10p | Ck20 | Omy09p | WS04 | 2 |
| Ots10q | | Omy08q | WS05 | 8 |
| Ots11p,q | Ck15 | Omy19p,q | WS22 | 7 |
| Ots12p | Ck18 | Omy11p+q | WS07 | 8 |
| Ots12q | | Omy26 | WS28 | 2 |
| Ots13p | Ck07 | Omy27 | WS16 | 8 |
| Ots13q | | Omy18 | WS19 | 4 |
| Ots14p | Ck10 | Omy18p | Ø | 11 |
| Ots14q | | Omy24 | WS17 | 11 |
| Ots15p,q | Ck23 | Omy21p,q | WS26 | 9 |
| Ots16p | Ck04 | Omy11p | WS07 | 3 |
| Ots16q | | Omy09q | WS04 | 3 |
| Ots17 | Ck01 | Omy15 | WS12 | 8 |
| Ots18 | Ck33 | Omy04q | WS01 | 6 |

| | | | | |
|---|---|---|---|---|
| Ots19 | Ck22 | Omy02q | WS18 | 12 |
| Ots20 | Ck28 | Omy05p | WS03 | 11 |
| Ots21 | Ck09 | Omy14q | WS10 | 6 |
| Ots22 | Ck34 | Omy16q | WS08 | 6 |
| Ots23 | Ck25 | Omy02p | WS18 | 2 |
| Ots24 | Ck27 | Omy16p | WS08 | 8 |
| Ots25 | Ck06 | Omy20p+q | WS15 | 6 |
| Ots26 | Ck21 | Omy22 | WS02 | 9 |
| Ots27 | Ck31 | Omy13q | WS29 | 2 |
| Ots28 | Ck24 | Omy28 | WS09 | 9 |
| Ots29 | Ck03 | Omy15 | WS12 | 5 |
| Ots30 | Ck29 | Omy10p | WS14 | 11 |
| Ots31 | Ck26 | Omy14p | WS10 | 6 |
| Ots32 | Ck30 | Omy13p | WS29 | 4 |
| Ots33 | Ck19 | OmySex | WS11 | 7 |
| Ots34 | Ck32 | Omy10q | WS14 | 3 |

Table 2.4 – Homologies between Chinook salmon and rainbow trout chromosome arms, corresponding linkage groups in Naish *et al.* (submitted) and Miller *et al.* (2012) and number of RAD markers supporting the homologies in this study as identified in Phillips *et al.* (submitted) and Phillips *et al.* (2009). There was no marker in common between Ots14p and Omy18p.

## Figures

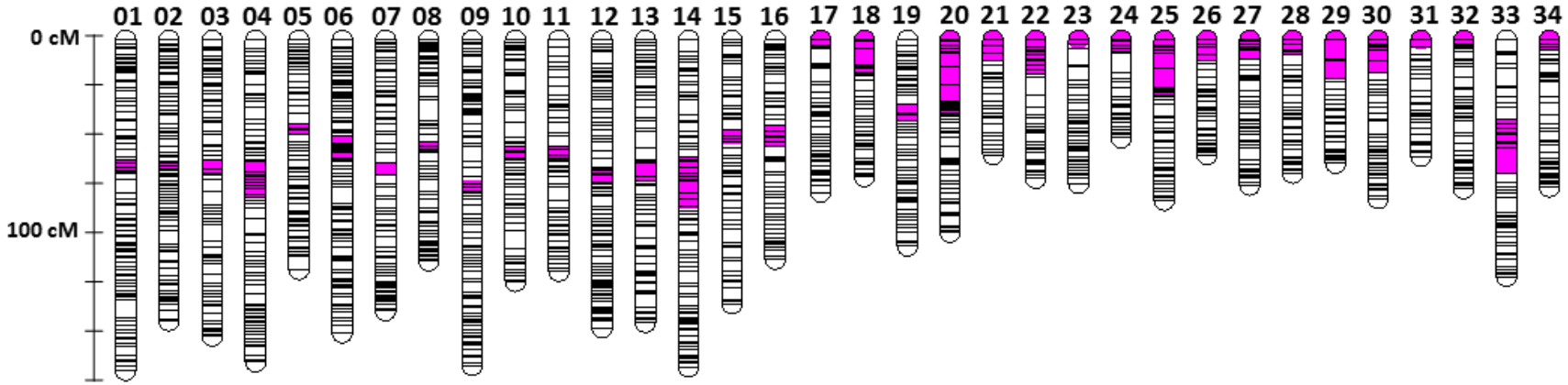


Figure 2.1 – Consensus Chinook salmon female linkage map – 34 linkage groups corresponding to the 34 Chinook salmon chromosomes. Ots01 to Ots16 are metacentric; Ots17 to Ots34 are acrocentric. The size of each linkage group varies from 58 to 173.2 cM. Each line corresponds to the location of one or more markers. The centromere is represented in pink. All the chromosomes are oriented with the shorter arm (p arm) before the centromere, longer arm (q arm) after the centromere.

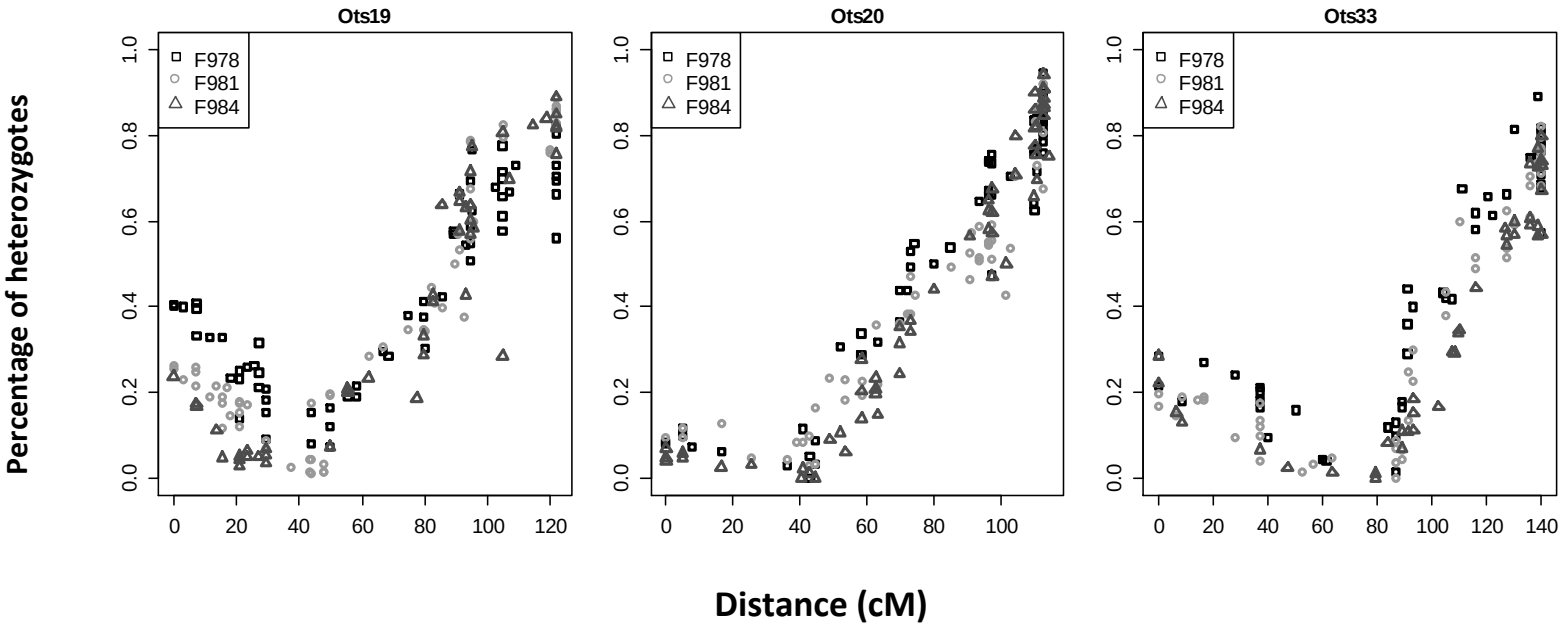


Figure 2.2 – Percentage of heterozygous offspring in the gynogenetic diploid crosses along three acrocentric chromosomes where the p arm has been identified: Ots19, Ots20 and Ots33. On the x axis the distances are oriented from the p arm. Three gynogenetic crosses were used (F978, F981 and F984). The centromere is located where the percentage of heterozygous offspring is about zero.

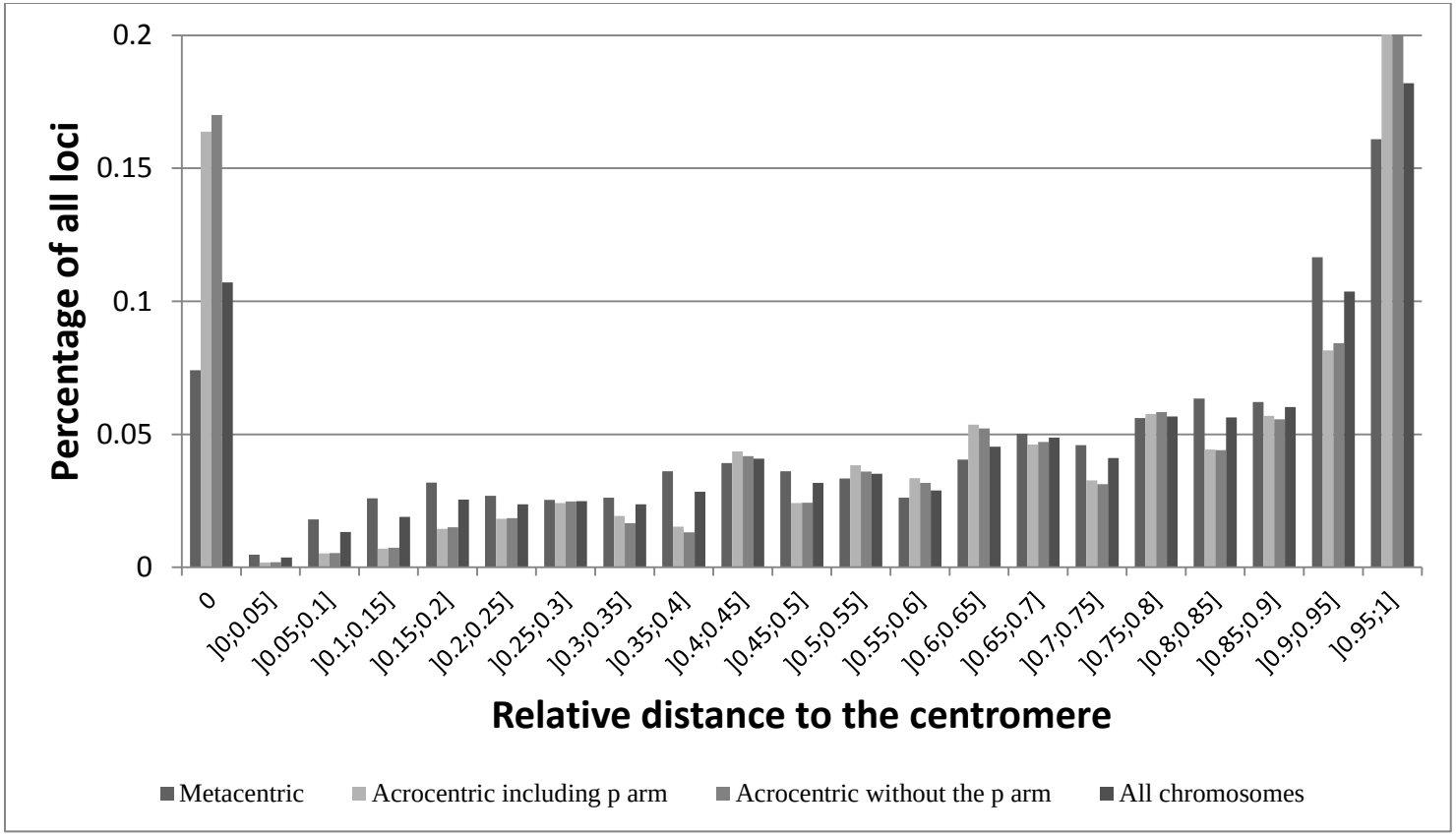


Figure 2.3 – Marker distribution across all chromosome arms, examined seperately for all chromosomes, metacentric chromosomes and acrocentric chromosomes (including the p arm or not for Ots19, Ots20 and Ots 33). Distances on the x axis are represented as the relative distance from the center of the centromeric region.

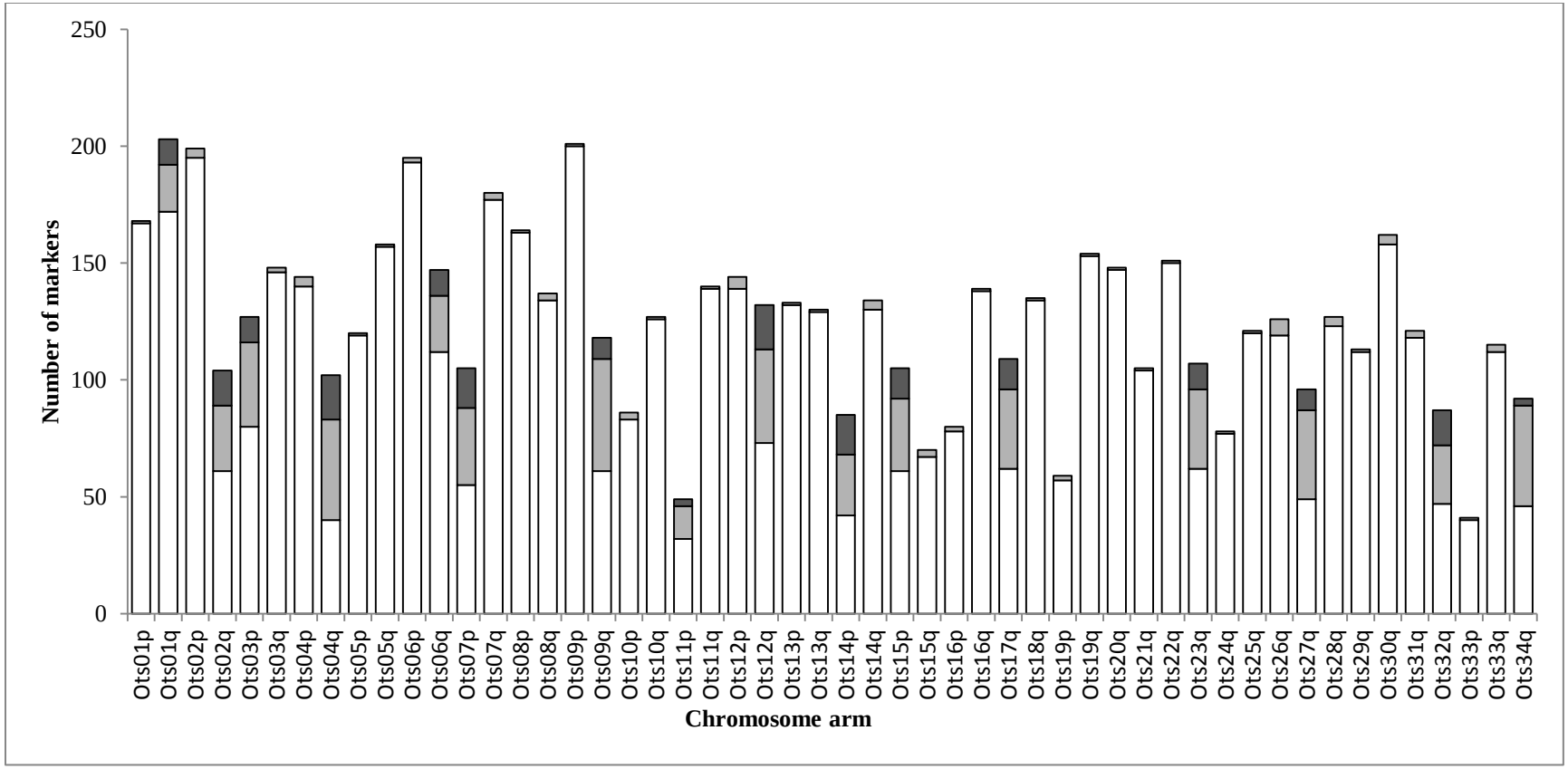


Figure 2.4 – Number of markers on each linkage group, designated by chromosome arm. Non duplicated loci (RAD loci or SNP loci) are represented by the white bars; Duplicated loci are represented by the light grey bars (loci with only one paralog polymorphic) or dark grey bars (both paralogs polymorphic).

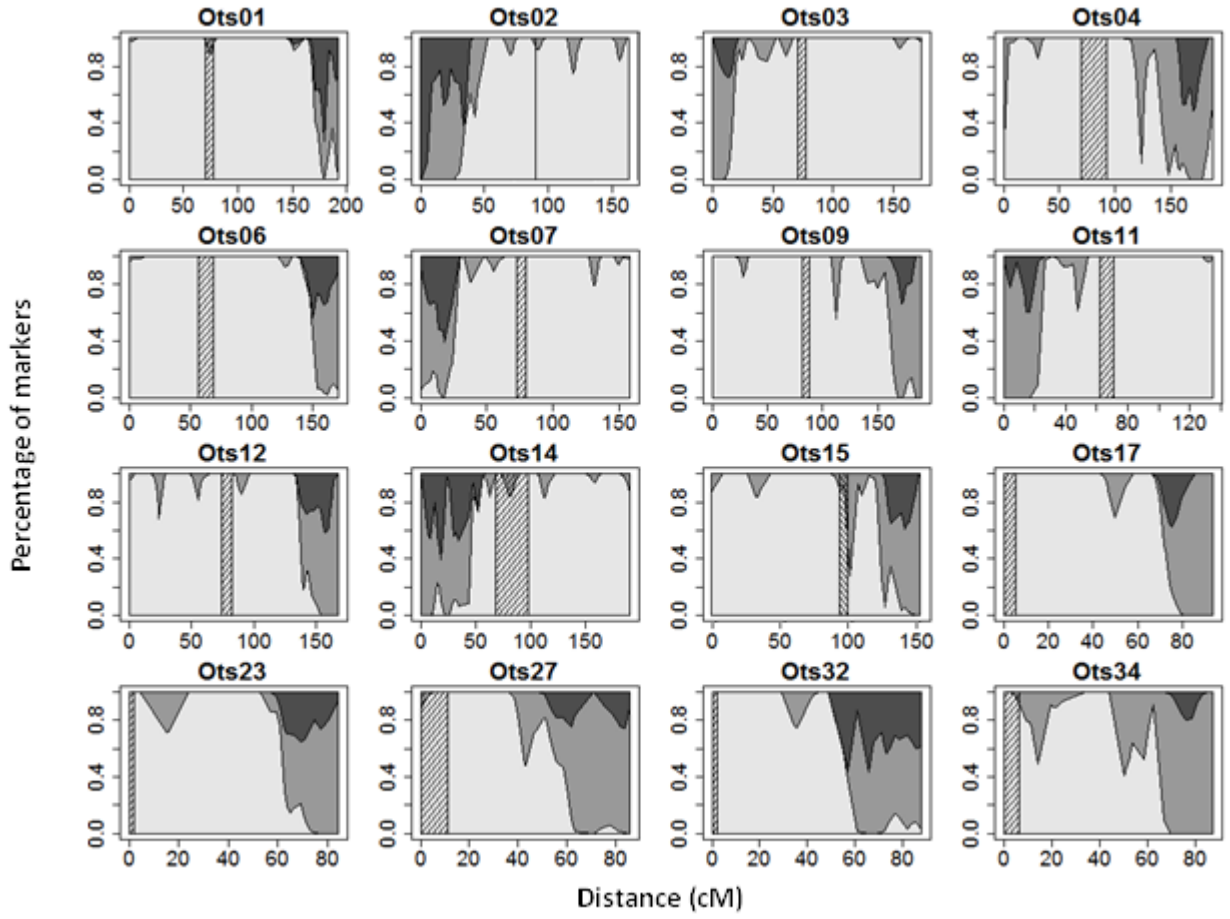


Figure 2.5 – Proportion of duplicated and non-duplicated loci along the 16 linkage groups (denoted by chromosome number) with a high number of duplicated loci. Non duplicated loci are represented in white; Duplicated loci are represented in grey (loci with one paralog polymorphic) or in dark grey (loci with both paralogs polymorphic). The centromere is represented by the cross-hatched area. All chromosomes are oriented with the short arm (p) on the left where relevant.

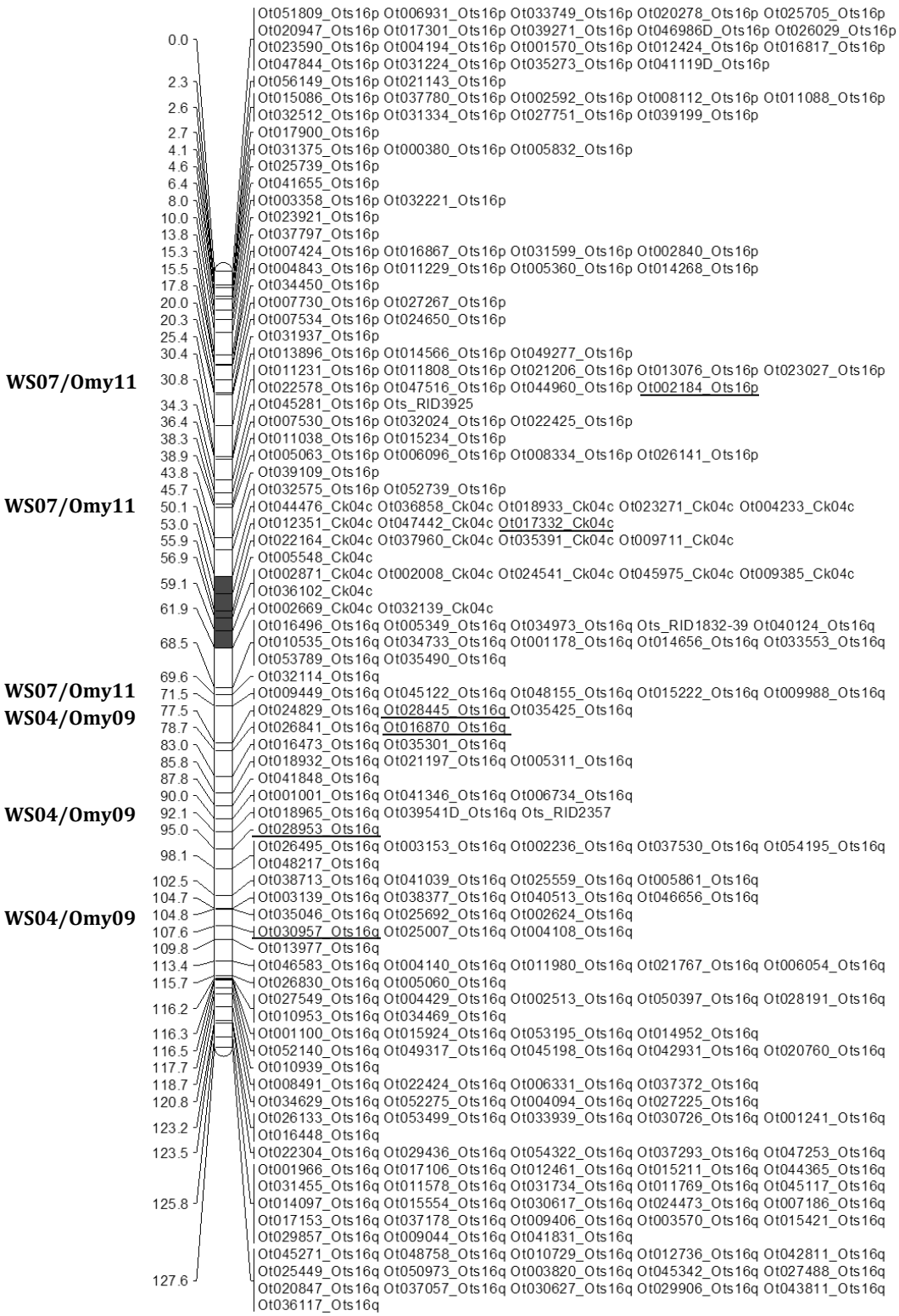


Figure 2.6 – Linkage map for Ots16, denoting loci that are homologous with rainbow trout. Loci in common between the two species are underlined and the position on the rainbow trout map is indicated on the left of the chromosome (WS: linkage groups from Miller *et al.* 2012; Omy: rainbow trout chromosomes). The centromere is represented in grey. The chromosome is oriented with the p arm on top and the q arm at the bottom. Distances are in centiMorgans.

## Supplementary Material

**S2.1:** Linkage maps

**S2.1.1**: Consensus haploid map

**S2.1.2**: Diploid map

**S2.2**: Percentage of heterozygous offspring in the gynogenetic diploid crosses along all chromosomes. On the x axis the distances are oriented from the p arm. Three gynogenetic crosses were used (F978, F981 and F984). The centromere is located where the percentage of heterozygous offspring is about zero.

**S2.3**: Number of individuals from Family A (n = 46) with double crossovers (DCO) for each chromosome arm. *: A crossover occurred in the centromere of Ots12. It was not possible to attribute the crossover to one arm or the other and so the range of individuals with double crossovers for each arm was estimated

## References


Abadia-Cardoso A, Clemento AJ, Garza JC (2011) Discovery and characterization of single-nucleotide polymorphisms in steelhead/rainbow trout, Oncorhynchus mykiss. *Molecular Ecology Resources* **11**, 31-49.

Allendorf FW, Danzmann RG (1997) Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* **145**, 1083-1092.

Allendorf FW, Seeb JE, Knudsen KL, Thorgaard GH, Leary RF (1986) Gene-centromere mapping of 25 loci in Rainbow-trout. *Journal of Heredity* **77**, 307-312.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. B.J. T). Plenum Press, New York.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Amish SJ, Hohenlohe PA, Painter S*, et al.* (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources* **12**, 653-660.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Bradbury IR, Hubert S, Higgins B*, et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, n/a-n/a.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Campbell NR, Narum SR (2008) Identification of novel single-nucleotide polymorphisms in Chinook salmon and variation among life history types. *Transactions of the American Fisheries Society* **137**, 96-106.

Chong ZC, Ruan J, Wu CI (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics* **28**, 2732-2737.

Cichon S, Craddock N, Daly M*, et al.* (2009) Genomewide Association Studies: History, Rationale, and Prospects for Psychiatric Disorders. *American Journal of Psychiatry* **166**, 540-556.

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**, 169-196.

Danzmann RG (2005) LINKMFEX: linkage analysis software for diploid and tetraploid outcrossed mapping panels. *Aquaculture* **247**, 10-10.

Danzmann RG, Cairney M, Davidson WS*, et al.* (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**, 1037-1051.

Danzmann RG, Davidson EA, Ferguson MM*, et al.* (2008) Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). *Bmc Genomics* **9**.

Danzmann RG, Gharbi K (2001) Gene mapping in fishes: a means to an end. *Genetica* **111**, 3-23.

Davey JW, Hohenlohe PA, Etter PD*, et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Everett MV, Miller MR, Seeb JE (2012) Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing. *Bmc Genomics* **13**.

Gharbi K, Gautier A, Danzmann RG*, et al.* (2006) A linkage map for brown trout (Salmo trutta): Chromosome homeologies and comparative genome organization with other salmonid fish. *Genetics* **172**, 2405-2419.

Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. *Nature Genetics* **3**, 266-272.

Good TP, Waples RS, Adams P (2005) Updated status of federally listed ESUs of West Coast salmon and steelhead eds. Good TP, Waples RS, Adams P). U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-66.

Guyomard R, Boussaha M, Krieg F, Hervet C, Quillet E (2012) A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. *Bmc Genetics* **13**.

Guyomard R, Mauger S, Tabet-Canale K*, et al.* (2006) A Type I and Type II microsatellite linkage map of Rainbow trout (Oncorhynchus mykiss) with presumptive coverage of all chromosome arms. *Bmc Genomics* **7**.

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90**, 33-38.

Hecht BC, Thrower FP, Hale MC, Miller MR, Nichols KM (2012) Genetic Architecture of Migration-Related Traits in Rainbow and Steelhead Trout, Oncorhynchus mykiss. *G3-Genes Genomes Genetics* **2**, 1113-1127.

Hess JE, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Molecular Ecology Resources* **11**, 137-149.

Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *Plos Genetics* **6**.

Houston RD, Davey JW, Bishop SC*, et al.* (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *Bmc Genomics* **13**.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3-17.

Johnson SL, Gates MA, Johnson M*, et al.* (1996) Centromere-linkage analysis and consolidation of the zebrafish genetic map. *Genetics* **142**, 1277-1288.

Jorgenson E, Tang H, Gadde M*, et al.* (2005) Ethnicity and human genetic linkage maps. *American Journal of Human Genetics* **76**, 276-290.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**.

Lien S, Gidskehaug L, Moen T*, et al.* (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**.

Lindner KR, Seeb JE, Habicht C*, et al.* (2000) Gene-centromere mapping of 312 loci in pink salmon by half-tetrad analysis. *Genome* **43**, 538-549.

Magwire MM, Fabian DK, Schweyen H*, et al.* (2012) Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster. *Plos Genetics* **8**.

Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas* **144**, 78-79.

Matala AP, Narum SR, Young W, Vogel JL (2012) Influences of Hatchery Supplementation, Spawner Distribution, and Habitat on Genetic Structure of Chinook Salmon in the

South Fork Salmon River, Idaho. *North American Journal of Fisheries Management* **32**, 346-359.

McClelland EK, Naish KA (2008) A genetic linkage map for coho salmon (Oncorhynchus kisutch). *Animal Genetics* **39**, 169-179.

Miller MR, Brunelli JP, Wheeler PA*, et al.* (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology* **21**, 237-249.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248.

Moen T, Hayes B, Baranski M*, et al.* (2008) A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *Bmc Genomics* **9**.

Moen T, Hoyheim B, Munck H, Gomez-Raya L (2004) A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes. *Animal Genetics* **35**, 81-92.

Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics & Development* **12**, 657-663.

Nichols KM, Edo AF, Wheeler PA, Thorgaard GH (2008) The genetic basis of smoltification-related traits in Oncorhynchus mykissi. *Genetics* **179**, 1559-1575.

Ohno S (1970) Enormous diversity in genome sizes of fish as a reflection of natures extensive experiments with gene duplicaiton. *Transactions of the American Fisheries Society* **99**, 120- &.

Phillips R, Rab P (2001) Chromosome evolution in the Salmonidae (Pisces): an update. *Biological Reviews* **76**, 1-25.

Phillips RB, Keatley KA, Morasch MR*, et al.* (2009) Assignment of Atlantic salmon (Salmo salar) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (Oncorhynchus mykiss). *Bmc Genetics* **10**.

Sakamoto T, Danzmann RG, Gharbi K*, et al.* (2000) A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates. *Genetics* **155**, 1331-1345.

Seeb JE, Pascal CE, Grau ED, et al. (2011) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. Molecular Ecology Resources 11, 335-348.

Shendure J, Ji HL (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145.

Slijepcevic P (1998) Telomeres and mechanisms of Robertsonian fusion. *Chromosoma* **107**, 136-140.

Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005a) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology* **14**, 4193-4203.

Smith CT, Seeb JE, Schwenke P, Seeb LW (2005b) Use of the 5 '-nuclease reaction for single nucleotide polymorphism genotyping in Chinook salmon. *Transactions of the American Fisheries Society* **134**, 207-217.

Smith CT, Templin WD, Seeb JE, Seeb LW (2005c) Single Nucleotide Polymorphisms (SNPs) provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* **25**, 944-953.

Templin WD, Seeb JE, Jasper JR, Barclay AW, Seeb LW (2011) Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. *Molecular Ecology Resources* **11**, 226-246.

Thorgaard GH, Allendorf FW, Knudsen KL (1983) Gene-centromere mapping in Rainbow-trout - High interference over long map distances. *Genetics* **103**, 771-783.

Tsumura Y, Uchiyama K, Moriguchi Y, Ueno S, Ihara-Ujino T (2012) Genome scanning for detecting adaptive genes along environmental gradients in the Japanese conifer, Cryptomeria japonica. *Heredity* **109**, 349-360.

Waples RS (1995) Evolutionarily Significant Units and the Conservation of Biological Diversity under the Endangered' Species Act. *American Fisheries Socicty Symposium* **17**, 8-27.

Wright JE, Johnson K, Hollister A, May B (1983) Meiotic models to explain classical linkage, pseudolinkage, and chromosome-pairing in tetraploid derivative Salmonid genomes. *Isozymes-Current Topics in Biological and Medical Research* **10**, 239-260.

Wu YH, Close TJ, Lonardi S (2011) Accurate Construction of Consensus Genetic Maps via Integer Linear Programming. *Ieee-Acm Transactions on Computational Biology and Bioinformatics* **8**, 381-394.

# Chapter 3 – Evaluation of the role of adaptive evolution of Chinook salmon (*Oncorhynchus tshawytscha*) in the Columbia River basin, USA.

Marine S.O. Brieuc and Kerry A. Naish
School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA

## Abstract

Geological events, including glaciations followed by ice melting and floods, shaped the population structure of salmon species in the Pacific Northwest. Chinook salmon (*Oncorhynchus tshawytscha*) populations in the Columbia River basin are believed to be derived from two main glacial refugia. However the role of adaptive divergence in the evolution of the species, specifically across the two main evolutionary scales, pre- and post-glaciation, is still unknown. High throughput DNA sequencing technologies and dense linkage maps now permit genome-wide detection of genomic regions that may have undergone adaptive evolution in non-model species. Using eleven populations from the Columbia River basin, and three populations from Puget Sound as an outgroup, we examined the role of adaptive divergence in the evolution of Chinook salmon in the Columbia River basin, the distribution of regions of elevated divergence across the genome, and determined whether there was molecular evidence of parallel evolution of run timing across populations. We identified 14105 RAD (restriction site associated DNA) loci in this system. We identified three main lineages in the Columbia River basin. The Interior Columbia Su/F and Lower Columbia Sp/F lineages were less differentiated from each other than from the Interior Columbia Sp/Su, supporting the hypothesis of colonization from two glacial refugia. Using approaches that detected outlier loci associated with adaptive divergence, and analyses that permitted hierarchical partitioning of these loci between the three lineages, we determined that 148 and 153 loci were putatively involved in pre- and post-glaciation adaptive divergence respectively. We identified regions of the genome with high divergence using a pre-existing Chinook linkage map, most of which were located in distal regions from the centromere. Although some regions of elevated divergence were

observed between the two evolutionary scales, many appeared to be specific to pre- or post-glaciation divergence. Finally, we did not find strong support for parallel evolution of run timing at the molecular level, but we did identify some loci that accurately predicted of run timing, two of which mapped to the same position on the Chinook linkage map. In this study we improved our understanding of Chinook salmon evolution. We also demonstrated the usefulness of dense linkage maps in identifying regions of the genome that may have been involved in adaptive evolution.

## Introduction

Past geological events contributed to the current population structure of Pacific salmon on the West Coast of the United States and Canada (Waples *et al.* 2001; Waples *et al.* 2008). The last glaciation event in the Pleistocene in particular reduced the habitat available for Pacific salmon, and only populations living in ice-free regions persisted. Following glaciation, ice melting and numerous floods reshaped the landscape, providing new habitat for colonization from these refugia (Utter *et al.* 1989; Waples *et al.* 2008). Following colonization, coho salmon (*Oncorhynchus kisutch*), steelhead trout (*O. mykiss*) and Chinook salmon (*O. tshawytscha*) occupied the widest geographical range in the Pacific Northwest and Canada compared to the other species of Pacific salmon (sockeye - *O. nerka*, chum - *O. keta* and pink salmon - *O. gorbuscha*) (Gustafson *et al.* 2007). Pacific salmon are of considerable conservation interest and in the contemporaneous United States, species have been subdivided into conservation units aimed recognize their evolutionary distinctness (Evolutionary Significant Units, ESUs). It is estimated that up to 30% of all conservation units of Pacific salmon have been extirpated following Euro-American contact because of natural and anthropogenic disturbances (Gustafson *et al.* 2007).

Population genetic studies of Chinook salmon across the Columbia River Basin suggest that the extant populations are descended from dispersers from two refugia, one northern and one southern (Beacham *et al.* 2006; Utter *et al.* 1989; Waples *et al.* 2001). However, the designated number of lineages has varied across studies. For example, Waples *et al.* (2004) and Beacham *et al.* (2006) identified four major lineages using allozyme and microsatellite

markers respectively: Lower Columbia, Willamette, Interior Columbia summer and fall-run, and Interior Columbia spring-run group. Subsequent studies with microsatellite markers determined that the populations from the Willamette River belonged to the Lower Columbia lineage (Moran *et al.* 2012; Narum *et al.* 2010; Seeb *et al.* 2007), and described a total of three lineages.

Chinook salmon exhibit a wide variety of life history traits, such as spawn timing, run timing or age at smoltification (Quinn 2005), many of which appear to reflect adaptation to local environmental conditions. Chinook salmon are anadromous. They migrate to sea in their first year or in their second year (referred to as ocean- and stream- type respectively). After several years spent at sea, adults return to their spawning grounds. The time at which they enter freshwater is referred to as the run timing. In the Columbia River it has recently been recommended that the three lineages detected using genetic analyses be characterized by their adult return timing (Moran *et al.* 2012; Waples *et al.* 2004). In the Columbia River Basin, the interior stream-type lineage returns to freshwater in spring and summer (Interior Columbia Su/F) whereas the interior ocean-type returns in summer and fall (Interior Columbia Sp/Su). Finally, the populations in the coastal lineage return to freshwater in spring or fall. Both of the interior lineages are sympatric but reproductively isolated (Beacham *et al.* 2006; Narum *et al.* 2007; Waples *et al.* 2004). Divergence of these two lineages is estimated to predate the last glaciation event (Waples *et al.* 2001) and to be the result of colonization from the two different refugia (Beacham *et al.* 2006; Waples *et al.* 2008).

Although population structure of Chinook salmon in the Columbia River basin has been extensively studied, the role of adaptive divergence in the evolution of the species in this system and across the range is still unknown. While it is widely recognized that life history diversity in salmon reflects their adaptation to diverse environments (Fraser *et al.* 2011; Taylor 1991), it has also been shown that Chinook salmon has the capacity to evolve rapidly (Quinn *et al.* 2000; Waples *et al.* 2004). Individuals from the Sacramento River, CA, USA, were introduced to the Waikiki River in New Zealand in the early 1900s. Many adjacent rivers were rapidly colonized following this introduction. Despite low genetic

diversity, population differentiation was evident in New Zealand after only thirty generations (Quinn *et al.* 1996). This rapid divergence was correlated with the evolution of life history traits such as run timing and freshwater growth rate (Quinn *et al.* 2001; Quinn *et al.* 2000). Rapid divergence of Chinook salmon has also been observed following its introduction in the Great Lakes, but the patterns observed are not as striking, possibly because of the lower of number of generations since introduction and the yearly supplementation from hatchery stocks (Weeder *et al.* 2005). In the Columbia River populations, mapping of adult return timing suggests that there has been parallel evolution in this trait across lineages (Waples *et al.* 2004). Parallel phenotypic evolution provides a unique opportunity to examine if the same evolutionary pathways were taken to obtain similar phenotypes (Conte *et al.* 2012; Elmer& Meyer 2011; Wood *et al.* 2005).

Genome-wide approaches, such as genome scans, have successfully been used to identify regions of the genome associated with adaptation (Bradbury *et al.* 2013; Hemmer-Hansen *et al.* 2013; Hohenlohe *et al.* 2010). Such approaches require a very large number of markers and, in species whose genomes have not been sequenced, the use of genetic linkage maps (e.g. Bourret *et al.* 2013; Tsumura *et al.* 2012). A very dense linkage map spanning the entire genome of Chinook salmon was recently created (Brieuc *et al.* Chapter 2). This map comprises over 7000 RAD (restriction site associated DNA) (Baird *et al.* 2008; Miller *et al.* 2007) markers representing all 34 chromosomes for Chinook salmon, and chromosome arms were identified by mapping the centromeres. One of the main challenges with using RAD sequencing in salmonids that have a residual duplicated genome due to a whole genome duplication event that occurred 25 to 100 million years ago (Allendorf& Thorgaard 1984) is identifying loci *de novo*. However, a reference database of mainly non duplicated RAD loci was recently created as part of these mapping efforts (Brieuc *et al.* Chapter 2), therefore greatly facilitating identification of validated loci. This resource can now be applied to genome-wide analyses of evolutionary divergence in Chinook salmon. Therefore we are poised to identify the relative role of adaptive processes involved in population divergence in the Columbia River compared to neutral evolution and to determine whether there is molecular evidence for parallel evolution.

The overall aim of this study was to determine the role of adaptive evolution in Chinook salmon population divergence over different evolutionary periods in the Columbia River Basin, and to examine the link between this divergence and trait diversity. We used eleven populations from the Columbia River basin, and three populations from the Puget Sound ESU as an outgroup, to assess overall population structure in the Columbia River. We then determined whether there was any evidence for elevated divergence consistent with adaptation across lineages in the Columbia River. Our last objective was to identify potential loci or regions of the genome that are linked to a life history trait, run timing, and to assess whether there is molecular evidence for parallel evolution in this trait.

## Material and Methods

### Sample Collection and RAD sequencing

A total of 570 individuals were sampled from 11 populations representing the three main lineages in the Columbia River basin and from 3 populations in Puget Sound (Table 3.1, Figure 3.1). All samples were collected between 1994 and 2009, stored as dried fin clips or in ethanol at the Northwest Fisheries Science Center (NOAA, Seattle, WA), at the Columbia River Inter-Tribal Fish Commission (Hagerman, ID) or at the University of Washington (Seattle, WA). Genomic DNA was extracted from each individual using the DNeasy extraction kit (QIAGEN, Valencia, CA, USA) following the manufacturer's recommendations. Libraries of 6 to 12 individuals were prepared for Restriction Associated site DNA (RAD) sequencing with the restriction enzyme *SbfI* following the protocol described in Etter *et al.* (2011). The libraries were subsequently pooled to make up 36 individuals per lane and sequenced on an Illumina GAIIx or on an Illumina HiSeq sequencer, producing 100-nucleotide single-end reads. The last 20 nucleotides were subsequently removed for quality purposes.

Loci were identified by aligning the reads to the reference database of non-duplicated *Sbf*I-RAD loci created for Chinook salmon (Brieuc *et al.* Chapter 2) using BOWTIE (Langmead *et al.* 2009). We removed from the analysis all reads that aligned to more than one locus in the

baseline, and the corresponding locus in the baseline, because these loci were putative paralogs that were not identified in the previous mapping study. Polymorphisms were identified using a maximum likelihood approach implemented in STACKS (Catchen *et al.* 2011). Genotypes were corrected to limit the bias against the detection of heterozygotes due to unequal sequencing depth between the alleles, as described in Brieuc *et al.* (Chapter 2). Samples with more than 50% of missing values and loci with a minor allele frequency less than 10% were removed from the analysis. Polymorphic loci were identified independently in the Columbia River Basin and in the Puget Sound dataset. When both datasets were concatenated for downstream analyses, we only kept loci polymorphic for the same alleles in both basins.

**Population structure and standard statistics**

We calculated observed and expected heterozygosity for each locus and across all loci using ARLEQUIN 3.5.1.3 (Excoffier& Lischer 2010). Pairwise population differentiation $F_{ST}$ (Weir& Cockerham 1984) and its statistical significance was estimated with ARLEQUIN (Excoffier& Lischer 2010).

A Bayesian approach (BAPS 6.0, Corander *et al.* 2004) was used to infer the number of groups (*K*) in the Columbia River basin, and the degree of admixture in populations and within individuals. The optimal number of groups was determined using a stochastic optimization process and the range of groups tested was 2 to 11, the latter corresponding to the total number of populations sampled.

Pairwise Cavalli-Sforza and Edwards chord distances (1967) between populations was computed using POWERMARKER (Liu& Muse 2005) , which was also used to construct 1000 neighbor joining trees bootstrapped over loci. The consensus tree, along with bootstrap support, was computed using PHYLIP 3.6 (Felsenstein, 2005) and the tree was drawn using TREEVIEW 1.6 (Page 1996).

The proportion of genetic variance explained by different hierarchical levels of structure in the Columbia River was estimated using an analysis of molecular variance (AMOVA;

Excoffier *et al.* 1992). We used HIERFSTAT (Goudet 2005), a package in R (R Core Team, 2012), because the program permits more than two hierarchical levels. Three levels were analyzed, based on the findings of the phylogenetic analysis: the deeper divergence between the Interior Columbia Su/Sp lineages and the remaining lineages ($F_{CT\text{-}deeper}$), the shallower divergence between the Interior Columbia Su/F and the Lower Columbia lineages ($F_{CT\text{-}shallower}$) and finally, divergence between all populations.

**Adaptive divergence on different evolutionary scales**

We used BAYESCAN (Foll& Gaggiotti 2008) to identify outlier loci, consistent with adaptive divergence, using a Bayesian approach. The program partitions the observed differentiation measure ($F_{ST}$) for each locus into a locus specific component, shared by all populations, and a population specific component, shared by all loci. A locus was identified as behaving non-neutrally if the locus-specific component was significantly different from 0. This approach has been showed to have a very low error rate, therefore limiting the risks of false positive detection due to multiple testing (Foll& Gaggiotti 2008; Narum& Hess 2011).

One aim of our study was to identify loci associated with different hierarchical levels of population structure within the Columbia River. However, it is not possible to perform a hierarchical analysis in BAYESCAN. Additionally, the power of the analysis performed by this program is limited by the number of populations (the recommended number is greater than six, Foll& Gaggiotti 2008). The BAPS analysis and the phylogenetic analysis described earlier revealed three main lineages; a deeper divergence between Interior Columbia Sp/Su and the two other lineages, and a shallower divergence between the Interior Columbia Su/F and the Lower Columbia. We therefore used two separate outlier analyses on two datasets. The first dataset comprised all 11 populations from the Columbia River Basin. The second included 9 populations derived from the southern refugium (Interior Columbia Su/F and Lower Columbia). All outlier loci were identified using a false discovery rate (FDR) of 1%. Outlier loci were grouped in three categories in order to identify the loci that could be attributed to two hierarchical levels. The first comprised outliers detected in

the 11 populations only (A), the second comprised the outlier loci that were in common between the two analyses (B) and the third reflected outlier loci unique to the 9 population comparison (C). To determine whether each group of outliers could be attributed to a specific divergence level (deeper or shallower), we conducted a locus-specific AMOVA for each category (A, B, C) of outlier loci. To achieve this goal, we estimated $F_{ST}$ and $F_{CT}$ due to the deeper divergence ($F_{CT\text{-}deeper}$) or to the shallower divergence ($F_{CT\text{-}shallower}$) within each category. $F_{CT}$ for each analysis was standardized over loci by dividing by the locus specific $F_{ST}$ value, since $F_{ST}$ values differed across loci. Loci involved in the deeper adaptive divergence are expected to have a high $F_{CT\text{-}deeper}$ /$F_{ST\text{-}11}$ ratio, whereas those involved in the shallow divergence are expected to have a high $F_{CT\text{-}shallower}$ /$F_{ST\text{-}9}$ ratio, where $F_{ST\text{-}11}$ and $F_{ST\text{-}9}$ are explained by the two different datasets.

To visualize the influence of the outlier loci on population divergence, we constructed phylogenetic trees (as described above) and conducted Principal Component Analysis (PCA) with the ADEGENET (Jombart 2008; Jombart& Ahmed 2011) and ADE4 (Dray& Dufour 2007) packages in R (R Core Team, 2012). Individual analyses were conducted for the outliers grouped in categories A, B or C above.

**Regions of the genome under selection**

We determined the distribution of the outlier loci across the genome using polymorphic markers identified in this study that also mapped to the available Chinook linkage map (Brieuc *et al.* Chapter 2). We adapted the Kernel-smoothing method from Hohenlohe *et al.* (2010) to identify regions of higher differentiation for the two comparisons previously described. A moving weighted average of $F_{ST}$ was calculated at regular 3 cM bandwidth intervals across each linkage group. The 95% confidence intervals for $F_{ST}$ values were estimated using 100000 bootstraps for each locus found within the window, drawing from the entire dataset of mapped loci. Some regions of the genome had no polymorphic RAD tags identified and thus measures of the kernel-smoothing average and confidence intervals were discontinuous at these points.

### Test for parallel evolution

We implemented a random forest analysis to determine whether there were loci that were good predictors of run-timing across populations. This statistical technique has proven useful in identifying predictors in datasets where there are more variables than observations (Cutler *et al.* 2007; Gutierrez *et al.* 2011). Random forest analysis is a robust classification tool that constructs many bootstrapped decision trees based on parameters (here, loci). We were specifically interested in loci predicting the run timing for Spring versus Fall individuals, because there were insufficient observations of summer fish in the dataset. Therefore, the response variable was binary. Like most regression techniques, this approach does not allow missing values: therefore, we restricted the analysis to 1075 loci that had been genotyped in 179 individuals. The resulting predictors were then applied to the broader data set to test the accuracy of classification. We constructed 10000 trees and calculated the percent classification error across all iterations (Cutler *et al.* 2007). Across all trees, we calculated the relative importance of individual loci in predicting run timing, based on a proximity matrix between individual observations (that is, the similarity between individuals with similar run timing at these loci). All analyses were performed using the RANDOMFOREST package (Liaw& Wiener 2002) in R (R Core Team 2012). We then used the 10 best predictive loci and constructed a phylogenetic tree for all the individuals in the dataset, as previously described, to confirm that the loci separated the populations by run timing rather than evolutionary history.

## Results

### Population structure and standard statistics

A total of 434 individuals were sequenced from the Columbia River. We identified 14105 polymorphic RAD loci with a maximum of two alleles per locus and a minor allele frequency greater than 10%, of which 3528 had been mapped to previously described Chinook salmon linkage groups (Brieuc *et al.* Chapter 2). We identified 10558 polymorphic loci in the Puget Sound populations. A total of 4636 loci were polymorphic for the same

alleles across both river basins, and of these 2046 had been placed on the linkage groups. Polymorphic loci across both dataset were subsequently used when considering populations from both regions.

There was no significant difference between observed and expected heterozygosity ($H_O$ and $H_E$ respectively) across most populations (Table 3.1). Two populations, Methow Su and Deschutes F, exhibited a heterozygote deficit. These populations corresponded to the oldest samples for which DNA quality was the lowest (data not shown). Additionally they had the highest number of missing values (27% and 21% average missing values per individual respectively). The same populations were studied in Narum *et al.* (2010), but did not exhibit this pattern. This suggests that the heterozygote deficit observed here might be the result of a bias against the detection of heterozygote in these populations because of degraded DNA, although biological explanations, such as inbreeding of Wahlund effect, cannot be ruled out. Heterozygote excess was observed for McKenzie Sp and Lyons Ferry F, but there was no apparent DNA degradation for these populations and the number of missing value was low (average 7.5% and 6% respectively). Heterozygote excess had not been previously reported (Narum *et al.* 2010), but could be the result of small population sizes. On average $H_E$ was lower in the Interior Columbia Sp/Su lineage (0.116) compared to the Interior Columbia Su/F lineage (0.150 ± 0.020) or the Lower Columbia lineage (0.148 ± 0.003). These results are consistent with those observed in a previous study, (Narum *et al.* 2010), which also observed reduced heterozygosity in the Interior Columbia Sp/Su lineage.

Bayesian clustering of individuals with BAPS revealed three main groups (Figure 3.2), corresponding to the three main lineages in the Columbia River Basin described in previous studies (Moran *et al.* 2012; Narum *et al.* 2010; Waples *et al.* 2004). These three lineages were used as the basis for subsequent analyses. Admixture results (Figure 3.2) revealed a few admixed populations: the Kalama Spring, the Lyons Ferry and McKenzie populations showed evidence for admixture from the Interior Columbia Sp/Su lineage (13.0%, 5% and 3% respectively); the Deschutes population was 5% admixed from the Lower Columbia. Admixture in the Kalama Spring population had previously been observed and likely explained by hatchery practices (Narum *et al.* 2010). A few individuals

in the Cowlitz Spring and Deschutes Fall population were potential migrants from the Interior Columbia Sp/Su lineage.

Patterns of pairwise population differentiation estimates ($F_{ST}$) estimated with GENEPOP (Raymond& Rousset 1995; Rousset 2008) within each lineage described by BAPS showed that the populations within the Interior Columbia Sp/Su lineage were the least differentiated (Table 3.2). The average differentiation between the Lower Columbia and the Interior Columbia Su/F lineages was much lower (0.075 ± 0.015) than between the Interior Columbia Spring type and the Lower Columbia or the Interior Columbia Su/F (0.273 ± 0.030 and 0.291 ± 0.027 respectively). The hierarchical AMOVA analysis revealed that 77% of the genetic variance could be explained by the divergence between the interior Columbia Sp/Su and the other two major lineages (the deeper divergence), whereas 14% was due to the differences between the Interior Columbia Su/F and coastal lineages (the shallower divergence; Table 3.3). These results are consistent with the view that there were two pre-glaciation refugia for the Columbia River Chinook.

Evolutionary relationships between populations based on a phylogenetic tree with Cavalli-Sforza and Edwards' distances (1967) with 4636 loci are shown in Figure 3.3, where all nodes but one were supported at 100% with 1000 bootstraps. The three lineages identified with BAPS are also represented on this tree. The Puget Sound populations were more closely related to the Interior Columbia Su/F and Lower Columbia lineages suggesting that they are derived from the same refugium.

## Detection of outlier loci and partitioning of these loci across different hierarchical levels

We identified 282 outlier loci in the entire dataset from the Columbia River (referred to as the 11 population analysis) and 153 outlier loci in the Interior Columbia Su/F and Lower Columbia populations only (referred to as the 9 population analysis), with a false discovery rate (FDR) of 1%. A total of 148 outlier loci were unique to the 11 population analysis (group A, Figure 3.4), 114 markers were outlier in both analyses (group B) and 39 were unique to the 9 population analysis (group C).

Most of the genetic variation for the outlier loci in the group A, was explained by the differences between the Interior Columbia Sp/Su and the remaining two lineages (high $F_{CT\text{-}deeper}$ /$F_{ST\text{-}11}$, Figure 3.5). A different pattern was observed for the outlier loci in the group B, where most of the variation was due to the difference between the two shallow lineages (high $F_{CT\text{-}shallower}$ /$F_{ST\text{-}9}$, Figure 3.5). Loci in the group A were therefore attributed to the deeper divergence, whereas loci from the group B were attributed to the shallower. The $F_{CT\text{-}deep}$ /$F_{ST\text{-}11}$ distribution for the outlier loci in group B appeared more uniform but with an over-representation of the loci with very low $F_{CT\text{-}deeper}$ /$F_{ST\text{-}11}$ value. This suggests that while most loci in this group could be attributed to the shallower divergence, some loci might be misclassified. Finally, outlier loci in the group C had high $F_{CT\text{-}deeper}$ /$F_{ST\text{-}9}$ values and were attributed to the shallower divergence.

The phylogenetic trees constructed with the outlier loci in the 11 or 9 population comparisons (A&B and B&C respectively, Figure 3.6a and 3.6c) had a similar topology compared to that of the entire dataset (Figure 3.3). However, the distances between all three lineages were greater in the 11 population analysis (Figure 3.6a, A&B). In contrast, the distances between the shallower lineages were increased with outlier loci in the 9 population analysis (B&C, Figure 3.6c). Using outlier loci unique to the 11 population analysis (A) resulted in a tree with a similar topology to the overall tree, but with very high distances between the deeper lineages (Figure 3.6b). All these observations were also supported by the PCA analyses (Figure 3.7). These results confirm that loci in the groups A were attributable to the deeper divergence, and loci in the group B and C to the shallower divergence.

Using the mapped markers only we determined that the outlier loci were not randomly distributed across chromosomes after correcting for the number of markers on each chromosome (all populations: $\chi 2 = 54.22$, df = 33, p = 0.011 ; ICSu/F & LC: $\chi 2 = 48.94$, df = 33, p = 0.037; Figure 3.8). We identified 20 regions of high divergence on 13 chromosomes in the eleven population comparison and 16 regions on 11 chromosomes in the nine population analysis (Table 3.4). Fourteen chromosomes had no region of high divergence, despite the fact that most of them had outlier loci identified by BAYESCAN.

Relying on the classification of outlier loci either involved in the deeper or the shallower divergence, we were able to identify six regions of high divergence linked to the deeper divergence on Ots02p, Ots04, Ots11, Ots15, Ots20 and Ots30 (Table 3.4, Figure 3.9). Similarly we identified 10 regions of high divergence linked to the shallower divergence on Ots01, Ots03, Ots07, Ots10, Ots13 (2), Ots18, Ots20 and Ots28 (2). The lengths of the regions involved in the deeper divergence ranged from 3.32 to 10.73cM and were supported by 1 to 5 outlier loci, whereas those involved in the shallower divergence ranged from 1.24 to 12.17cM and were generally supported by fewer loci (1 to 4). A region of high divergence on Ots13 was identified in the 9 population analysis only, but the outlier locus in this region was attributed to the shallower divergence. In contrast, a region of high divergence on Ots04 was identified in the 11 population analysis only, but the outlier locus in this region appeared to be involved in the shallower divergence. Some chromosomes had more than one region of high divergence, and Ots13 and Ots28 were the chromosomes with the highest number of such regions, with 4 and 3 respectively. Most regions of high divergence (71%) were located in distal regions from the centromere. Among the mapped outlier loci, 50.7% of the deep divergence outlier loci were located in a region of high divergence or in close proximity (±9cM as defined when calculating the moving average of $F_{ST}$) compared to 40.6% of the shallower divergence outlier loci.

**Parallel evolution and identification of candidate loci linked to run timing**

Using 179 spring- and fall-run timing individuals and 1075 loci, the random forest approach successfully identified loci that correctly predicted the run-timing with a 7.26% error rate (Table 3.5). All the fall-run timing individuals but one were correctly predicted. Twelve spring-run timing individuals were incorrectly predicted as being fall-run individuals. All these individuals belonged to the Cowlitz Spring population from the lower Columbia F/Sp lineage. The list of the 20 best predictors and their relative contribution is reported in Figure 3.10. The neighbor joining tree constructed with all spring- and fall-individuals in the dataset using the ten best predictive loci separated the spring- and fall-populations (Figure 3.11), except for Cowlitz Spring, therefore further supporting that these loci may be linked to run timing. Finally two of the 10 best predictive loci

(Ot024147_Ots30 and Ot000978_Ots30) map to the same position on the located on the same linkage group (Ots30, 74.5cM). These loci map to a highly divergent region on Ots30 attributed to pre-glaciation divergence (Figure 3.9).

## Discussion

Here, we aimed to examine the incidence of adaptive divergence across different lineages in the Columbia Basin, to determine the distribution of non-neutral divergence across the Chinook salmon genome, and to determine whether there was molecular support for parallel evolution in run timing. Using 14103 polymorphic RAD loci, our results confirmed the presence of three main lineages in the Columbia River. High levels of population divergence between Interior Columbia River Su/Sp lineage versus two other lineages in the basin supported the previously described hypothesis of colonization of the Columbia River Basin from two main refugia following the last glaciation event. Examination of outlier loci in two datasets – the first one comprising all populations in the Columbia River basin, the second on populations representing a more recent divergence between Interior Columbia Su/F and Coastal Columbia – permitted the identification of 148 and 153 outlier loci associated with the older and the more recent divergences between the lineages respectively. The placement of 135 of these outlier loci on the Chinook linkage map revealed that they were not evenly distributed across linkage groups, and many were clustered in the telomeric regions of chromosomes. The distribution of regions consistent with adaptive divergence varied across lineages: for example, outliers on Ots02, Ots20 or Ots30, separated the interior Columbia Sp/Su from the other two lineages, while others separated the more recently diverged lineages, (*e.g.* Ots09, Ots13, Ots18, Ots28 or Ots29). Several loci were found to be associated with spring and fall return timing in a subset of individuals, and accurately predicted individual traits in a larger dataset. Of these loci, two mapped to the same position and were located in a region of high divergence on Ots30, associated with the divergence between Interior Columbia Sp/Su and the remaining two lineages.

Interestingly, the Bayesian clustering analysis did not identify the individual eleven populations in the dataset, despite higher locus number, but rather the three main lineages in the Columbia River. Previous studies have shown that BAPS could successfully distinguish populations with $F_{ST}$ values as low as 0.02 (*e.g.* Hauser *et al.* 2006). In this study most of the $F_{ST}$ were higher than 0.02 (Table 3.2), but the fine scale population structure was not resolved with BAPS. This result could be due to the small sample sizes in our study, where each population comprised between 26 and 41 individuals. The observed clustering of the data could also be the result of the hierarchical population structure in the data related to pre- and post-glaciation divergence, where 77% and 14% of the observed variation in the dataset were due to differences among deep lineages and shallow lineages respectively, whereas only 9% was due to the differences between populations. A study by Kalinowski (2011) recently demonstrated that hierarchical population structure with long divergence times could affect the power of clustering analysis programs to detect fine scale population structure. Although the analysis in this paper were performed with the software STRUCTURE (Pritchard *et al.* 2000), it is reasonable to assume that the same limitation could be true with other clustering approaches, such as the one used in this study.

The population structure computed in this study confirmed phylogenetic relationships identified in previous studies (Moran *et al.* 2012; Narum *et al.* 2010). As previously stated, the relationships between the three lineages support the hypothesis of colonization from two main refugia, and subsequent persistent reproductive isolation in the Interior Columbia between populations derived from the two refugia. Moreover the $F_{ST}$ and heterozygosity values were lower in the Interior Columbia Sp/Su lineage compared to the other lineages, suggesting limited gene flow between the two populations from the Interior Columbia Sp/Su. Although these observations are based on only two populations in this study, high philopatry among the Interior Columbia Sp/Su has been suggested (Narum *et al.* 2010; Quinn 1993) and our results do support this hypothesis.

The power to detect outlier loci linked to genes under selection can be affected by several factors. For example, the time since the selection event, the distance between the gene under selection and the observed locus, and the recombination rate can limit the power to

detect outlier loci (Nielsen 2005). As the time since the selection event increases, the informative region around the gene under selection gets narrower because of recombination and drift. We therefore expected to detect fewer outlier loci attributable to a more ancient divergence, (here the Interior Columbia Su/Sp) than to a more recent one. However, we identified 148 loci attributable to the older divergence compared to 153 to the more recent one. It is possible that the power of the outlier analyses was affected by the number of populations involved in our efforts to partition divergence across the hierarchical population structure. We therefore conducted an additional test for selection with BAYESCAN using the two populations from the Interior Columbia Sp/F and seven populations, randomly selected, from the other two lineages as to have a total of 9 populations. The results were not affected by the lower number of populations (data not shown). Moreover, although more loci attributable to pre-glaciation divergence (50.7%) were located in a region of high divergence compared to the outlier loci attributable to post-glaciation divergence (40.6%), there were fewer regions associated with to the deeper divergence than to the shallower divergence. This result would suggest that the high number of outlier loci identified in the deeper divergence might be in part due to the fact that they mapped to the same regions. Some regions of high divergence could not be attributed to the deeper or shallower divergence exclusively, therefore suggesting that they may be under selection on several evolutionary scales. Regions of high divergence on linkage groups have previously been identified in other teleost fish, such as Atlantic salmon (Bourret *et al.* 2013), threespine stickleback (Hohenlohe *et al.* 2010) and Atlantic cod (Bradbury *et al.* 2013; Hemmer-Hansen *et al.* 2013). Annotation of the loci in the regions of high divergence and of the outlier loci themselves would help gain further insights in the mechanisms of adaptive divergence in the Columbia River.

Most of the regions of high divergence were located in distal regions from the centromere. However, 16 chromosome arms comprised mainly duplicated loci in distal regions (Brieuc *et al.* Chapter 2) and these regions have therefore not been genotyped in this study, since we explicitly targeted non-duplicated loci. Many regions of interest might not have been identified in the present study. The use of markers, such as microsatellite markers or SNPs,

where alleles specific to each paralog have been identified, would allow the investigation of adaptive divergence in these regions.

We were able to successfully to identify predictive loci of run timing, specifically spring and fall, using the random forest approach. The fact that two of the best ten predictive loci mapped to the same position would suggest that similar genes might be associated with run timing in the different lineages. This would support the hypothesis of molecular basis for parallel evolution within each lineage in the Columbia River and in Puget Sound. Missing values were the main limitation of this approach. Reducing the number of missing values is therefore necessary to successfully implement this method to identify loci that might be involved in parallel evolution. This goal could be achieved by increasing the sequencing depth for each individual or by inferring the missing values. This latter approach has been implemented in human genomic studies to address the high rate of missing values inherent to high throughput sequencing (e.g. Howie *et al.* 2009; Li *et al.* 2009; Marchini *et al.* 2007; Scheet& Stephens 2006) and may be explored for other species, such as salmonids (D. Drinan, pers. comm.).

Here we have gained new insights in Chinook salmon evolution in the Columbia River. Beyond identifying putative loci involved in pre- and post-adaptive divergence in this system, we have demonstrated the usefulness of dense linkage maps in identifying regions of the genome that may have been involved in adaptive evolution. Although the evidence of parallel evolution of run timing at the molecular level is still not extensive we have identified a promising approach to further investigate this intriguing topic.

## Acknowledgments

This work was funded by a grant from the Washington Sea Grant Program, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA07OAR4170007, Project No. R/F-51 awarded to KAN. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of their sub-agencies. We thank Linda Park, Robin Waples, Paul Moran, David Teel and Michael Ford,

Northwest Fisheries Science Center, NOAA, John Hess and Sean Narum at the Columbia River Inter-Tribal Fish Commission, and Steve Stout, Marblemount Hatchery, for giving us access to the samples used in this study. We thank Isadora Jimenez-Hidalgo, Melissa Baird, Carita Pascal, James Seeb, Steven Roberts and Jon Hess for their help with the lab work and paper review. We thank Lorenz Hauser, Eric Ward, Kotaro Ono for their invaluable advice on data analysis.

## Tables

| Population Number | Population | Lineage | Run timing | Year sampled | Sample size | Ho | He |
|---|---|---|---|---|---|---|---|
| *Columbia River (14105 loci)* | | | | | | | |
| 1 | Cowlitz_Sp | Lower Columbia | Spring | 2002 | 30 | 0.142 | 0.140 |
| 2 | Kalama_Sp | Lower Columbia | Spring | 2004 | 27 | 0.149 | 0.151 |
| 3 | Cowlitz_F | Lower Columbia | Fall | 2002 | 32 | 0.15 | 0.145 |
| 4 | Kalama_F | Lower Columbia | Fall | 2004 | 41 | 0.145 | 0.140 |
| 5 | McKenzie_Sp | Lower Columbia | Spring | 2004 | 23 | 0.173 | 0.149 |
| 6 | Methow_Su | Interior Columbia Su/F | Summer | 1994 | 27 | 0.085 | 0.134 |
| 7 | Wenatchee_Su | Interior Columbia Su/F | Summer | 2004 | 33 | 0.149 | 0.142 |
| 8 | Deschutes_F | Interior Columbia Su/F | Fall | 1998 | 26 | 0.118 | 0.141 |
| 9 | Lyons_Ferry_F | Interior Columbia Su/F | Fall | 2007 | 32 | 0.238 | 0.184 |
| 10 | Chiwawa_Sp | Interior Columbia Sp/Su | Spring | 2007 | 37 | 0.118 | 0.119 |
| 11 | Upper_Salmon_Sp | Interior Columbia Sp/Su | Spring | 2007 | 29 | 0.113 | 0.113 |
| *Puget Sound (4635 loci)* | | | | | | | |
| 12 | Skagit_Sp | Puget Sound | Spring | 2009 | 34 | 0.267 | 0.267 |
| 13 | Skagit_Su | Puget Sound | Summer | 2009 | 34 | 0.301 | 0.276 |
| 14 | Skate_F | Puget Sound | Fall | 2009 | 28 | 0.246 | 0.277 |

Table 3.5 – Chinook salmon populations sampled in this study with their lineage, run timing, sampling year, sample size, observed (Ho) and expected (He) heterozygosity. The population number corresponds to that used on Figure 3.1. Su/F: Summer Fall run timing; Sp/Su: Spring or Summer run timing

| | Cowlitz Sp | Kalama Sp | Cowlitz F | Kalama F | McKenzie Sp | Methow Su | Wenatchee Su | Deschutes F | Lyons Ferry F | Chiwawa Sp |
|---|---|---|---|---|---|---|---|---|---|---|
| **Kalama Sp** | 0.0292 | | | | | | | | | |
| **Cowlitz F** | 0.0149 | 0.0279 | | | | | | | | |
| **Kalama F** | 0.0253 | 0.0341 | 0.0072 | | | | | | | |
| **McKenzie Sp** | 0.0605 | 0.0288 | 0.057 | 0.0627 | | | | | | |
| **Methow Su** | 0.0796 | 0.0838 | 0.0735 | 0.0802 | 0.1077 | | | | | |
| **Wenatchee Su** | 0.0773 | 0.0764 | 0.0689 | 0.0749 | 0.103 | 0.0124 | | | | |
| **Deschutes F** | 0.0579 | 0.0593 | 0.0495 | 0.0565 | 0.088 | 0.0214 | 0.0197 | | | |
| **Lyons Ferry F** | 0.073 | 0.0684 | 0.0639 | 0.0709 | 0.0894 | 0.0419 | 0.0323 | 0.0242 | | |
| **Chiwawa Sp** | 0.2841 | 0.2125 | 0.2685 | 0.273 | 0.265 | 0.3024 | 0.2901 | 0.2777 | 0.2471 | |
| **Upper Salmon Sp** | 0.3103 | 0.2361 | 0.2938 | 0.298 | 0.2912 | 0.3307 | 0.3139 | 0.3041 | 0.2652 | 0.0365 |

Table 3.6 –Pairwise genetic differentiation $F_{ST}$ (Weir& Cockerham 1984) calculated with 14105 RAD loci for all population pairs in the Columbia River basin. All pairwise comparisons were significant ($p<0.01$).

| | **Variance component** | **Proportion of variance** |
|---|---|---|
| Among interior Columbia Spring/Su and all other populations (deep lineage) | 555.67 | 0.77 |
| Among interior Columbia Su/F and coastal S/F within the deep lineage | 104.22 | 0.14 |
| Among populations within lineages | 65.74 | 0.09 |
| Within populations | 0 | 0 |

Table 7.3 – Partitioning of the genetic variation in Chinook salmon in the Columbia River at each level using AMOVA. Lineages are defined by the phylogenetic analyses and the assignment test. Negative values have been replaced by 0.

| Chromosome | Chromosome arm | Position (cM) | Relative distance from the centromere | Number of outlier loci | | | Comparison | |
|---|---|---|---|---|---|---|---|---|
| | | | | A | B | C | 11 pop | 9 pop |
| Ots01 | Ots01q | 162.78-174.95 | 0.81-0.92 | | 1 | | | x |
| Ots02 | Ots02q | 74.96-85.69 | 0.02-0.15 | 4 | | | x | |
| Ots02 | Ots02p | 8.46-21.33 | 0.8-1 | 7 | 2 | | x | |
| Ots03 | Ots03q | 125.81-129.98 | 0.41-0.48 | | 1 | | | x |
| Ots04 | Ots04c | 77.42-80.2 | c | | 1 | | x | |
| Ots04 | Ots04q | 134.33-141.27 | 0.54-0.61 | 2 | | | x | |
| Ots07 | Ots07p | 23.15-30.04 | 0.71-0.81 | | 1 | | x | x |
| Ots08 | Ots08q | 115.41-121.04 | 0.81-1 | 2 | 2 | | x | x |
| Ots09 | Ots09p | 25.81-30.19 | 0.78-0.82 | 1 | 1 | | x | |
| Ots09 | Ots09q | 147.11-152.96 | 0.56-0.63 | 1 | 1 | | x | |
| Ots10 | Ots10q | 131.28-132.52 | 0.99-1 | | 1 | | | x |
| Ots11 | Ots11q | 108.4-111.72 | 0.7-0.76 | 1 | | | x | |
| Ots12 | Ots12q | 116.98-119.6 | 0.48-0.51 | 2 | 2 | | | x |
| Ots13 | Ots13p | 8.27-9.73 | 0.98-1 | 3 | 1 | | x | |
| Ots13 | Ots13p | 37.61-42.01 | 0.58-0.64 | | 1 | | | x |
| Ots13 | Ots13q | 134.44-137.38 | 0.68-0.72 | 1 | | | | x |
| Ots13 | Ots13q | 144.71-154.98 | 0.83-0.99 | | 3 | 1 | x | x |
| Ots15 | Ots15p | 18.11-27.04 | 0.8-0.9 | 1 | | | x | |
| Ots15 | Ots15c | 92.09-98.47 | c | 1 | 1 | | x | |
| Ots18 | Ots18c | 8.86-19.62 | c | | 2 | | x | x |
| Ots18 | Ots18q | 58.83-67.05 | 0.79-0.92 | 1 | 1 | 2 | x | x |
| Ots20 | Ots20c | 40.62-47.43 | c | | 2 | | x | x |
| Ots20 | Ots20q | 99.98-105.82 | 0.94-1 | 5 | | | x | |
| Ots28 | Ots28q | 18.56-20.95 | 0.21-0.39 | | 1 | | | x |
| Ots28 | Ots28q | 64.87-68.46 | 0.92-0.98 | 1 | 2 | | x | x |
| Ots28 | Ots28q | 49.34-54.71 | 0.69-0.77 | | 1 | | | x |
| Ots29 | Ots29q | 60.39-64.38 | 0.93-1 | 1 | 4 | 1 | x | x |
| Ots30 | Ots30q | 70.69-77.9 | 0.81-0.91 | 5 | | | x | |

Table 3.8 –Location of regions of high divergence within linkage groups for Chinook salmon in the Columbia River identified from 3528 mapped loci. The position, the relative distance from the centromere (c: centromere), the number of outlier loci from each category of outlier loci in each region (±9cM) and the comparison in which the region was identified (11 or 9 population comparison) are indicated for each linkage group.

| | | Predicted | | |
|---|---|---|---|---|
| | | Fall | Spring | Error |
| **Observed** | Fall | 85 | 1 | 0.0116 |
| | Spring | 12 | 81 | 0.129 |

Table 3.9 – Confusion matrix from the random forest analysis: prediction of run timing for 86 fall individuals and 93 spring individuals using 1075 loci. One fall individual was misclassified as a spring individual and 12 spring individuals were misclassified as fall individuals with the random forest prediction. Most individuals were correctly predicted (92.7%).

# Figures

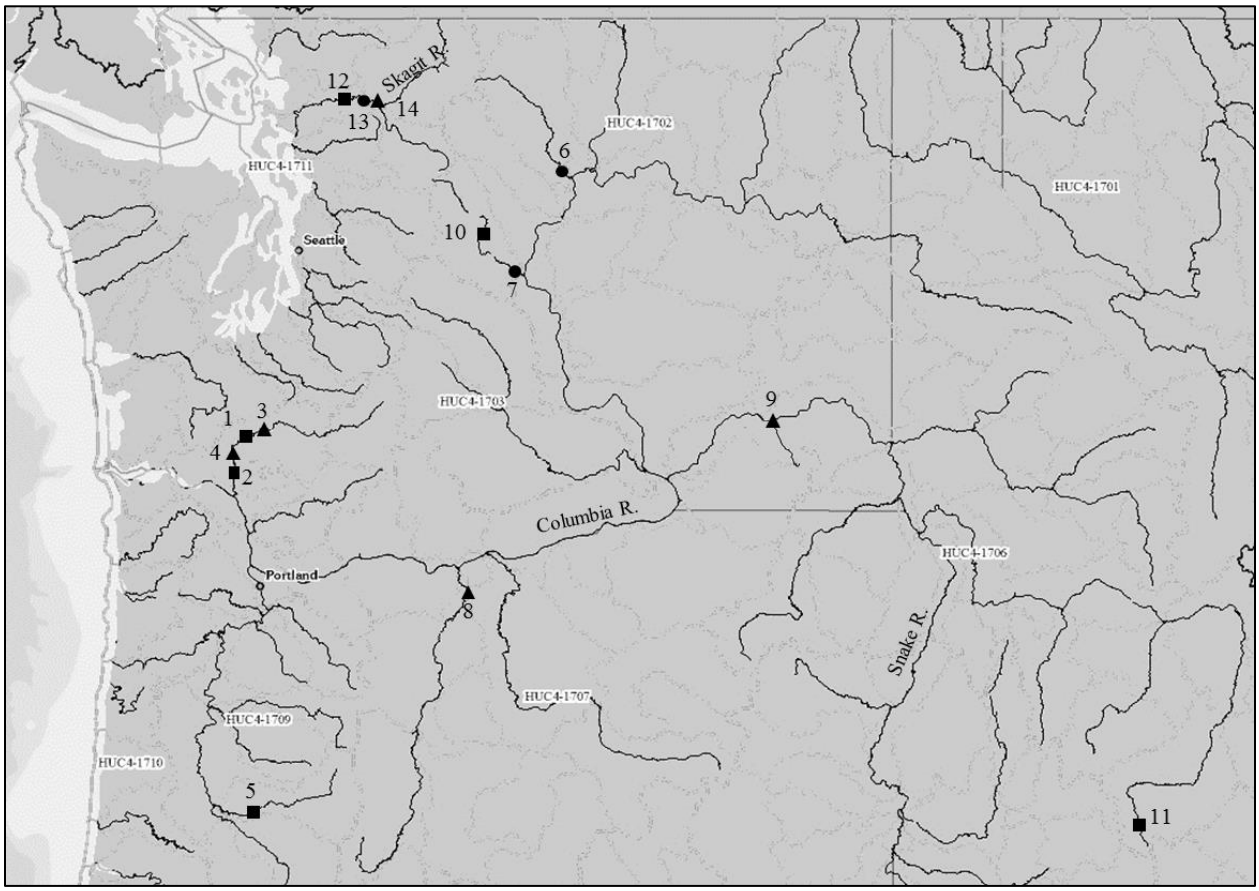


Figure 3.7 – Geographic locations of the Chinook salmon populations sampled in this study. The numbers correspond to those described in Table 1. Run timing for each population is indicated: square: Spring, round: Summer, triangle: Fall.

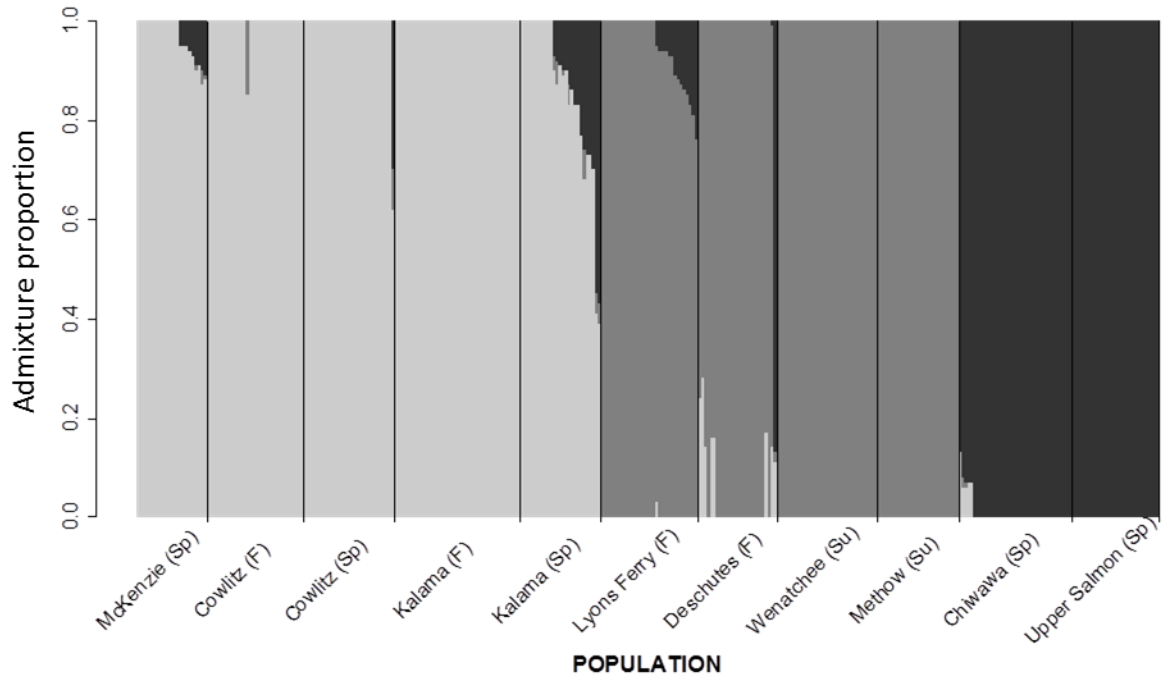


Figure 3.8 - Bayesian clustering (BAPS) analysis of individual Chinook salmon in the Columbia River basin using RAD markers across 11 populations (14105 loci). Three main groups are identified, corresponding to three main lineages: Interior Columbia Su/F (light grey), Lower Columbia (medium grey), and Interior Columbia Sp/Su (dark grey). Each bar represents an individual and its relative composition from each of the three lineages.

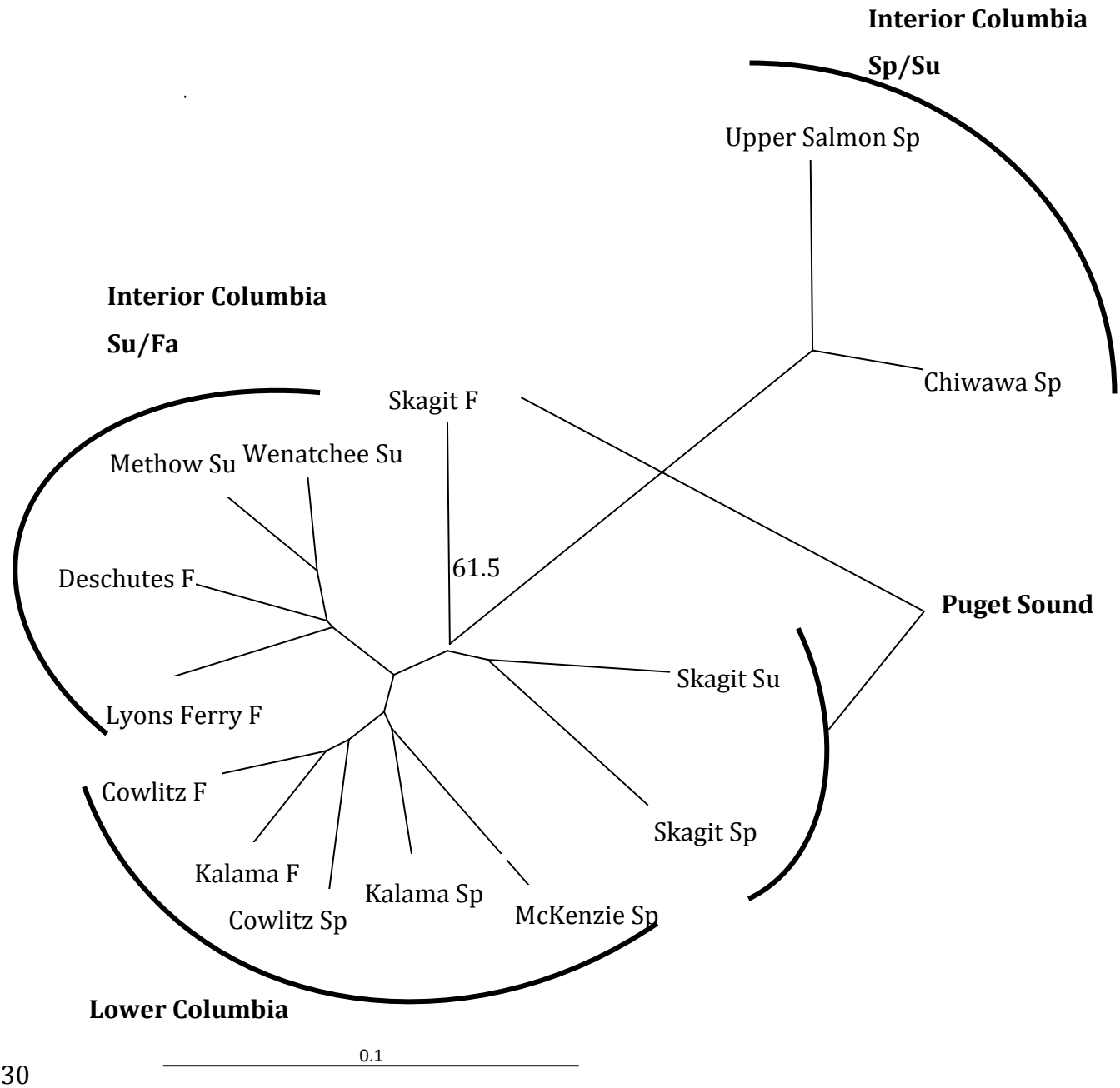


Figure 3.9 - Neighbor-joining dendogram for 11 populations from the Columbia River and 3 populations from Puget Sound, based on Cavalli-Sforza and Edwards (1967) chord distances (4636 loci). All nodes had 100% bootstrap support (1000 replicates), except for one node that had a bootstrap support of 61.5 % (shown). F: Fall run timing, Sp: Spring run timing, Su: Summer run timing

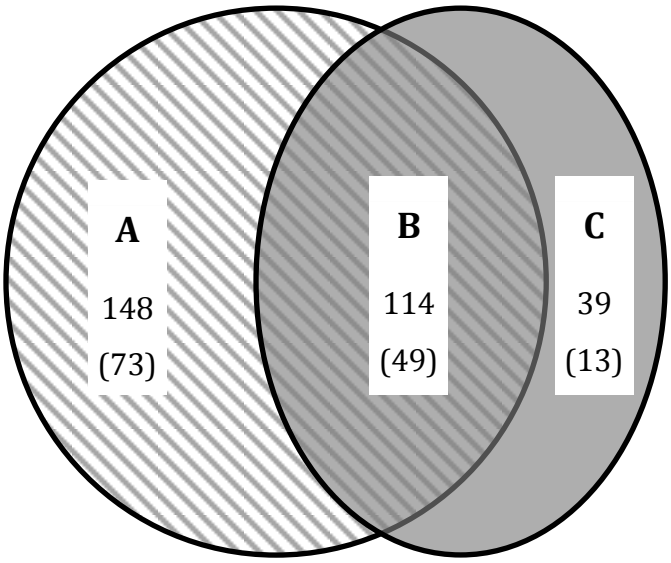


Figure 3.10 – Number of unmapped and mapped (in parentheses) outlier loci identified within the entire Columbia River dataset (hashed circle) and in the Interior Columbia Su/F and Lower Columbia (grey circle). Outlier loci were separated in three categories: A: 148 outlier loci in analysis based on 11 populations; B: 114 outlier loci in common between the two analyses; C: 39 outlier loci based on the analyses of the Interior Columbia Su/F and Lower Columbia only. The number of mapped outlier loci is indicated in parenthesis

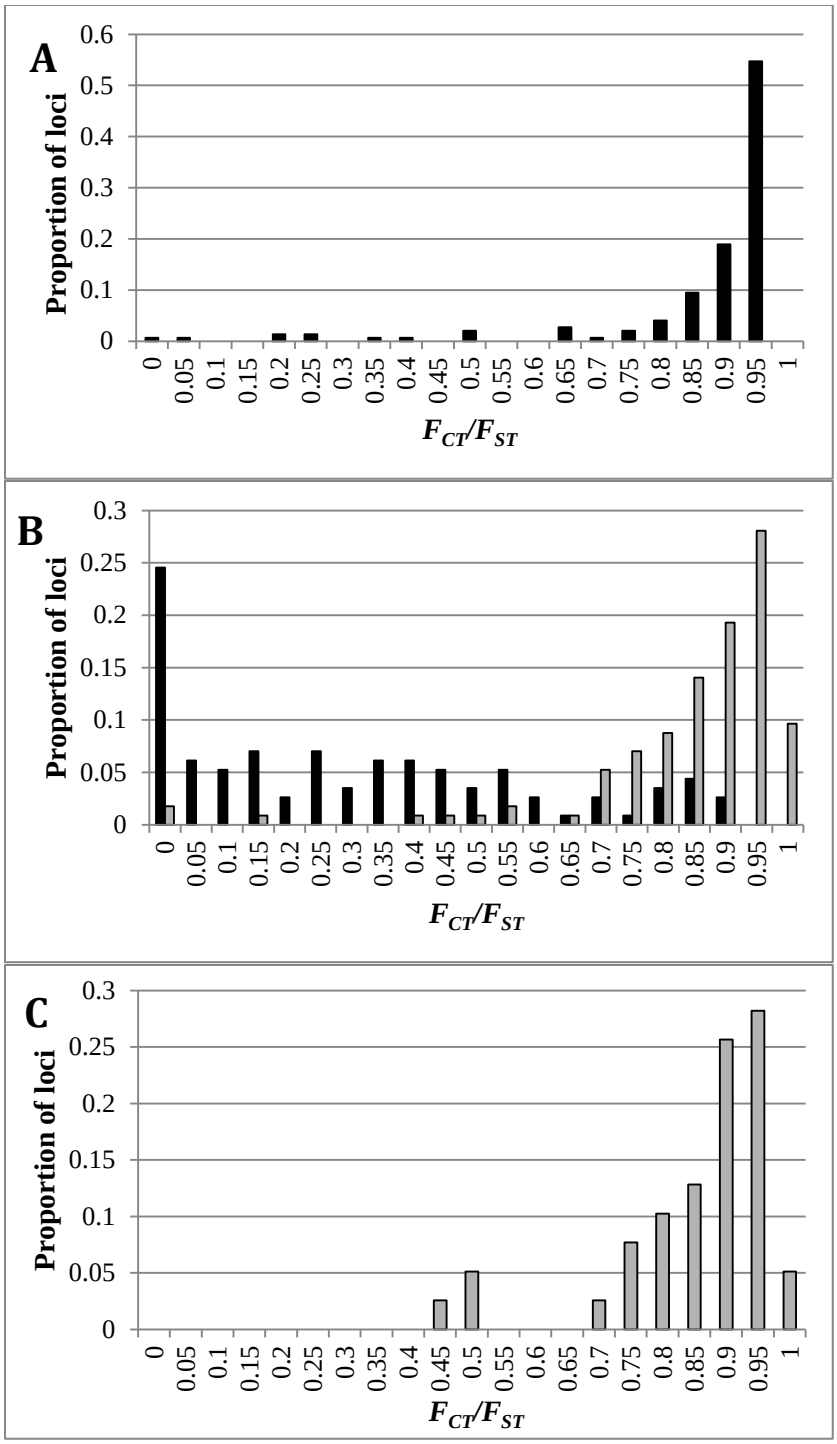


Figure 3.11 – Partitioning of outlier loci among two hierarchical levels in the Columbia River basin. Analyses are based on three categories, group A, group B and group C defined in Figure 4. $F_{CT}/F_{ST}$ values are given for each category of outlier loci, where $F_{CT\text{-}deeper}/F_{ST\text{-}11}$ is represented in black and $F_{CT\text{-}shallower}/F_{ST\text{-}9}$ is represented in grey. Only loci identified as outliers in each comparison are represented.

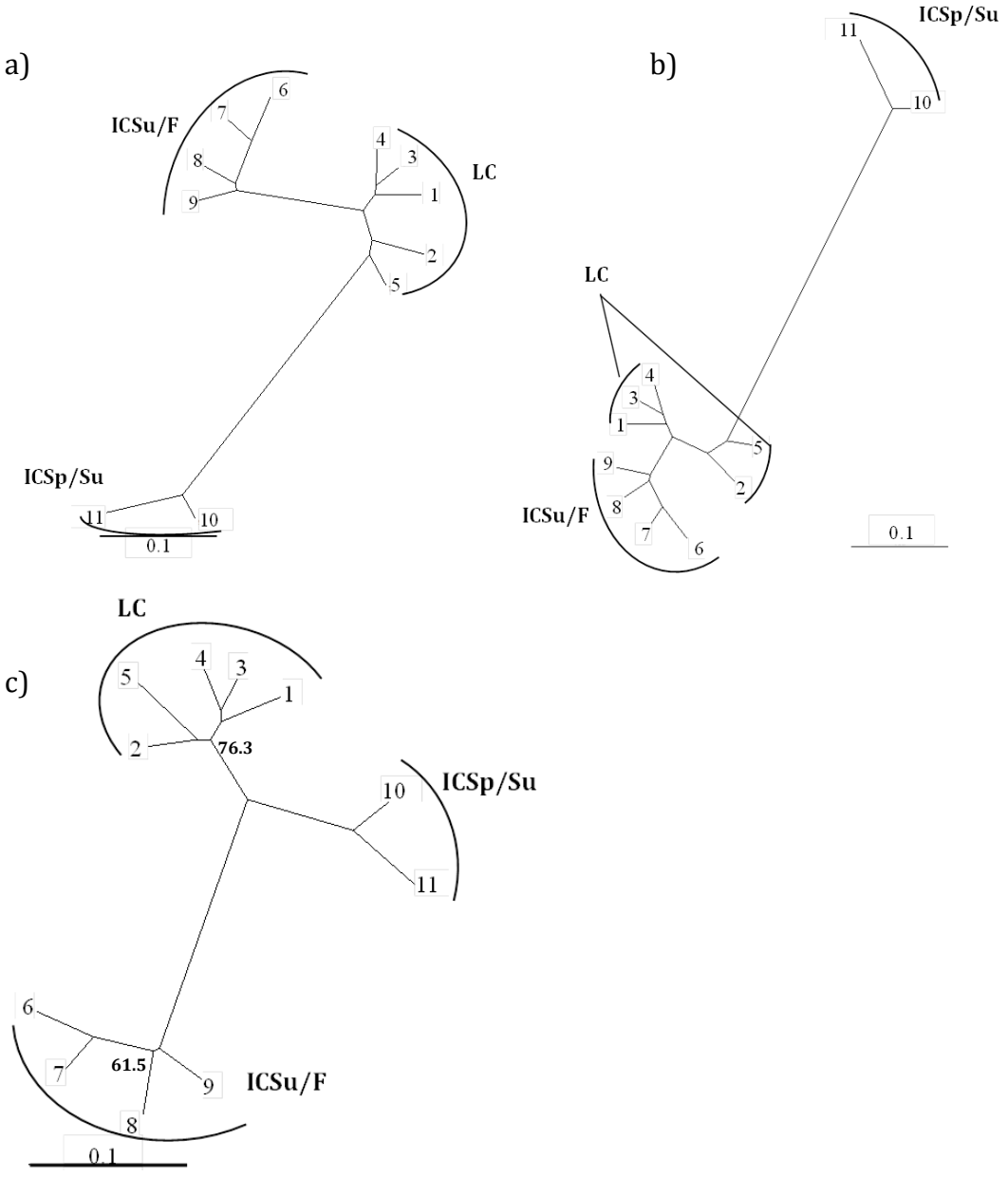


Figure 3.12 - Neighbor-joining dendograms for 11 populations from the Columbia River, based on Cavalli-Sforza and Edwards (1967) chord distances using mapped and unmapped outlier loci: a) 282 outlier loci detected when using all populations (A&B); b) 148 outlier loci detected when post-glacial lineages are excluded (A); c) 114 outlier loci detected when using all populations from the Interior Columbia Su/F and Lower Columbia lineages (B&C);. The population numbers correspond to those on Table 1. LC: Lower Columbia, ICSu/F: Interior Columbia Su/F, ICSp/Su: Interior Columbia Sp/Su. All nodes had a bootstrap support of 80% or more, except for those indicated on the trees.

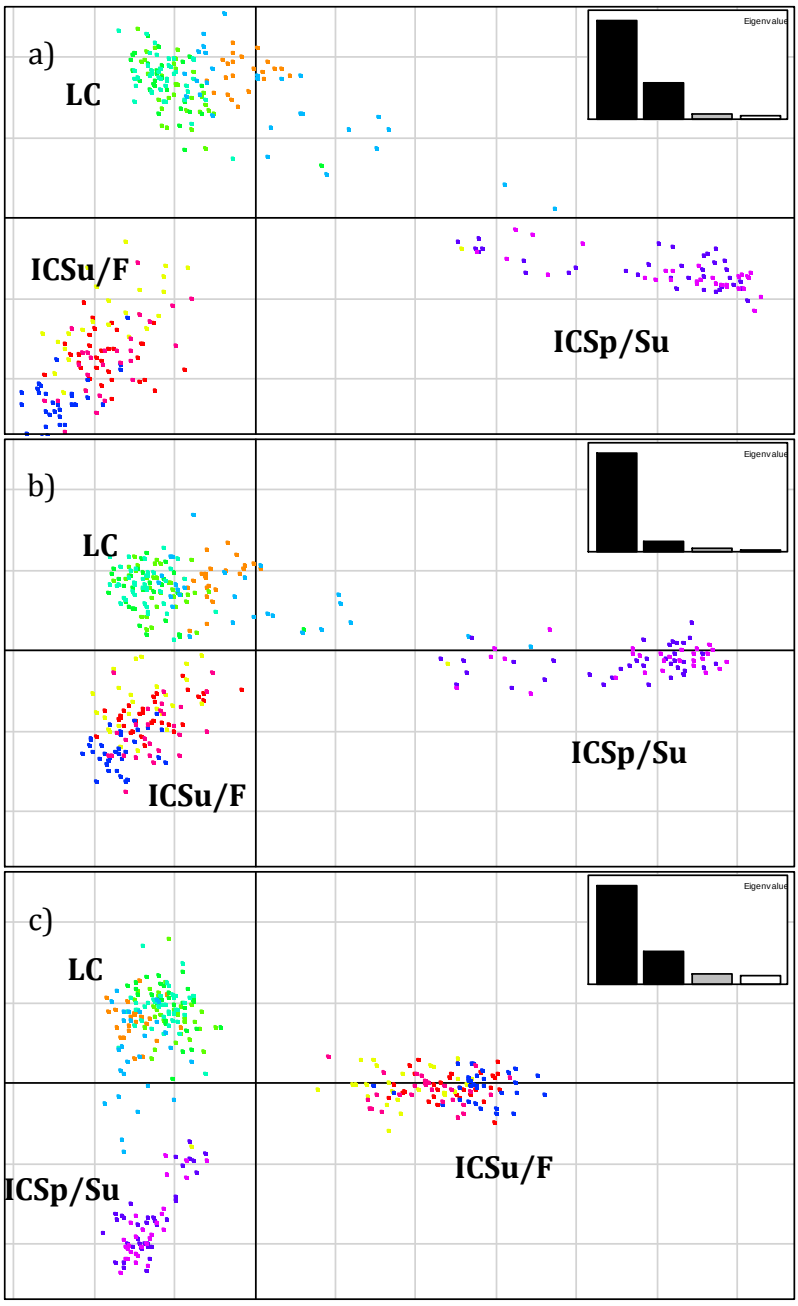


Figure 3.13 – Principal component analysis and Eigenvalues for 11 populations from the Columbia River using: a) all outlier loci (282) detected among all populations (A&B); b) outlier loci (148) unique to the 11 population comparison (A); c) all outlier loci detected among the shallower lineages (114, B&C). LC: Lower Columbia, ICSu/F: Interior Columbia Su/F, ICSp/Su: Interior Columbia Sp/Su.

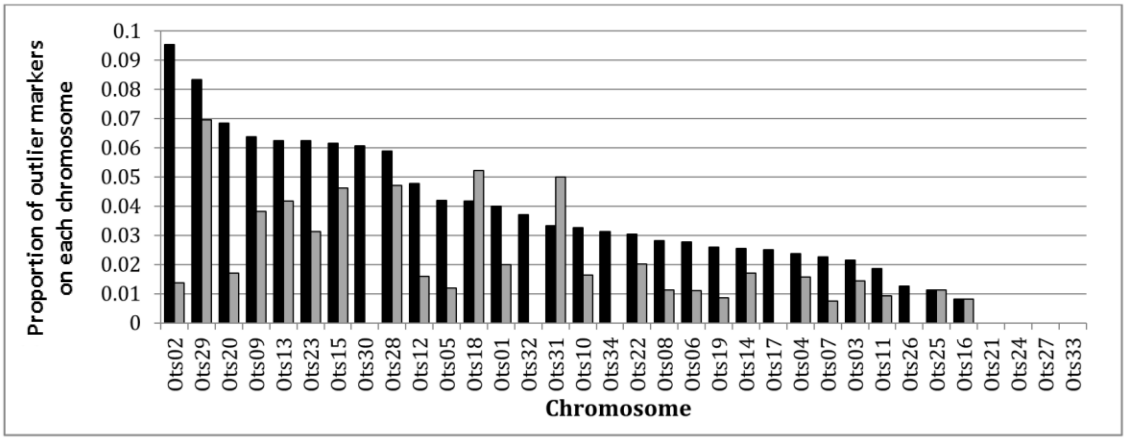


Figure 3.14 - Proportion of mapped outlier markers on each linkage group for the analysis with all populations (black) and for the analysis with populations from the Interior Columbia Su/F and the Lower Columbia only (grey) (3528 mapped loci).

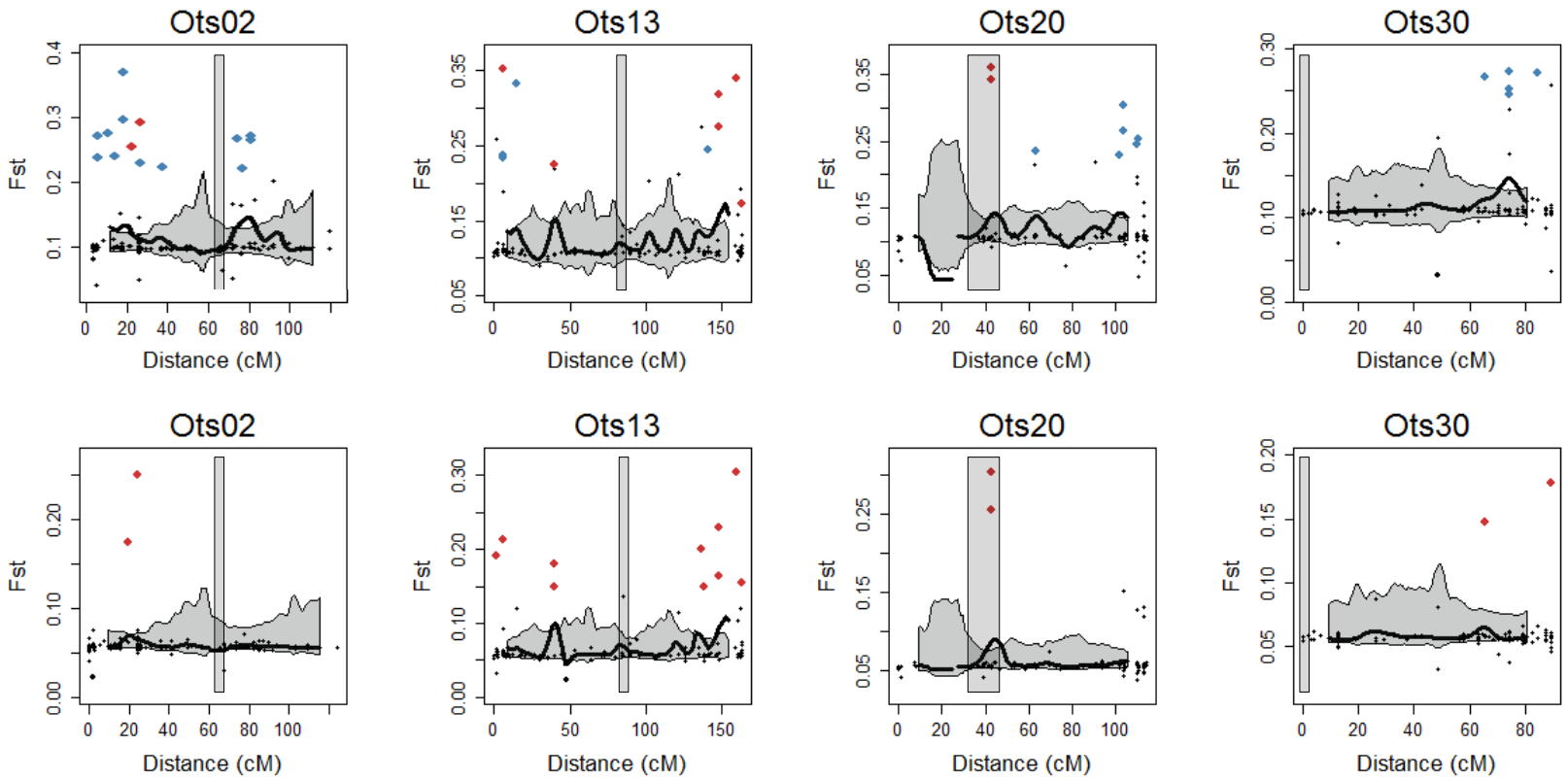


Figure 3.15 - Moving average of $F_{ST}$ for mapped loci along some linkage groups comparing analyses based on all the populations from the Columbia River (top) or populations from the Interior Columbia Su/F and Lower Columbia lineages (bottom). The black line corresponds to the moving average calculated from $F_{ST}$ values estimated with BAYESCAN. The grey area is the 95% Confidence Interval obtained with 100000 bootstraps. Each dot corresponds to a locus. Red and blue dots represent outlier loci classified as involved in pre- and post-glaciation divergence respectively as identified with BAYESCAN (FDR=0.01). Discontinuities of the moving average correspond to regions with no polymorphic locus. The centromere of each chromosome is represented by the grey bar.

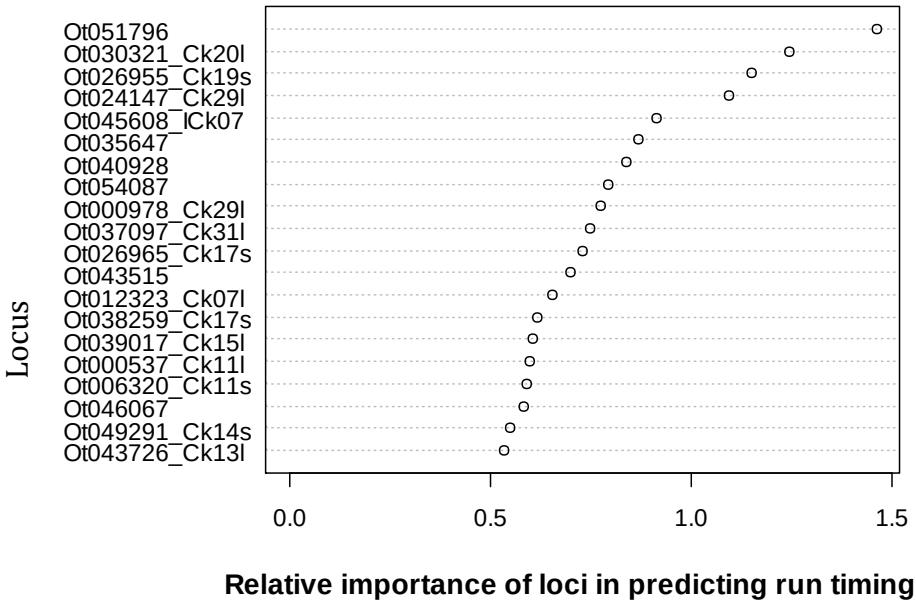


Figure 3.10 – Twenty best predictive loci for run timing (Fall vs. Spring) in the Columbia River and Puget Sound, and their relative ranking in importance, identified with the Random Forest approach.

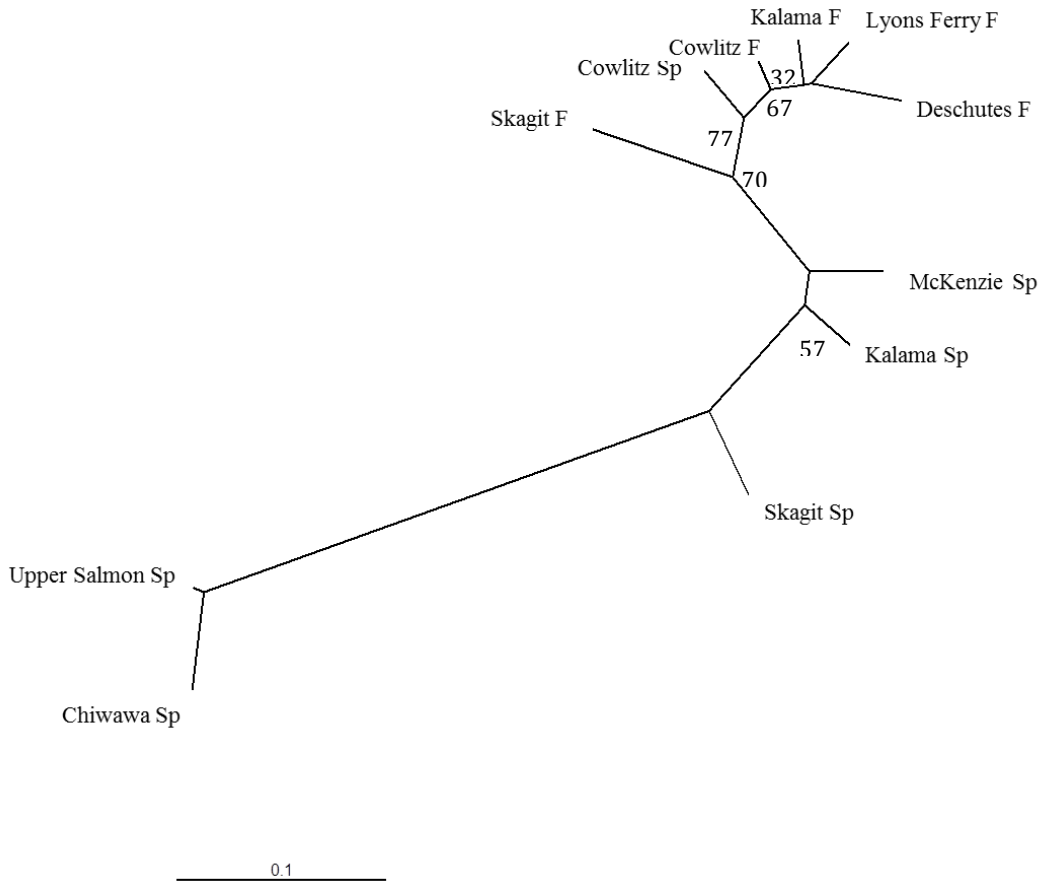


Figure 3.11- Neighbor-joining dendogram of Cavalli-Sforza and Edwards (1967) chord distances for 11 populations from the Columbia River and Puget Sound constructed using the best 10 best putative descriptors of run timing (Fall and Spring only) identified with the Random Forest approach. Bootstrap values 80% of less are indicated on the tree.

## References


Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. B.J. T). Plenum Press, New York.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Beacham TD, Jonsen KL, Supernault J*, et al.* (2006) Pacific Rim population structure of Chinook salmon as determined from microsatellite analysis. *Transactions of the American Fisheries Society* **135**, 1604-1621.

Bourret V, Kent MP, Primmer CR*, et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology* **22**, 532-551.

Bradbury IR, Hubert S, Higgins B*, et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, n/a-n/a.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Cavallis.Ll, Edwards AWF (1967) Phylogenetic analysis – Models and estimation procedures. *Evolution* **21**, 550- &.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B-Biological Sciences* **279**, 5039-5047.

Corander J, Waldmann P, Marttinen P, Sillanpaa MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363-2369.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT (2007) Random forests for classification in ecology. *Ecology* **88**, 2783-2792.

Dray S, Dufour AB (2007) The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* **22**, 1-20.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution* **26**, 298-306.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol Biol* **772**, 157-178.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes – Application to human microchondrial-DNA restriction data. *Genetics* **131**, 479-491.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**, 977-993.

Fraser DJ, Weir LK, Bernatchez L, Hansen MM, Taylor EB (2011) Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. *Heredity* **106**, 404-420.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184-186.

Gustafson RG, Waples RS, Myers JM*, et al.* (2007) Pacific salmon extinctions: Quantifying lost and remaining diversity. *Conservation Biology* **21**, 1009-1020.

Gutierrez NL, Hilborn R, Defeo O (2011) Leadership, social capital and incentives promote successful fisheries. *Nature* **470**, 386-389.

Hauser L, Seamons TR, Dauer M, Naish KA, Quinn TP (2006) An empirical verification of population assignment methods by marking and parentage data: hatchery and wild steelhead (Oncorhynchus mykiss) in Forks Creek, Washington, USA. *Molecular Ecology* **15**, 3157-3173.

Healey MC (1991) Life history of chinook salmon (Oncorhynchus tshawytscha). In: *Pacific Salmon Life Histories* (eds. Groot C, Margolis L), pp. 311-393. UBC Press, Vancouver.

Hemmer-Hansen J, Nielsen EE, Therkildsen NO*, et al.* (2013) A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology* **22**, 2653-2667.

Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *Plos Genetics* **6**.

Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *Plos Genetics* **5**.

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

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.

Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* **106**, 625-632.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**.

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype Imputation. In: *Annual Review of Genomics and Human Genetics*, pp. 387-406.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* **2**, 18-20.

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128-2129.

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906-913.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**, 240-248.

Moran P, Teel DJ, Banks MA*, et al.* (2012) Divergent life-history races do not represent Chinook salmon coast-wide: the importance of scale in Quaternary biogeography. *Canadian Journal of Fisheries and Aquatic Sciences* **70**, 415-435.

Narum SR, Hess JE (2011) Comparison of F-ST outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**, 184-194.

Narum SR, Hess JE, Matala AP (2010) Examining Genetic Lineages of Chinook Salmon in the Columbia River Basin. *Transactions of the American Fisheries Society* **139**, 1465-1477.

Narum SR, William DA, Talbot AJ, Powell MS (2007) Reproductive isolation following reintroduction of Chinook salmon with alternative life histories. *Conservation Genetics* **8**, 1123-1132.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Page RDM (1996) TreeView: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357-358.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Quinn TP (1993) A review of homing and straying of wild and hatchery-produced salmon. *Fisheries Research* **18**, 29-44.

Quinn TP (2005) *The behavior and ecology of Pacific salmon and trout* American Fisheries Society, Bethesda, Md.

Quinn TP, Kinnison MT, Unwin MJ (2001) Evolution of chinook salmon (Oncorhynchus tshawytscha) populations in New Zealand: pattern, rate, and process. *Genetica* **112**, 493-513.

Quinn TP, Nielsen JL, Gan C*, et al.* (1996) Origin and genetic structure of chinook salmon, Oncorhynchus tshawytscha, transplanted from California to New Zealand: Allozyme and mtDNA evidence. *Fishery Bulletin* **94**, 506-521.

Quinn TP, Unwin MJ, Kinnison MT (2000) Evolution of temporal isolation in the wild: Genetic divergence in timing of migration and breeding by introduced chinook salmon populations. *Evolution* **54**, 1372-1385.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raymond M, Rousset F (1995) Genepop (version-1.2) – Population-genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629-644.

Seeb LW, Antonovich A, Banks AA*, et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries* **32**, 540-552.

Taylor EB (1991) A review of local adaptation in salmonidae, with particular reference to Pacific and Atlantic salmon. *Aquaculture* **98**, 185-207.

Tsumura Y, Uchiyama K, Moriguchi Y, Ueno S, Ihara-Ujino T (2012) Genome scanning for detecting adaptive genes along environmental gradients in the Japanese conifer, Cryptomeria japonica. *Heredity* **109**, 349-360.

Utter F, Milner G, Stahl G, Teel D (1989) Genetic population-structure of Chinook salmon, *Oncorhynchus tshawytscha*, in the Pacific Northwest. *Fishery Bulletin* **87**, 239-264.

Waples RS, Gustafson RG, Weitkamp LA*, et al.* (2001) Characterizing diversity in salmon from the Pacific Northwest. *Journal of Fish Biology* **59**, 1-41.

Waples RS, Pess GR, Beechie T (2008) Evolutionary history of Pacific salmon in dynamic environments. *Evolutionary Applications* **1**, 189-206.

Waples RS, Teel DJ, Myers JM, Marshall AR (2004) Life-history divergence in Chinook salmon: Historic contingency and parallel evolution. *Evolution* **58**, 386-403.

Weeder JA, Marshall AR, Epifanio JM (2005) An assessment of population genetic variation in Chinook salmon from seven Michigan rivers 30 years after introduction. *North American Journal of Fisheries Management* **25**, 861-875.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358-1370.

Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation: when evolution repeats itself. *Genetica* **123**, 157-170.

## Acknowledgments

I would like to thank Washington Sea Grant, the School of Aquatic and Fishery Sciences, the Roy Jensen Research Fellowship, the Melvin G. Anderson Scholarship in Fisheries, the H. Mason Keeler Endowments and the Fidalgo San Juan Island - Puget Sound Anglers for their support.

I would like to sincerely thank my friends in Seattle and in France, who have supported me all these years, especially Kristi Straus who, through her passion for science, inspired me to pursue my career in research.

I am thankful to have had Kerry Naish as an advisor and as a mentor. Her knowledge, advice and patience have been invaluable.

I sincerely thank my parents who have been so understanding of my projects. Letting me move across the world wasn't easy, but they supported me every step in the way. I am thankful for my sister and my brother-in-law who always manage to be around at times that really matter. Finally, I am deeply grateful for the love and support of my husband, Kotaro Ono, without whom I would not have started this amazing journey in the first place!

## Vita

Marine Brieuc was born in Rennes, France. Growing up, she spent her vacation in the country side, where she enjoyed helping milk cows and going on long bike rides, and on the North coast of Brittany where she couldn't get enough of fishing, clamming and swimming. Her passion for the land and the sea naturally led her to pursue her education in an agronomic school, Ecole Nationale Superieure Agronomique de Rennes (France), where she received an engineering degree with a minor in aquatic and fishery sciences and a focus in aquaculture in 2007. Her interest for research and molecular biology developed during an internship with Dr. Kerry Naish and Dr. Kristina Straus in the School of Aquatic and Fishery Sciences at the University of Washington in 2006. She then returned to this institution in 2007 to start working on Chinook salmon evolution and adaptation with Dr. Kerry Naish. She earned her PhD from this school in 2013.