ECOLOGICAL GENOMICS OF KOKANEE SALMON

by

Matthew Alexander Lemay

B.Sc., University of Victoria, 2002
M.Sc., University of Guelph, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLEGE OF GRADUATE STUDIES

(Biology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

July 2014

© Matthew Alexander Lemay, 2014

Abstract

Divergent natural selection across a heterogeneous landscape can drive the evolution of locally adapted populations in which phenotypic variation is fine-tuned to the environment. At the molecular level, such processes can be inferred by identifying correlations between genetic variation and environmental variables. In this dissertation, I used multiple complementary approaches to investigate the genetic basis of adaptation in natural populations of kokanee, the freshwater form of sockeye salmon (*Oncorhynchus nerka*). In Chapter 2, I found that the frequency and length of alleles in the circadian regulation gene, *OtsClock1b*, displays a predictable distribution with respect to latitude among lakes sampled from British Columbia to Alaska, providing evidence that variation at this locus may be locally adapted to divergent photoperiod regimes. In Chapter 3, I tested for transcriptome-wide patterns of sequence divergence in reproductive ecotypes of kokanee within Okanagan Lake and found evidence for differential gene expression and asymmetrical pathogen load between ecotypes. In Chapter 4, I used restriction site associated DNA sequencing to identify ~6,000 single nucleotide polymorphisms (SNPs) from multiple spawning populations of kokanee within Okanagan Lake; statistical outlier tests revealed 20 SNPs that were putatively under divergent selection between ecotypes, many of which annotated to genes associated with early development. While there was no evidence for neutral genetic divergence, outlier SNPs demonstrated significant structure with respect to ecotype and had high assignment accuracy (>99%) in mixed composition simulations, suggesting the utility of these loci for genetic stock identification. These data support the hypothesis that kokanee ecotypes are in the early stages of ecological differentiation, making them an ideal system for investigating the genomic basis of adaptation. Results from this study will be used to assist conservation

and management initiatives by providing molecular tools for in-season monitoring of ecotype abundance.

Preface

The following sections of this thesis have been published or submitted to peer reviewed journals:

Chapter 2:

Lemay, M.A. and Russello, M.A. (2014) Latitudinal cline in allele-length provides evidence for selection in a circadian rhythm gene. *Biological Journal of the Linnean Society* 111: 869-877.

Michael Russello and I designed the study. I carried out the laboratory assays, performed the data analyses, and drafted the manuscript.

Chapter 3:

Lemay, M.A., Donnelly, D.D., and Russello, M.A. (2013) Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon. *BMC Genomics* 14: 308.

Michael Russello and I designed the study. I carried out the field sampling, laboratory assays, data analyses, and drafted the manuscript. Dave Donnelly assisted with laboratory assays.

Chapter 4:

Lemay, M.A. and Russello, M.A. (in revision) Genomic evidence for early ecological divergence in kokanee salmon. *Molecular Ecology*

Michael Russello and I designed the study. I carried out all laboratory preparations, performed the data analyses, and drafted the manuscript.

During the course of my PhD research I undertook a number of collaborative side-projects that were conceptually linked to the research chapters of this dissertation. While these manuscripts are not included as chapters in my thesis, they were informed by my genomic research with kokanee:

1. Lemay, M.A. and Russello, M.A. (2012) Neutral loci reveal population structure by geography, not ecotype, in Kootenay Lake kokanee. *North American Journal of Fisheries Management* 32(2): 282-291.
2. Lemay, M.A., Henry, P., Lamb, C.T., Robson, K.M., and Russello, M.A. (2013) Novel genomic resources for a climate change sensitive mammal: characterization of the American pika transcriptome. *BMC Genomics* 14: 311.
3. Shalev, T., Lemay, M.A., Conrad, K., Steinfartz, S., and Russello, M.A. (in prep.) Genomic divergence between stream and pond ecotypes of fire salamander. *Journal of Zoology*.

All animal collections and tissue sampling complied with University of British Columbia animal care protocol #A11-0127 and BC Ministry of Environment collection permit #PE10-66394.

Table of Contents

Abstract ii

Preface iv

Table of Contents vi

List of Tables ix

List of Figures xi

Acknowledgements xiii

Chapter 1: Introduction 1

1.1 THE GENETIC BASIS OF ADAPTATION 1

1.2 TOP DOWN TESTS FOR SELECTION 2

1.3 BOTTOM UP TESTS FOR SELECTION 3

1.4 REDUCED REPRESENTATION DNA SEQUENCING 4

1.5 ADAPTIVE DIVERGENCE IN POST-GLACIAL LAKE FISHES 5

1.6 KOKANEE SALMON 6

1.7 THESIS OBJECTIVES 8

Chapter 2: Latitudinal cline in allele length provides evidence for selection in a circadian rhythm gene 10

2.1 BACKGROUND 10

2.2 METHODS 12

2.2.1 Tissue sampling and molecular methods 12

2.2.2 Data analyses 13

2.2.3 Ecotype differentiation 15

2.3 RESULTS 16

2.3.1 *OtsClock1b* diversity 16

2.3.2 Latitudinal clines 17

2.3.3 Ecotype differentiation 18

2.4 DISCUSSION 18

Chapter 3: Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon 29

3.1 BACKGROUND 29

3.2 METHODS 32

3.2.1 Sample collection, RNA extraction, and next-generation sequencing 32

3.2.2 Transcriptome assembly 33

3.2.3 Mitochondrial genome 33

3.2.4 Transcriptome analysis and annotation 34

3.2.5 SNP discovery 35

3.2.6 SNP validation 36

3.3 RESULTS AND DISCUSSION 38

3.3.1 Sequencing and assembly 38

3.3.2 Transcriptome analyses and annotation 39

3.3.3 Mitochondrial genome 40

3.3.4 SNP detection 40

3.3.5 SNP validation 42

3.3.6 Ecotype unique data 44

3.3.7 Conclusions 46

Chapter 4: Genomic evidence for ecological divergence in kokanee 56

4.1 BACKGROUND 56

4.2 METHODS 59

4.2.1 Sample collections 59

4.2.2 Molecular methods 59

4.2.3 Assembly and SNP discovery 60

4.2.4 Population genetic analyses 62

4.2.5 Mixed stock analyses 63

4.2.6 Genetic stock identification 64

4.3 RESULTS 65

4.3.1 Sequencing and assembly 65

4.3.2 Outlier loci 65
4.3.3 Population genetic analyses 66
4.3.4 Mixed composition analyses 67
4.3.5 Genetic stock identification 68
4.4. DISCUSSION 68
4.4.1 Ecological divergence 68
4.4.2 Fisheries management 70
4.4.3 Summary 72

Chapter 5: Conclusions 85

5.1 RESEARCH FINDINGS 85
5.2 LIMITATIONS AND FUTURE DIRECTIONS 86
5.3 SIGNIFICANCE 88

Bibliography 89

Appendices 102

APPENDIX A: Diversity of the bacterial pathogen, *Flavobacterium sp*., infecting reproductive ecotypes of kokanee salmon. 102

List of Tables

Table 2.1: Kokanee samples from British Columbia (BC), Alaska (AK), and Russia. Length variation is reported as the frequency of allele 370 and mean allele length (MAL). 26

Table 2.2: Analysis of molecular variance (AMOVA) results showing the percentage of variation at each hierarchical level of organization for *OtsClock1b* and the neutral microsatellites. For this analysis, Group = latitude (categorized as high, mid, low; see main text for more details related to grouping strategy), and population = lake 27

Table 2.3: Pairwise F_{ST} for each population at (A) four neutral loci and (B) *OtsClock1b*. For each dataset, *p*-values are above the diagonal and F_{ST} -values are below the diagonal. F_{ST}-values in bold are significant at $p<0.05$ (uncorrected for multiple tests). 28

Table 3.1: Collection information of all individuals used for each pooled cDNA library 51

Table 3.2: Summary of next-generation sequence data obtained for each ecotype of Okanagan Lake kokanee. 52

Table 3.3: Summary of the contigs present in each kokanee data set. 53

Table 3.4: Genetic diversity estimates from loci that were successfully genotyped using High Resolution Melt Analysis (HRMA). 54

Table 3.5: Number of next-generation sequencing reads that aligned to reference sequences from four salmonid pathogens. 55

Table 4.1: Distribution of samples used to generate RAD-libraries for Illumina sequencing. Samples were collected from seven locations in Okanagan Lake, encompassing two ecotypes (stream & shore) and two spawning years (2007 & 2010). 76

Table 4.2: Single nucleotide polymorphisms (SNPs) retained at each step of the quality filtering process. 77

Table 4.3: Summary of outlier loci identified using BAYESCAN at false discovery rate (FDR) = 0.2; loci that were identified as outliers at FDR=0.05 are indicated with an asterisk. 78

Table 4.4: Pairwise F_{ST} for each comparison between sampling locations and years. Comparisons between sites from different ecotypes are indicated in blue font. Comparisons between different years of the same ecotype are indicated with red font. Values of F_{ST} that significantly differ from zero are indicated with an asterisk (uncorrected $p < 0.05$). 81

Table 4.5: Analysis of molecular variance (AMOVA) results showing the percentage of variation at each hierarchical level of organization for neutral and outlier loci. For this analysis, 'group' = ecotype (stream and shore), and 'population' is each sampling location from each year (n = 14). 82

Table 4.6: Assignment accuracy for the 100% mixture simulations following the method of Anderson *et al*. (2008) as implemented in ONCOR. 83

Table 4.7: Estimated proportion of stream and shore-spawners from each of seven simulated mixture compositions carried out using realistic fisheries simulations implemented in ONCOR. 84

List of Figures

Figure 2.1: Sampling localities used to quantify genetic variation at the *OtsClock1b* locus in kokanee salmon from North America and Russia (inset). Additional location information is provided in Table 1. 22

Figure 2.2: Multiple sequence alignment of the two *OtsClock1b* alleles identified in kokanee salmon along with the eight alleles previously published by O'Malley and Banks (2008). Kokanee_1 is allele 370 from the current study; calibrated to O'Malley et al. (2010) it would be allele 353. Kokanee_2 is allele 391 from the current study; calibrated to O'Malley et al. (2010) it would be allele 374. 23

Figure 2.3: Frequency of the most common OtsClock1b allele (allele 370) for each sampled population of kokanee salmon as a function of latitude. 24

Figure 2.4: Graphical representation of the distribution of (a) R-squared values, and (b) regression coefficients (slopes) for the relationship between allele frequency and latitude at each allele sampled. There are two alleles from the candidate locus, *OtsClock1b* (indicated as shaded circles), and 75 alleles from neutral microsatellite loci. 25

Figure 3.1: Characterization of the contigs present in the high coverage data set**.** Histograms represent (A) average coverage of each contig (mean = 37.0), (B) number of reads (mean = 77.9), (C) contig lengths (mean = 594.8 bases), and (D) the number of SNPs for each of the high coverage contigs. 48

Figure 3.2: Functional annotation of the high coverage contigs. The frequency (%) of each observed gene ontology (GO) term is given for the three GO domains (biological process, cellular component, and molecular function). 49

Figure 3.3: Functional annotation of contigs that were unique to each ecotype. The frequency (%) of each observed gene ontology (GO) term is presented for both ecotypes. 50

Figure 4.1: A map of Okanagan Lake identifying each of the seven sampling locations used in this study. Detailed information about each site is included in Table 1. 73

Figure 4.2: Frequency distribution of F_{ST} values obtained across all loci (neutral and outlier; n = 5976) using the method of Weir and Cockerham (1984) followed by Fisher's exact tests for statistical significance..... 74

Figure 4.3: Results of the Bayesian clustering analysis carried out for all individuals (n = 96) of known ecotype as implemented in the program STRUCTURE: (A) visual representation of *Ln P(K)* for all neutral loci (5956 loci); (B) visual

representation of *Ln P(K)* for all outlier loci (20 loci). Both analyses were run with 100,000 steps and a burn in of 50,000, with five replicates of each value of *K* ranging from 1-15. Based on the delta-*K* method of Evanno et al. 2005 an optimal *K* = 2 was retained (ΔK = 1385.4); (C) A barplot displaying the percent membership of each individual to the inferred number of clusters (*K* = 2) using outlier loci. 75

Figure A:1 Unrooted neighbor-joining tree showing the relationship between the unique *Flavobacterium* haplotypes amplified from each ecotype (n = 13 stream, n = 8 beach) and nine previously published reference sequences (indicated with bold font). Genbank accession numbers are included next to species names. Branch labels are bootstrap support values (%)..... 107

Figure A:2 Relative proportion of *Flavobacterium sp*. DNA amplified from the operculum tissue of kokanee sampled at stream and beach spawning sites in Okanagan Lake, BC. Samples are separated by year: (A) 2007 (n = 48) and (B) 2010 (n = 48). 108

Acknowledgements

This research would not have been possible without the biological insight, lab assistance, field assistance, analytical advice, and editorial suggestions from my advisory committee, fellow graduate students, and researchers at UBC.

In particular, Mike Russello has been an exceptional advisor and mentor. He has lead by example in terms of how to be a happy and productive scientist, and how to integrate multidisciplinary approaches in order to ask interesting biological questions. I have greatly benefited from his insightful feedback and support at every stage of my graduate work at UBC.

I have deep respect for the members of my advisory committee, whose diverse expertise has fostered interesting discussion along the way. Paul Askey provided essential insight into kokanee natural history as well as much needed advice on experimental design and statistical analyses. Similarly, Jim Seeb's expertise in salmon genomics has been an invaluable resource during my graduate studies. John Klironomos and Bob Lalonde have greatly improved the quality of this dissertation through their insightful feedback during committee meetings and on earlier drafts of these chapters.

I am thankful to the past and present members of the Ecological and Conservation Genomics Lab for helpful discussions along the way. In particular, feedback from Philippe Henry, Evelyn Jensen, and Karen Frazer has significantly improved this thesis.

Eric Taylor, the Beaty Biodiversity Museum DNA Archive, and Jim Seeb provided DNA samples. Jason Pither provided helpful statistical advice. Dave Donnelly, Tal Shalev, and Karen Frazer provided laboratory assistance. Mark Rheault and Deanna Gibson provided the use of laboratory equipment. Ana Kuzmin from the biodiversity sequencing facility at

UBC was very helpful during the library construction in Chapter 4, and carried out the Illumina sequencing. Julian Catchen provided prompt troubleshooting advice during the data analyses in Chapter 4. I thank Eric Duchaud for the kind donation of *Flavobacterium sp*. DNA standards described in the appendix. Special thanks go to Barb Lucente and Jennifer Janok who provided the logistical support essential for navigating the bureaucratic aspects of grad school.

I thank Katy Hind for her never-ending support, patience, and insight.

This research was funded by several sources including an NSERC Discovery Grant, an NSERC Engage Grant, and a grant from Genome BC, all of which were to Michael Russello. I was partially supported by University Graduate Fellowships from UBC and by an NSERC Postgraduate Scholarship.

Chapter 1: Introduction

1.1 THE GENETIC BASIS OF ADAPTATION

Population differentiation, and ultimately speciation, requires the formation of barriers to gene flow. Understanding the mechanisms responsible for creating and maintaining these reproductive barriers has long been a central question in evolutionary biology (Coyne 1992; Boughman *et al*. 2005), with early research focused primarily on the impact of physical barriers to gene flow (Coyne & Orr 2004). For example, Mayr (1942) believed that geographical isolation was essential for population divergence, and that in the absence of gene flow, allopatric populations would differentiate as the result of accumulated genetic differences over time (Coyne 1994). However, he conceded that at smaller geographical scales, adaptation to divergent ecological conditions could also lead to population differentiation, even when there is some level of gene flow (Mayr 1947; Nosil 2008). This latter idea of an ecological mechanism to population differentiation has received much attention in recent literature (Schluter 1996a; 2001; McKinnon *et al*. 2004; Egan *et al*. 2008; Nosil *et al*. 2009; Schluter 2009; Nosil 2012). Yet the extent to which adaptation drives population divergence remains poorly understood and challenging to quantify (Freeland *et al*. 2010).

Demonstrating the role of adaptation in population divergence ultimately requires identifying a causal link between genotype, phenotype, and fitness (Gould & Lewontin 1979; Barrett & Hoekstra 2011). However for many natural systems this is not easily accomplished; factors such as a lack of information on the genetic basis of phenotypic diversity or the unfeasibility of carrying out experimental transplants to test for fitness

differences can preclude direct tests for adaptation. When these factors exist, the identification of a correlation between either genotypic or phenotypic variation and environmental variables can provide indirect evidence for the presence of selection in natural populations (Endler 1986). At the molecular level, tests for adaptive population divergence generally take one of two different forms: the top down approach tests for habitat-specific patterns in the segregation of genetic variation at genes that are *a priori* candidates for selection (Dalziel *et al*. 2009); while the alternative, bottom up approach, uses genome scans to statistically identify loci that deviate from a neutral model of evolution (Lewontin & Krakauer 1973; Beaumont & Balding 2004).

1.2 TOP DOWN TESTS FOR SELECTION

Candidate gene studies infer natural selection by identifying a correlation between environmental variables and the frequency of alleles present at genes of known function (Barrett & Hoekstra 2011). These studies provide strong evidence for selection by linking both phenotypic and genotypic variation with landscape heterogeneity (Fraser *et al*. 2011). A classic example where candidate genes have been used to study adaptive evolution comes from marine and freshwater populations of threespine stickleback (*Gasterosteus sp*.). In the ancestral marine environment, stickleback are heavily armored with lateral bony plates that protect them from marine predators. In freshwater populations, reduced predation pressure has resulted in a reduction in the number of lateral plates, which confers a growth advantage (Barrett *et al*. 2008). The genetic basis of lateral plate coverage in sticklebacks has been linked to variation at the *Ectodysplasin* (*Eda*) locus (Colosimo *et al*. 2005); individuals homozygous for the ancestral *Eda* allele have complete coverage of lateral plates, individuals

homozygous for the derived (low) allele have few lateral plates, and heterozygous individuals have an intermediate number of plates (Schluter *et al*. 2010). Parallel evolution of the derived form in freshwater lakes provides strong evidence for natural selection (Rundle *et al*. 2000; Colosimo *et al*. 2005; Schluter *et al*. 2010). A direct relationship between genotype, phenotype, and fitness has since been confirmed through the use of experimental transplant experiments (Barrett *et al*. 2008).

While the use of a top-down approach can provide compelling evidence for the presence of natural selection, it relies on the presence of overt phenotypic differences among habitats, and corresponding knowledge of the genetic architecture underlying this phenotypic variation. For these reasons, the use of candidate genes is often limited to well-studied systems with pre-existing genetic resources.

1.3 BOTTOM UP TESTS FOR SELECTION

The alternate molecular approach for inferring natural selection is the use of genome scans to identify regions of divergence in wild populations that are undergoing adaptive differentiation (Turner *et al*. 2005; Nosil *et al*. 2009; Andrew & Rieseberg 2013; Renaut *et al*. 2013). Unlike the candidate gene method that focuses on the segregation of genetic diversity at targeted loci, genome scans utilize a bottom up approach in which potentially adaptive genes are identified statistically from a large baseline of neutral loci, precluding the need for *a priori* hypotheses about the phenotypic basis of selection (Luikart *et al*. 2003).

Central to this approach is the identification of outlier loci, which are genes that deviate from a neutral model of evolution when compared among different populations (Luikart *et al*. 2003; Storz 2005; Allendorf *et al*. 2010). Patterns of nucleotide divergence at

these loci provide evidence that natural selection, rather than neutral processes, govern their distribution (Lewontin & Krakauer 1973; Nielsen 2001; Beaumont & Balding 2004; Nosil *et al*. 2009). Previous research has shown that the use of putatively adaptive outlier loci can identify fine-scale genetic structure beyond what can be observed with only neutral loci (Westgaard & Fevolden 2007; André *et al*. 2011; Freamo *et al*. 2011; Russello *et al*. 2012). Further, with the rapidly expanding literature of genomic data, it is increasingly possible to annotate outlier loci to genes of known function (Galindo *et al*. 2010; Prunier *et al*. 2012; Frazer & Russello 2013).

1.4 REDUCED REPRESENTATION DNA SEQUENCING

Despite the enhanced data-gathering ability provided by next-generation sequencing technology, obtaining entire genome sequences for non-model organisms is still prohibitively expensive for most labs (Peterson *et al*. 2012). Further, obtaining sequence data from every nucleotide in the genome would generally provide more data than necessary to address basic questions in ecology and evolution. For these reasons, researchers in the field of population genomics employ reduced representation approaches to genome sequencing, which provide an abundance of molecular data across a biologically informative subset of the genome (Narum *et al*. 2013). This dissertation utilizes two methods for generating reduced representation libraries, transcriptome sequencing (RNA-seq) and Restriction Site Associated DNA sequencing (RAD-seq).

A transcriptome is the set of all RNA sequences that have been transcribed from DNA at a given point in time, providing a snapshot of gene expression in the sampled cells. Given that the identity and abundance of genes expressed will vary depending on the tissue

being sampled and the stimulus being experienced (i.e. factors such as homeostatic stress can alter patterns of gene expression), RNA-seq facilitates comparisons of both nucleic acid sequence and level of gene expression between different treatments (Wang *et al*. 2009). Further, by targeting expressed regions of the genome, RNA-seq focuses on genomic regions that directly influence the phenotype of the study organism, providing a powerful tool to study adaptation (Jeukens *et al*. 2010; Renaut *et al*. 2010).

Unlike transcriptome sequencing, which specifically targets functional regions of the genome, alternative methods of reduced representation sequencing that use restriction enzymes (REs) provide a tool for randomly subsampling the genome. As with transcriptome sequencing, this method provides a rapid, cost effective, method for generating genome-wide sequence data in wild populations of non-model species (Baird *et al*. 2008; Etter *et al*. 2011; Peterson *et al*. 2012); these methods work by digesting genomic DNA using restriction enzymes and then ligating sequencing adapters to the resulting RE recognition sequence at the terminus of the cleaved DNA. Given that RE recognition sequences are abundant, are homogeneously distributed across the genome, and are expected to occur at the same location among individuals of the same species, this methods provides a framework for reduced representation sequencing that provides consistent loci among different individuals (Peterson *et al*. 2012). Consequently, RAD-seq is becoming a standard approach for population genomic analysis of large numbers of individuals from natural populations (Davey *et al*. 2011; Davey *et al*. 2012; Narum *et al*. 2013).

1.5 ADAPTIVE DIVERGENCE IN POST-GLACIAL LAKE FISHES

Fish inhabiting post-glacial lakes provide ideal systems for studying the effect of ecological

factors on the evolution of reproductive isolation (Schluter 1996a; b; Hendry *et al*. 2000). Following the last glacial maximum, newly formed lakes were colonized by many species of fish that have since undergone rapid adaptive radiations to fill available niches (reviewed by Schluter 1996b). In these systems, parallel evolution of the same traits in replicate lakes provides strong evidence for natural selection (Hendry 2009). For example lake whitefish (*Coregonus sp*.) and threespine stickleback have both undergone trophic divergence in freshwater lakes, in which each species have evolved a larger benthic ecotype and a small planktivorous limnetic ecotype, with divergent phenotypes locally adapted to their respective ecological niches (Schluter 1995; 1996b; Rundle *et al*. 2000; Bernatchez *et al*. 2010). The trophic partitioning observed in the two forms of these species provides strong evidence for ecologically mediated population divergence.

1.6 KOKANEE SALMON

The Pacific salmonid, *Oncorhynchus nerka*, occurs in a variety of divergent forms, which deviate in their reliance on the freshwater stage of their life cycle (reviewed by Wood *et al*. 2008). Sockeye salmon is the anadromous form, which has an initial freshwater stage in nursery lakes before migrating to sea, whereas kokanee are a non-anadromous form that spend their entire life cycle in freshwater (Godbout *et al*. 2011). Both forms of *O. nerka* are semelparous and highly philopatric.

Kokanee populations are polyphyletic, having evolved from anadromous sockeye salmon through multiple independent freshwater colonization events (Taylor *et al*. 1996; McGurk 2000; Godbout *et al*. 2011). Within kokanee, two divergent reproductive ecotypes have been identified in lakes across their range. The ‘stream-spawning’ ecotype exhibits

typical sockeye behaviour in that they migrate into tributaries to spawn in early autumn. Conversely, the 'shore-spawning' ecotype forgoes a tributary migration and instead spawns directly on the submerged shoreline of lakes (Taylor *et al*. 2000). Outside of the spawning season, both ecotypes co-occur in many lakes and are morphologically indistinguishable (Taylor *et al*. 1997).

Reproductive ecotypes of kokanee are distinct from other post-glacial lake fish species in that their divergence does not appear to be associated with trophic differentiation (Taylor *et al*. 1997). Rather the only observable differences between ecotypes are associated with behavioral differences in the choice spawning habitat and the timing of spawning. However, the parallel evolution of these divergent reproductive ecotypes provides evidence that they are the result of natural selection (Hendry 2009), and may therefore have a genetic basis.

In British Columbia, kokanee are a heavily managed recreational fishery of great socioeconomic and ecological importance (Thompson 1999). In many lakes, kokanee populations have experienced sharp declines resulting from factors such as competition with invasive species, reduced nutrient input, and loss of spawning habitat (Thompson 1999; Askey & Johnston 2013). Conservation strategies for kokanee tend to manage the reproductive ecotypes as distinct stocks (Taylor *et al*. 2000; Askey & Johnston 2013), yet a lack of morphological differences hinders in-season monitoring of relative stock abundance.

To improve the accuracy of population estimates, kokanee management would be significantly enhanced by the identification of molecular markers capable of assigning individuals to the correct ecotype, however the identification of sufficiently fine-scale markers has been hindered by the recent divergence (<12,000 years) and correspondingly

weak genetic structure of the two ecotypes (Lemay & Russello 2012; Russello *et al*. 2012; Frazer & Russello 2013). Russello *et al*. (2012) carried out a population genomic assessment of ecotype differentiation in Okanagan Lake. Using a panel of 52 expressed sequence tag (EST)-linked and non-EST-linked microsatellites, they identified eight EST-linked loci that were statistical outliers. While they found no evidence for genetic divergence at neutral loci, the outlier loci showed significant population genetic structuring between ecotypes and had high assignment accuracy for stock identification, providing preliminary evidence for restricted gene flow and adaptive population divergence. Yet the limited genomic coverage provided by a small number of microsatellite markers makes it difficult to draw definitive conclusions.

1.7 THESIS OBJECTIVES

This thesis uses three complimentary approaches to investigate the genetic basis of adaptation in kokanee salmon. In Chapter 2, I apply a top down, candidate locus approach in order to test for patterns of adaptation to divergent photoperiod regimes. Using population samples collected across a latitudinal transect from southern British Columbia to Alaska, I demonstrate that the frequency and length of alleles present at the circadian regulation gene, *OtsClock1b*, displays a predictable distribution with respect to latitude, providing evidence that genetic variation at this locus is locally adapted to divergent photoperiod regimes.

In Chapter 3, I use a bottom up approach in order to test for divergence between reproductive ecotypes of kokanee in Okanagan Lake. This study used the Roche 454 platform in order to sequence transcriptomes from pooled samples of each ecotype. A subset of SNPs identified using this approach were validated on independent samples using High

Resolution Melt Analysis. I found evidence for differences in gene expression between ecotypes that were sampled during spawning, as well as evidence for differing levels of pathogen load between ecotypes (Appendix 1). This is the first time that transcriptome sequencing has been carried out for kokanee, providing novel resources for future studies of kokanee population genomics.

Chapter 4 builds on the bottom up approach of Chapter 3, by expanding from the transcriptome to a large random subset of the genome. This study employed RAD-seq in order to generate genome-wide sequence data for 48 individuals of each ecotype from Okanagan Lake. Using a large baseline of ~6,000 SNPs, I identified and annotated a suite of 20 high F_{ST} outlier loci. As proof of concept, these loci were then used to genotype an independent sample of kokanee collected by trawl in Okanagan Lake, illustrating their potential utility as a fine-scale molecular tool for conservation and management.

These findings contribute to a growing body of research on the genetic basis of adaptation in natural populations, and suggest that kokanee may be an ideal system for future research aimed at identifying the early genomic changes that give rise to population divergence. This study also provides data of significant utility to the conservation and management of kokanee stocks, providing a powerful tool for in season monitoring of relative ecotype abundance.

Chapter 2: Latitudinal cline in allele length provides evidence for selection in a circadian rhythm gene

2.1 BACKGROUND

Environmental heterogeneity can result in localized differences in selection pressure across the range of a species. When gene flow is minimized, these differences may drive the evolution of locally adapted populations in which phenotypic variation is fine-tuned to the environment (Taylor 1991; Garcia de Leaniz *et al*. 2007; Fraser *et al*. 2011). Several methods have been described for detecting the presence of adaptation in wild populations (Endler 1986; Fraser *et al*. 2011). The most powerful evidence comes from transplant experiments where fitness differences can be directly observed when individuals are placed outside of their native local habitat (Kawecki & Ebert 2004; Barrett & Hoekstra 2011). For many species, however, the use of experimental transplants may be unfeasible. An alternative approach is to infer adaptation by demonstrating a correlation between environmental factors and either phenotypic or genotypic variation. In molecular studies this can be achieved by identifying non-random patterns of variation at genes that are candidates for divergent selection (Garcia de Leaniz *et al*. 2007; Fraser *et al*. 2011).

Genes associated with circadian rhythm are promising candidates for inferring the genetic basis of adaptation in natural populations. The endogenous circadian clock, which is coordinated by the day-night cycle, is nearly ubiquitous among eukaryotes, and has been linked to the timing of many aspects of development including sexual maturation, reproduction, migration, and hibernation (Bradshaw & Holzapfel 2007). Recent studies using latitude as a surrogate for photoperiod have inferred the presence of divergent selection

acting on a wide range of phenotypes under circadian regulation (reviewed by Hut *et al*. in press), suggesting that the affected phenotypes are locally adapted to specific daylight regimes.

The genetic architecture underlying circadian rhythms has been characterized across a diversity of taxa (Harmer *et al*. 2001; Kyriacou *et al*. 2008). In vertebrates, the circadian gene (*clock*) codes for a protein (CLOCK) that heterodimerizes with a second protein (BMAL1) to produce a transcription-activating complex. Length polymorphisms resulting from variation in the number of polyglutamine (polyQ) repeats within the *clock* gene have a subsequent effect on the transcription of *clock*-mediated gene products, and may alter the corresponding phenotype (Caprioli *et al*. 2012).

Among anadromous salmonids, major developmental changes occur both during the juvenile migration from freshwater to the ocean and during the subsequent migration of mature adults from the ocean to their natal freshwater spawning grounds (Groot & Margolis 1991; Pearcy 1992). These life-history events involve highly synchronized behavioral, physiological, and morphological changes, suggesting that aspects of their development may be under circadian regulation (Groot & Margolis 1991; O'Malley *et al*. 2010b). Previous research has quantified genetic variation in the salmonid *clock* gene, *OtsClock1b*, from four anadromous Pacific salmonid species, sampled along latitudinal transects (O'Malley & Banks 2008; O'Malley *et al*. 2010a). These studies identified a significant correlation between the latitude of spawning grounds and the length of alleles present at the *OtsClock1b* locus in two of the four species tested, providing evidence that patterns of genetic diversity at this locus may be driven by divergent selection.

Freshwater relatives of anadromous species may provide additional insights into the evolutionary relationship between *clock* variation and photoperiod. While anadromous salmonids are born in freshwater, they grow and mature at sea; in such cases, individuals from the same natal spawning grounds mature at a variety of latitudes in the open ocean (Royce *et al*. 1968). Conversely, non-anadromous salmonids inhabit fixed latitudes (within freshwater lakes) throughout their entire life cycle. This resident lifestyle should favor adaptation across all stages of growth and maturity, potentially resulting in increased divergence at genes involved in circadian regulation.

Kokanee are resident freshwater populations of sockeye salmon (*Oncorhynchus nerka*) that inhabit lakes throughout western North America and northeast Asia (Groot & Margolis 1991). Kokanee populations are polyphyletic, having evolved from anadromous sockeye though multiple independent freshwater-colonization events following the last glaciation (Schluter 1996b; Taylor *et al*. 1996). Here, we quantified genetic diversity at the *OtsClock1b* locus from kokanee populations sampled across a latitudinal gradient in North America with paired longitudinal sites in British Columbia and across the Pacific Ocean (Kamchatka, Russia). We identified a strong correlation between latitude and *OtsClock1b* length-variation, providing preliminary evidence for adaptation to divergent photoperiod regimes.

2.2 METHODS

2.2.1 Tissue sampling and molecular methods

DNA samples collected from spawning adult kokanee were obtained from a series of 12 lakes along a latitudinal gradient ranging from southern British Columbia (49˚N) to Alaska

(62˚N), as well as an additional lake from the Kamchatka Peninsula, Russia (55˚N) (Figure 2.1; Table 2.1). All samples were originally collected as part of previous research, and detailed sampling methods and DNA extraction protocols can be found in the original publications (Taylor *et al*. 1996; Lemay & Russello 2012; Russello *et al*. 2012). Samples were genotyped using *OtsClock1b* primers described by O'Malley et al. (2007), but modified to facilitate automated genotyping (Brownstein *et al*. 1996; Schuelke 2000). Polymerase chain reaction (PCR) contained 1μl of DNA template, 1.25μl of 10 X PCR buffer, 1.25μl of 2mM dNTP mix, 0.5μl of 1μM forward primer, 0.5μl of 10μM reverse primer, 0.5μl of 10μM M13 fluorescently labeled primer, and 0.5 units of *Taq* polymerase (KAPA Biosystems) in a total volume of 13.5μl. A touchdown PCR was carried out with initial denaturation of 94˚C for 2 minutes, followed by 10 cycles at 94˚C for 30 seconds, 60˚C for 30 seconds, 72˚C for 30 seconds, with the annealing step decreasing by 1˚C per cycle to 50˚C. The annealing temperature was then maintained at 50˚C for an additional 30 cycles, followed by a final extension at 72˚C for 2 minutes. Fragment length analysis was carried out using an Applied Biosystems 3130XL automated sequencer and analyzed using GeneMapper 4.0.

2.2.2 Data analyses

In order to calibrate our results with previous studies, we Sanger-sequenced one homozygous individual for each *OtsCock1b* length variant. Translated amino acid sequences were then aligned to the eight Chinook salmon *OtsClock1b* alleles reported by O'Malley and Banks (2008) using a Clustal W protein alignment implemented in Geneious 6.1 (Biomatters Ltd.) with the default parameters.

To generate a baseline measurement of neutral variation as a function of latitude, we genotyped all individuals for four neutral microsatellite loci [*EV377149* (Koop *et al*. 2008), *One8* (Scribner *et al*. 1996), *One112* (Olsen *et al*. 1996), *Ots14* (Wright *et al*. 2008)] that have been previously used in population genetic studies of kokanee (Lemay & Russello 2012; Russello *et al*. 2012). Genotyping for these loci was carried out using the same methods described for the *OtsClock1b* gene. All loci were screened for null alleles and large allelic drop-out using the software program, Micro-Checker (Van Oosterhout *et al*. 2004). Deviation from Hardy-Weinberg equilibrium (HWE) and the presence of linkage disequilibrium (LD) were assessed for all pairs of loci in each population using the software program, Genepop 4.2 (Raymond & Rousset 1995), and corrected for multiple comparisons using the sequential Bonferroni (Rice 1989).

We tested for significant associations between latitude and *OtsClock1b* allele frequency as well as latitude and mean allele length (MAL) using a linear regression analysis implemented in R (R Development Core Team 2011); MAL was defined as the average size of alleles (in base pairs) present at each population. Following the methods of O'Malley and Banks (2008), we then tested whether the observed association between *OtsClock1b* allele frequency and latitude differed from a null hypothesis of neutrality by comparing the regression coefficients (slope of the regression) and R-squared values for each *OtsClock1b* allele (n = 2 alleles) against a neutral distribution of regression values from each of 75 microsatellite alleles (obtained from four loci) using box-plots.

To further test for differences in the distribution of genetic variation at *OtsClock1b* relative to neutral expectations, we calculated pair wise fixation index (F_{ST}) values for each population-pair using (a) only *OtsClock1b* genotypes, and (b) using combined data from the

four neutral microsatellite loci. Pair wise F_{ST} was estimated using the method of Weir and Cockerham (1984) as implemented in FSTAT 2.9.3 (Goudet 1995). We also carried out an analysis of molecular variance (AMOVA) separately for *OtsClock1b* and the neutral dataset. The AMOVA was implemented in Arlequin 3.5 (Excoffier & Lischer 2010) with 1000 permutations. To carry out this analysis we introduced hierarchical groupings in the data, where 'site' corresponded to each lake (n = 12), and 'group' was based on latitude as follows: High latitude (sites > 60˚N; n = 2); mid latitude (sites 52˚N – 55˚N; n = 4); and low latitude (sites < 51˚N; n = 6). If the distribution of genetic variation at *OtsClock1b* is the result of divergent natural selection among latitudes, then we expect low levels of *OtsClock1b* genetic differentiation among populations sampled at the same latitude groupings and high differentiation across latitude groupings; further, we would expect this distribution to occur independent of the patterns observed at neutral loci.

2.2.3 Ecotype differentiation

Within kokanee salmon, two divergent reproductive ecotypes (stream- and shore-spawning) have been identified, which vary in a number of life history characteristics including spawning timing (Taylor *et al*. 1996), with the shore ecotype spawning 2-4 weeks after the stream ecotype (Taylor *et al*. 1997). We tested for directional patterns of *OtsClock1b* gene divergence within ecotypes in the two populations for which we had samples of both stream- and shore-spawners [Okanagan Lake: n = 22 (shore), n = 24 (stream); and Kootenay Lake West: n = 27 (shore), n = 21 (stream)]. Both populations exist at 49˚ N, making them acceptable replicates for investigating patterns of ecotype divergence. Using population and ecotype as factors, MAL was compared using an ANOVA implemented in R (R Development Core Team 2011).

2.3 RESULTS

2.3.1 *OtsClock1b* diversity

We identified two alleles at the *OtsClock1b* locus with fragment lengths of 370 and 391 bases, respectively, corresponding to the presence/absence of four glutamine and three proline amino acids. Alignment with eight published Chinook salmon *OtsClock1b* sequences from O'Malley and Banks (2008) allowed us to calibrate our results with previous studies of Pacific salmonid allelic diversity (Figure 2.2). In their study of *OtsClock1b* length-variation in four anadromous Pacific salmonids, O'Malley et al. (2010a) also standardized their allele sizes to the same eight Chinook salmon alleles, which allowed us to compare the allele length (but not sequence variation) of our kokanee alleles with all four Pacific salmonids for which length-variation has previously been measured. This comparison suggests that allele 370 from our present study is equal in length to allele 353 which was previously identified in a single population of coho salmon (see Table 2 from O'Malley et al. (2010a)). Further, this comparison suggests that allele 391 from our study has not been previously identified in Pacific salmonids.

Across all loci, there were two within-population instances of homozygote excess suggesting the presence of null alleles (*EV377149* from OK and *One8* from QUE), one instance of deviation from HWE (*EV377149* from OK), and one instance of LD (*One112* and *Ots14* from OK). Given that these loci adhered to linkage equilibrium and HWE, and lacked evidence for null alleles in all other sampled populations and in previous studies (Lemay & Russello 2012; Russello *et al*. 2012), they were retained for further analyses.

2.3.2 Latitudinal clines

Linear regression analysis identified a significant correlation between the frequency of each *OtsClock1b* allele and latitude (each allele: slope ± SE = 0.038 ± 0.005, $R^2 = 0.86$ $p = 1.25 \times 10^{-5}$; Figure 2.3). Given that there were only two alleles present, MAL is directly correlated with, and will not differ from, the results of allele frequency. However, we report these values in the interest of fitting our MAL data in the context of previous work (O'Malley & Banks 2008; O'Malley *et al*. 2010a). Within that context, the linear regression revealed a significant negative correlation of *OtsClock1b* MAL with latitude (MAL: Slope ± SE = -0.8 ± 0.1, $R^2 = 0.86$, $p = 1.25 \times 10^{-5}$).

To test whether the association between latitude and *OtsClock1b* allele frequency deviated from a null hypothesis of neutrality, we compared the regression fit of each *OtsClock1b* allele against a neutral distribution of 75 microsatellite alleles. None of the microsatellite alleles (n = 75; maximum slope = 0.036, maximum $R^2 = 0.5$) exceeded the regression values (slope and R-square) observed at the circadian regulation locus, *OtsClock1b* (Figure 2.4). Indeed, following Bonferroni correction, only the *OtsClock1b* allele frequencies were significantly correlated with latitude.

Results from the AMOVA indicated that for *OtsClock1b* there is reduced genetic differentiation among populations sampled at the same latitude and increased differentiation across latitudes (Table 2.2). Conversely, among the neutral loci, substantially more genetic variation was distributed among populations within latitude than among latitudes (Table 2.2). Pairwise comparisons of F_{ST} mirror the results of the AMOVA (Table 2.3).

2.3.3 Ecotype differentiation

No statistically significant differences in MAL were detected between population ($p = 0.78$), ecotype ($p = 0.72$), or population x ecotype interactions ($p = 0.18$) in Okanagan Lake and Kootenay Lake [mean allele length ± SD: OK stream = 379.6 ± 10.6; OK shore = 378.1 ± 10.3; KLW stream = 377.9 ± 10.3; KLW shore = 380.5 ± 10.6].

2.4 DISCUSSION

Pacific salmonids provide an ideal system for investigating adaptation across a range of spatial scales; however, the vast majority of studies have focused on anadromous forms (Taylor 1991; Garcia de Leaniz *et al*. 2007; Fraser *et al*. 2011). Here, we observed that genetic variation at the circadian *OtsClock1b* locus displays a predictable distribution with respect to latitude in freshwater resident kokanee. Specifically, the shorter *OtsClock1b* allele (allele 370) increases in frequency within lakes from 50% in southern British Columbia to complete fixation in the Alaskan population. No such pattern was observed among the neutral microsatellite markers used in this study. In addition, the kokanee population from Russia fits within the North American latitudinal cline, suggesting the possibility that this pattern may be convergent across large longitudinal spatial scales. Similarly, the results from the AMOVA suggest that *OtsClock1b* variation is similar within latitudes and divergent among latitudes, a pattern that was not observed using the neutral microsatellite loci. These results provide further evidence that variation at this locus is shaped by selection rather than demographic processes.

A significant correlation between *OtsClock1b* allele-length and latitude has previously been observed in two anadromous Pacific salmonids (chum salmon, *O. keta*, and

Chinook salmon, *O. tshawytscha*) (O'Malley *et al*. 2010a). However, in these two species, the direction of the correlation was variable. Our results with kokanee are similar in magnitude and direction to the cline previously observed in chum salmon (O'Malley *et al*. 2010a); we hypothesize that this similarity may have a phylogenetic basis, given that kokanee have a closer evolutionary relationship to chum than to Chinook salmon (Murata *et al*. 1993; Kitano *et al*. 1997). Quantification of *OtsClock1b* length variation in anadromous sockeye salmon would provide additional insights in to the historical context of these observed patterns.

Correlation between genotypes and environmental variables provide evidence for adaptation (Fraser *et al*. 2011). However, given the broad regulatory activity of the vertebrate *clock* gene, it is unknown what phenotype may be under selection in this system. O'Malley et al. (2010a) speculated that clinal variation at the *OtsClock1b* locus may be associated with latitudinal variation in the timing of reproductive behavior, however subsequent statistical analysis did not support this hypothesis (O'Brien *et al*. 2010). Yet, several other behavioral phenotypes related to maturation, migration, and spawn timing have been shown to correlate with photoperiod (Pearcy 1992; Beacham *et al*. 1994; Clarke *et al*. 1994; Taylor *et al*. 2005), and may be regulated by *clock* genes (Leder *et al*. 2006; O'Malley *et al*. 2010b). For example, it has been shown that the *OtsClock1b* localizes to a quantitative trait locus (QTL) for growth rate and body length in coho salmon (O'Malley et al. 2010b). Interestingly, the QTL effects were developmentally stage-specific, with the *clock* gene co-locating with the QTL for growth rate and length at 479 days post-hatching. This timing corresponds to the period directly following smoltification when coho salmon are out-migrating to the ocean

(O'Malley *et al*. 2010b). Further research is needed in order to test whether there is a relationship between juvenile migration and *OtsClock1b* length variation.

In kokanee, it has been shown that body length is strongly correlated with fecundity, and that average body length and fecundity both decrease at increasing latitudes (McGurk 2000). This pattern is hypothesized to reflect a latitudinal gradient in the duration of the growing season, and may also explain why kokanee populations are relatively rare at higher latitudes (McGurk 2000). If population-specific patterns of growth and fecundity are related to photoperiod, then they may be candidates for phenotypes affected by the kokanee *OtsClock1b* locus. However, clines in body size may also be the result of differences in growth opportunity from environmental factors (i.e. phenotypic plasticity). Given that many phenotypes are likely to show latitudinal clines, it is hazardous to infer causation based on the common correlation of traits with geographic factors (O'Brien *et al*. 2010). Indeed, elucidating the phenotypic basis of the latitudinal cline in *OtsClock1b* genetic variation would require controlled experimental approaches in order to disentangle correlation from causation.

Our comparison of *OtsClock1b* length variation between reproductive ecotypes of kokanee in Okanagan and Kootenay Lake found no significant differences in MAL in either lake. Indeed, differences in *OtsClock1b* length variation between ecotypes within the same lake would have been counter to our hypothesis that variation at this gene is shaped by latitude (as a surrogate for photoperiod). We speculate that the lack of difference in *OtsClock1b* diversity between ecotypes suggests that the either the phenotype under selection is not involved in reproductive-timing or alternatively that the recent evolutionary divergence

of reproductive ecotypes of kokanee (<12,000 years) may not have allowed for sufficient time for the segregation of genetic diversity between ecotypes.

In addition to salmonids, there have been several recent studies testing for latitudinal clines in *clock* length-variation in birds (for example: Fidler *et al*. 2006; Johnsen *et al*. 2007; Dor *et al*. 2011; Caprioli *et al*. 2012). Significant latitudinal clines in *clock* length-variation have been observed in the migratory barn swallow (Caprioli *et al*. 2012) and in the non-migratory blue tit (Johnsen *et al*. 2007). Among species where a significant cline has been observed, there is correlative evidence suggesting that these patterns are associated with breeding phenology (Liedvogel *et al*. 2009; Caprioli *et al*. 2012). Yet, as with Pacific salmonids, the lack of generalities across taxa highlights the need for research aimed at understanding the functional significance of these latitudinal clines.

While it is premature to conclude that populations are locally adapted before a direct causal relationship between genotype, phenotype, and fitness has been characterized (Gould & Lewontin 1979; Barrett & Hoekstra 2011), the identification of a correlation between gene frequency and an environmental variable provides indirect evidence for natural selection, warranting future study (Endler 1986). We provide evidence for non-random distribution of allele-length with respect to latitude at the *OtsClock1b* locus in kokanee salmon, suggesting adaptation to divergent photoperiod regimes. Future research is necessary to identify the trait(s) that is putatively under selection.

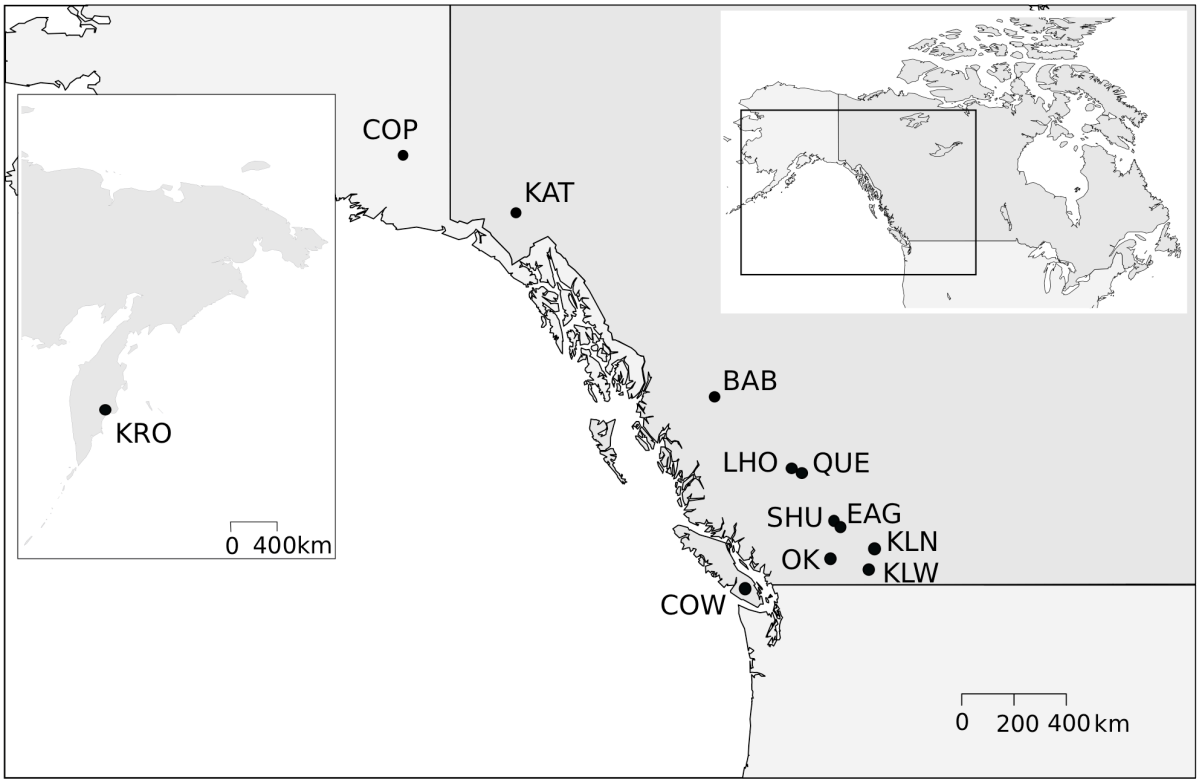

Figure 2.1: Sampling localities used to quantify genetic variation at the *OtsClock1b* locus in kokanee salmon from North America and Russia (inset). Additional location information is provided in Table 1.

```
                    1         10        20        30        40        50
Kokanee_1           HQQQQQRQQQPQPQRQQQQPPPQQQQ---PPPQQQQPP-PQQQQPPP-------P
Kokanee_2           HQQQQQRQQQPQPQRQQQQPPPQQQQ---PPPQQQQPP-PQQQQPPPQQQQPPPP
Chinook_allele383   HQQQQQRQQQPQPQRQQQQP---------PPPQQQQPPPPQQQQPPP-------P
Chinook_allele362   HQQQQQRQQQPQPQRQQQQPQPQRQQQQPPPPQQQQPPPPQQQQPPP-------P
Chinook_allele359   HQQQQQRQQQPQPQRQQQQPPPPQQQQ-PPPPQQQQPPPPQQQQPPP-------P
Chinook_allele356   HQQQQQRQQQPQPQRQQQQPPPPQQQQ--PPPQQQQPPPPQQQQPPP-------P
Chinook_allele338   HQQQQQRQQQPQPQRQQQQP---------PPPQQQQPPPPQQQQPPP-------P
Chinook_allele335   HQQQQQRQQQPQPQRQQQQP---------PPPQQQQPPPPQQQQPPP-------P
Chinook_allele329   HQQQQQRQQQPQPQRQQQQP---------PPPQQQQPPPPQQQQPP---------
Chinook_allele293   HQQQQQRQQQPQPQRQQQQPPP------------------PQQQQPPP-------P
                          60        70        80        90       100  103
Kokanee_1           QQQQPPPQQQ----------------QQPPPEQQPPQQQQLQIQPSVQ
Kokanee_2           QQQQPPPQQQ----------------QQPPPEQQPPQQQQLQIQPSVQ
Chinook_allele383   QQQQPPPPQQQQPPPPQQQQPPPQQQQQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele362   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele359   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele356   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele338   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele335   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele329   QQQQPPPQQQ----------------QQPPQEQQPP-QQQLQIQPSVQ
Chinook_allele293   QQQQPPPPQQ----------------QQPP-------QQQLQIQPSVQ
```

Figure 2.2: Multiple sequence alignment of the two *OtsClock1b* alleles identified in kokanee salmon along with the eight alleles previously published by O'Malley and Banks (2008). Kokanee_1 is allele 370 from the current study; calibrated to O'Malley et al. (2010) it would be allele 353. Kokanee_2 is allele 391 from the current study; calibrated to O'Malley et al. (2010) it would be allele 374.

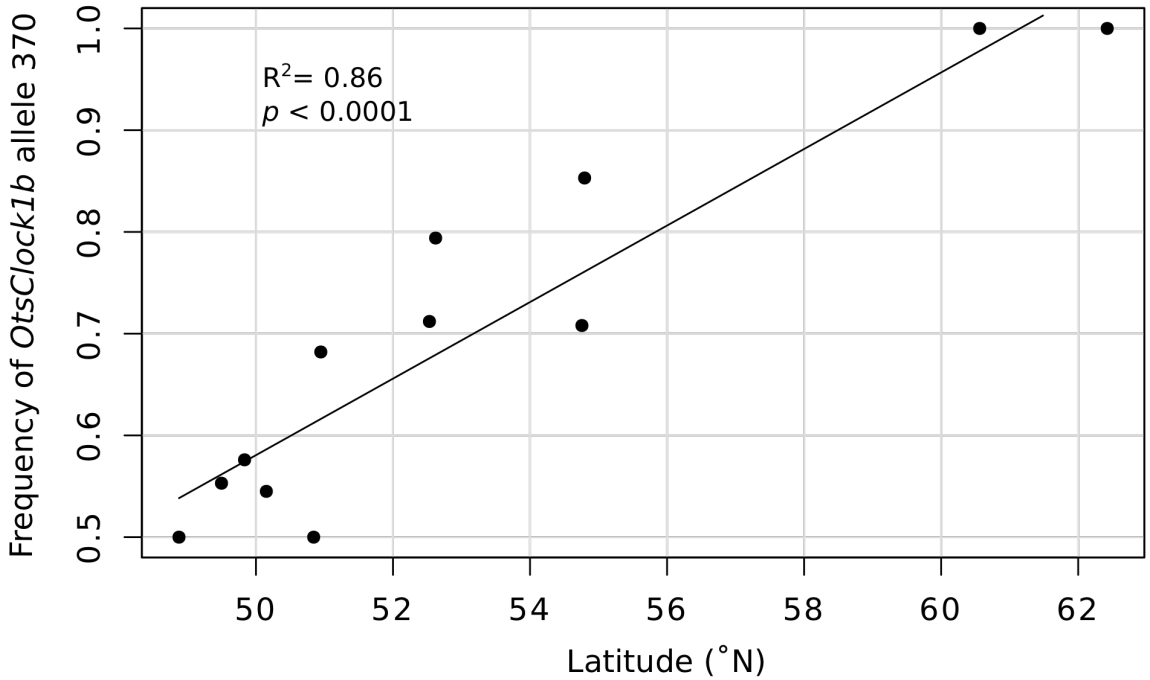

Figure 2.3: Frequency of the most common OtsClock1b allele (allele 370) for each sampled population of kokanee salmon as a function of latitude.

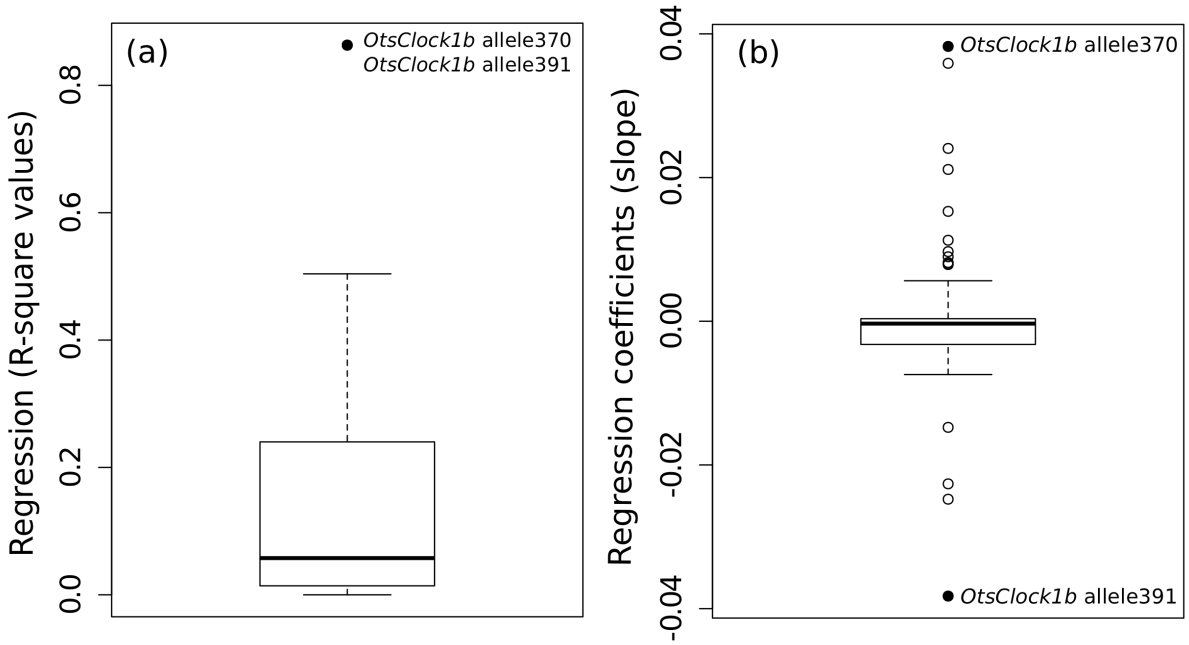

Figure 2.4: Graphical representation of the distribution of (a) R-squared values, and (b) regression coefficients (slopes) for the relationship between allele frequency and latitude at each allele sampled. There are two alleles from the candidate locus, *OtsClock1b* (indicated as shaded circles), and 75 alleles from neutral microsatellite loci.

Table 2.1: Kokanee samples from British Columbia (BC), Alaska (AK), and Russia. Length variation is reported as the frequency of allele 370 and mean allele length (MAL).

Population	Location ID	Latitude	Longitude	Sample size	Allele 370	MAL
Cowichan Lake, BC	COW[1]	48.877	-124.266	11	0.500	380.5
Kootenay Lake, BC West arm	KLW[2]	49.497	-117.321	48	0.553	379.4
Okanagan Lake, BC	OK[3]	49.834	-119.524	46	0.576	378.9
Kootenay Lake, BC North arm	KLN[2]	50.152	-116.933	23	0.545	379.6
Eagle River, BC	EAG[1]	50.843	-119.012	12	0.500	380.5
Shuswap Lake, BC	SHU[1]	50.947	-119.259	11	0.682	376.7
Quesnel Lake, BC	QUE[1]	52.532	-121.114	26	0.712	376.1
Lower Horsefly River, BC	LHO[1]	52.622	-121.612	17	0.794	374.3
Kronotsky Lake, Russia	KRO[1]	54.757	160.284	12	0.708	376.1
Babine Lake, BC	BAB[1]	54.797	-126.027	17	0.853	373.1
Kathleen Lake, BC	KAT[1]	60.562	-137.331	11	1.000	370.0
Copper Lake, AK	COP[4]	62.424	-143.567	24	1.000	370.0

[1] Donated by E.B. Taylor (UBC), described in Taylor et al. (1996).
[2] Sampling described in Lemay and Russello (2012).
[3] Sampling described in Russello et al. (2012).
[4] Donated by J.E. Seeb (University of Washington)

Table 2.2: Analysis of molecular variance (AMOVA) results showing the percentage of variation at each hierarchical level of organization for *OtsClock1b* and the neutral microsatellites. For this analysis, Group = latitude (categorized as high, mid, low; see main text for more details related to grouping strategy), and population = lake.

	Percentage of variation		
	Among groups (latitude)	Among populations within groups	Within populations
OtsClock1b	17.88*	-0.84	82.96*
Neutral loci	4.81*	8.73*	86.46*

* $p < 0.05$

Table 2.3: Pairwise *FST* for each population at (A) four neutral loci and (B) *OtsClock1b*. For each dataset, *p*-values are above the diagonal and *FST* -values are below the diagonal. *FST*-values in bold are significant at $p<0.05$ (uncorrected for multiple tests).

		COP	KAT	BAB	KRO	LHO	QUE	SHU	EAG	KLN	OK	KLW	COW
A.	COP		0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076
NEUTRAL	KAT	**0.3488**		0.00152	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076
LOCI	BAB	**0.1965**	**0.1487**		0.02879	0.00152	0.00379	0.20076	0.82273	0.00985	0.00076	0.00076	0.03485
(N=4)	KRO	**0.2569**	**0.2051**	**0.0624**		0.00076	0.00303	0.02045	0.075	0.00076	0.00076	0.00076	0.00076
	LHO	**0.1559**	**0.1646**	**0.0281**	**0.0339**		0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076
	QUE	**0.2303**	**0.1551**	**0.0237**	**0.0508**	**0.0234**		0.00076	0.03258	0.00076	0.00076	0.00076	0.00076
	SHU	**0.2158**	**0.1674**	0.0075	**0.0717**	**0.0457**	**0.0184**		0.39015	0.00227	0.00303	0.00076	0.00152
	EAG	**0.2127**	**0.0931**	0.0018	0.0562	**0.0355**	**0.0108**	-0.0058		0.01364	0.00076	0.00076	0.08636
	KLN	**0.1872**	**0.2148**	**0.0139**	**0.0989**	**0.0468**	**0.0391**	**0.0265**	**0.0428**		0.00076	0.00076	0.0053
	OK	**0.2087**	**0.2294**	**0.0404**	**0.0747**	**0.0499**	**0.0506**	**0.0288**	**0.059**	**0.0378**		0.00076	0.00076
	KLW	**0.3295**	**0.3123**	**0.0974**	**0.2006**	**0.1428**	**0.0902**	**0.1114**	**0.1336**	**0.076**	**0.1029**		0.00076
	COW	**0.262**	**0.1742**	**0.0193**	**0.0907**	**0.0561**	**0.0411**	**0.0354**	0.0264	0.0311	**0.0702**	**0.0894**	

		COP	KAT	BAB	KRO	LHO	QUE	SHU	EAG	KLN	OK	KLW	COW
B.		COP	KAT	BAB	KRO	LHO	QUE	SHU	EAG	KLN	OK	KLW	COW
CLOCK	COP		NA	0.00985	0.00303	0.00152	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076	0.00076
LOCUS	KAT	0		0.05152	0.0447	0.03788	0.00152	0.00227	0.00076	0.00076	0.00076	0.00076	0.00227
	BAB	**0.1541**	0.0952		0.24167	0.74697	0.1553	0.10455	0.00606	0.00455	0.00909	0.00152	0.00985
	KRO	**0.3544**	**0.2353**	0.0212		0.60076	1	1	0.27348	0.22348	0.30152	0.27727	0.30985
	LHO	**0.2162**	**0.1396**	-0.0185	-0.0265		0.46667	0.50606	0.02955	0.03182	0.02955	0.01439	0.06364
	QUE	**0.2687**	**0.2038**	0.0353	-0.0315	-0.0046		1	0.07879	0.08258	0.13409	0.07803	0.11439
	SHU	**0.4159**	**0.3**	0.0584	-0.0447	0.0004	-0.022		0.17424	0.22348	0.50606	0.32424	0.35076
	EAG	**0.5918**	**0.4731**	**0.2398**	0.0455	**0.151**	0.0712	0.0389		0.76364	0.76364	0.65985	0.66818
	KLN	**0.4569**	**0.3693**	**0.179**	0.0238	**0.1075**	0.0424	0.0146	-0.0197		0.85682	1	0.79091
	OK	**0.3473**	**0.2926**	**0.1341**	0.0043	**0.0754**	0.023	-0.007	-0.0168	-0.0156		0.77121	0.65076
	KLW	**0.3687**	**0.3147**	**0.1561**	0.0195	**0.0952**	0.0372	0.0061	-0.0207	-0.0162	-0.011		0.82955
	COW	**0.6005**	**0.4727**	**0.2345**	0.0333	0.1437	0.0657	0.0273	-0.0415	-0.0264	-0.0215	-0.0252	

Chapter 3: Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon

3.1 BACKGROUND

Biology is currently undergoing a genomics revolution (Wolfe & Li 2003). With the increasing ubiquity of high-throughput sequencing technology, it is now possible to rapidly generate millions of base pairs of DNA sequence data for a fraction of the per-base cost required for chain-termination Sanger sequencing. The result is an exponential increase in the abundance of genomic resources available for studying non-model organisms (Allendorf *et al*. 2010). Single nucleotide polymorphisms (SNPs) have emerged as a marker of choice for population-level studies in the genomics era (Brumfield *et al*. 2003; Garvin *et al*. 2010; Helyar *et al*. 2011). Due to their broad genomic distribution, direct association with functional significance, and ease of genotyping, SNPs represent an improvement over conventional markers such as amplified length polymorphisms (AFLPs) and expressed sequence tag (EST)-linked microsatellites. The efficiency of SNP discovery and validation may be further increased through targeting of the transcriptome [RNA-seq;(Wang *et al*. 2009)] or restriction-site associated DNA tags [RAD-tags; (Baird *et al*. 2008; Peterson *et al*. 2012)].

The emerging field of population genomics has evolved in tandem with these novel technologies, incorporating genome-wide sequence data into the study of systems that historically would have been limited to a small number of neutral markers (Luikart *et al*. 2003). Central to a population genomic approach is the identification of statistical outlier loci that exhibit locus-specific patterns that are highly divergent from the rest of the genome,

representing candidate regions under selection (Cavalli-Sforza 1966; Lewontin & Krakauer 1973; Narum & Hess 2011). The combined signal from high density neutral and putatively adaptive SNPs throughout the genome offers great potential for investigating evolutionary and ecological questions in natural populations. For example, several recent studies have found that the use of outlier loci can reveal fine-scale population structure beyond what was previously inferred from conventional neutral markers (e.g. Atlantic herring (André *et al*. 2011), Atlantic cod (Westgaard & Fevolden 2007), Atlantic salmon (Freamo *et al*. 2011)). Moreover, population genomic approaches that incorporate statistical outlier loci offer great potential for delineating conservation units (Bonin *et al*. 2007; Funk *et al*. 2012) and informing fisheries management (Russello *et al*. 2012).

Kokanee salmon are recently diverged land-locked populations of sockeye salmon (*Oncorhynchus nerka*) that exist in lakes throughout western North America and northeast Asia. Two reproductive ecotypes of kokanee have been described based on spawning behaviour and location. Stream-spawning kokanee migrate into tributaries in early autumn, display pronounced secondary sex characteristics, site defence, and redd formation. Conversely, shore-spawning populations (also known as beach-spawners) breed later in the autumn, lack secondary sex characteristics and site defence, and mate directly on the shoreline rather than migrating into tributaries (Dill 1996; Shepherd 2000). In many lakes, the two ecotypes occur in sympatry. Although kokanee ecotypes display different reproductive traits and experience varying incubation environments, outside of the spawning season the ecotypes co-occur within lakes and morphologically indistinguishable.

Throughout their range, kokanee constitute a heavily managed recreational fishery of great socioeconomic and ecological importance (Thompson 1999). The ability to distinguish

stream- and shore-spawning kokanee is an important goal for fisheries managers, as ecotypes may be differentially impacted by alternative conservation and management regimes with regard to water use, the protection of spawning habitat, and recreational harvest regulations. While visual surveys of kokanee during the spawning season are quite accurate in streams, it is considerably more difficult to obtain visual measurement of shore-spawner abundance, especially in lakes where shore populations spawn at greater depth or at night. Kokanee management would be significantly enhanced by the identification of molecular markers capable of assigning individuals to the correct ecotype, however the identification of sufficiently fine-scale markers has been hindered by the recent divergence and correspondingly weak genetic structure of the two ecotypes (Russello *et al*. 2012).

Previous studies using neutral loci have found evidence for low levels of genetic differentiation based on mitochondrial DNA haplotype and microsatellite loci, yet individual assignment probabilities to ecotype were generally low (< 80%; Taylor *et al*. 1997; Taylor *et al*. 2000; Lemay & Russello 2012). More recently, work using a panel of EST-linked microsatellite loci has identified a set of markers with a high association of genetic diversity with reproductive ecotype (>90% in individual assignment tests), suggesting that there may indeed be a genetic basis for the evolution of kokanee ecotypes (Russello *et al*. 2012). However these markers suffer from poor coverage across the genome and provide little information on the specific genes involved in the divergence of the two ecotypes.

Here, we evaluated an RNA-seq approach for preferentially identifying SNPs of putatively adaptive significance within and among kokanee ecotypes. We used the Roche 454 GS FLX Titanium platform to generate transcriptome-wide sequence data for pooled cDNA libraries of shore- and stream-spawning kokanee. These data were used for SNP

discovery and subsequent validation in natural populations, enabling explicit investigations of ecotype-specific patterns of genetic variation.

3.2 METHODS

3.2.1 Sample collection, RNA extraction, and next-generation sequencing

A total of eight individuals per ecotype were collected during spawning from Okanagan Lake, British Columbia (see Table 3.1). Each individual was immediately sacrificed upon collection and dissected in the field; five different tissues (heart, liver, muscle, gonad, and olfactory bulb) were harvested and immersed in separate 5ml screw-cap vials containing 2.5ml of RNALATER® (Life Technologies) solution. Samples were held at 4°C for 24hrs and then stored at -20°C until needed. RNA was extracted from each tissue type using the RNEASY UNIVERSAL MINIKIT (Qiagen) following the manufacturer's protocol.

Two normalized cDNA libraries (Evrogen, Russia) were constructed using pooled RNA from all shore-spawners (5 tissues x 8 individuals) and all stream-spawners (5 tissues x 8 individuals). The two resulting cDNA libraries were each subject to a full run of 454 GS FLX Titanium sequencing at the Genome Quebec core facility. Pooling of multiple individuals in each sample was used to provide a preliminary indication of sequence variation within and among ecotypes; the combination of five tissue types for each individual was used to maximize the diversity of expressed genes present in each library. RNA samples were normalized to increase the likelihood that rare transcripts were detected in the sequence data. The raw read data from each library was deposited at the NCBI Sequence Read Archive (project # SRP021088) with the following accessions: Stream reads SRR827512, SRR827513; Shore reads SRR827572, SRR827573).

3.2.2 Transcriptome assembly

The CLC GENOMICS WORKBENCH (CLC Bio) v.4.8 was used for initial quality filtering. A *de novo* assembly was then carried out using CLC GENOMICS WORKBENCH v.4.8 (similarity = 0.96, length fraction 0.5) in order to generate reference contigs for subsequent analyses. A conservatively high similarity value was used in our assembly in order to minimize the incorporation of paralogous sequence variants (PSVs) within these reference contigs (Renaut *et al*. 2010; Seeb *et al*. 2011).

To facilitate a comparison of sequence variation between the two ecotypes, the consensus sequence from each contiguous DNA sequence (contig) created during the *de novo* assembly was used as a reference to map the stream and shore reads separately (CLC GENOMICS WORKBENCH v.5.5; similarity = 0.96, length fraction 0.5). We retained only those contigs with a minimum length of 200 bases and an average depth of coverage greater than 5 reads per nucleotide (5x) for each ecotype (hereafter referred to as the high coverage data set).

We also generated data sets containing those contigs that were composed of reads from only a single ecotype (same parameters as above, minimum coverage = 5x; minimum length = 200 bases). These two 'ecotype-unique' data sets may suggest target genes for subsequent studies examining differences in gene expression among ecotypes.

3.2.3 Mitochondrial genome

While there is not yet a fully assembled and annotated salmon nuclear genome, an annotated mitochondrial genome sequence is available for most salmonid species. To assess the prevalence of mitochondrial genes in the transcriptome data, we used the mitochondrial

genome for sockeye salmon [GENBANK: EF055889] as a reference to map the kokanee transcriptome reads (CLC GENOMICS WORKBENCH v.4.8, similarity = 0.90).

3.2.4 Transcriptome analysis and annotation

The high coverage data set was subject to sequence similarity searches using the software program, Blast2Go v.2 (Conesa *et al*. 2005; Gotz *et al*. 2008). For this analysis, a translated nucleotide (Blastx) search was performed using the NCBI non-redundant database (e-value threshold = 10^{-6}, HSP length cut-off = 33). The top five hits for each contig were retained. Following the Blastx search, gene ontology (GO) analysis was carried out in order to obtain hierarchical structure information with respect to the three GO domains (molecular function, biological process, and cellular component). This analysis was carried out using Blast2Go v. 2 (e-value threshold = 10^{-6}, HSP length cut-off = 20, GO weight = 5).

Sequence similarity searches were also carried out for all contigs in each of the ecotype unique data sets. Both Blastx (same parameters as above) and nucleotide (Blastn; maximum e-value threshold = 10^{-5}) searches were performed. Blast2Go v.2 was again used to assign GO annotations to contigs (same parameters as above).

A *de novo* assembly (CLC GENOMICS WORKBENCH v5.5; similarity = 0.90, length fraction = 0.5) was carried out using all contigs in the high coverage data set in order to identify contigs with overlapping portions. Redundancies among the contigs may indicate the presence of paralogous sequence variants (PSVs) or be indicative of alternative splicing within the transcriptome data. Contigs that overlapped with one or more other contigs were removed from the high coverage data set.

3.2.5 SNP discovery

The working data set of high coverage contigs was screened for SNPs using the CLC GENOMICS WORKBENCH v. 5.5 software package (minimum coverage 8x, minimum variant frequency 10%, minimum reads per allele = 2, minimum central quality 20). By carrying out mapping and SNP detection separately for the two ecotypes and then combining the resulting SNP tables, it was possible to determine the number and percent of reads that matched the reference contig sequence at each SNP site. The absence of a polymorphism (fixation) in one ecotype was evident as either 100% of reads matching the reference (in this case no SNP would be observed in the table for this ecotype) or 100% of reads possessing a nucleotide other than that of the reference sequence. For example, a fixed difference for a given location would be scored when one ecotype was 100% different from the reference contig sequence, while the other ecotype displayed no polymorphism at that site (i.e. 100% match to the reference). Using this approach we were able to characterize all SNPs as being either: (a) polymorphic in both ecotypes; (b) fixed in one ecotype, polymorphic in the other; or (c) fixed for different alleles in each ecotype. Variation in the form of insertions or deletions (indels) was not examined in this study.

A divergence value (based on the index implemented in Juekens et al. (Jeukens *et al*. 2010)) was calculated for each SNP, defined as the absolute value of the frequency difference of the major allele between ecotypes. For example, a SNP that was fixed (frequency = 100%) for a different allele in each ecotype would have a divergence value = 1.0; a SNP where the major-allele was fixed in shore and had a frequency of 60% in stream would have a divergence value = 0.40. The approach was used to putatively identify the most divergent polymorphic sites within the transcriptome.

3.2.6 SNP validation

A subset of 36 SNPs was used to genotype an independent sample of kokanee in order to validate this ascertainment procedure. Validation of candidate SNPs was carried out following a pipeline similar to that implemented by Seeb et al. (2011). Briefly, primers were designed using PRIMER3 (Rozen & Skaletsky 2000) such that they would amplify a ~60-200bp fragment that encompassed a single SNP. Loci that produced a single clean PCR-amplified fragment with no detectable introns (larger PCR products than expected) were used to genotype 88 individuals using HRM analysis (see below).

Each PCR reaction contained 1.25μl of 10x buffer, 1.25μl of 2 mM dNTP mix, 0.5μl of 10mM forward and reverse primer, 0.5 units of Kapa *Taq* polymerase (Kapa Biosystems), 20-100 ng of DNA template, and ultra pure water for a total reaction volume of 12.5 μl. For each reaction, a touchdown PCR procedure was implemented using a Veriti thermal cycler (Applied Biosystems). The program had an initial denaturation at 94˚C for 2 minutes, followed by 10 cycles at 94˚C for 30 seconds, 60˚C for 30 seconds, and 72˚C for 30 seconds, with the annealing temperature decreasing by 1.0˚C per cycle. This was followed by 25 cycles at 94˚C for 30 seconds, 50˚C for 30 seconds, and 72˚C for 30 seconds. The final cycle had an extension of 72˚C for 2 minutes and was then held at 4˚C. PCR products were run on a 1.5% agarose-gel in order to obtain a preliminary assessment of the quality and size of the amplicon. Loci that showed evidence for the presence of introns, or had multiple bands were not retained for subsequent analyses.

Loci that produced a single clean PCR product of the anticipated size were then evaluated by HRM analysis using DNA samples from 32 stream-spawning and 32 shore-spawning kokanee from Okanagan Lake. For detailed sampling methods for the kokanee

DNA samples see Russello et al. (Russello *et al*. 2012) . In addition, 24 anadromous sockeye salmon DNA samples were included in this analysis. These samples were collected by the Okanagan Nation Alliance from Osoyoos Lake, British Columbia, which is the closest extant population of sockeye salmon from the same drainage system as Okanagan Lake. Fin clips were removed from mature adults during the spawning season in September 2011. DNA was extracted using a NucleoSpin Tissue Kit (Macherey Nagel) following the manufacturers suggested protocol for 96 well plates.

Each HRM reaction contained 7.2μl of Precision Melt Supermix (Bio-Rad), 0.4μl of each primer, 20-100 ng of DNA template, and ultra pure water for a total reaction volume of 20μl. HRM analyses were run in 96 well plates on a Bio-Rad CFX96 Touch™ real time PCR detection system. A two-step touchdown PCR protocol was used starting with an initial denaturation step at 95˚C for 2 minutes, followed by 9 cycles of 95˚C for 10 seconds, 60˚C for 30 seconds, with the annealing temperature decreasing by 1˚C per cycle. This was followed by 43 cycles of 95˚C for 10 seconds and 50˚C for 30 seconds. The final PCR cycle consisted of 95˚C for 30 seconds followed by 55˚C for 1 minute. A plate read was obtained at the end of every PCR cycle. The melt curve data were obtained starting at 70˚C and increasing by 0.2˚C every 10 seconds to a maximum of 95˚C. During the melting stage a plate read was obtained at every 0.2˚C increment. Melt curve data were analyzed using the Bio-Rad CFX proprietary software.

SNPs that demonstrated multiple clusters of HRM curves (i.e. polymorphic loci) were subjected to Sanger re-sequencing on an ABI 3130 Genetic Analyzer to confirm their genotype. Given that each HRM cluster should represent a single SNP genotype, 1-2

individuals from each cluster were sequenced in order to determine the genotype of each cluster.

For each locus that was successfully genotyped, we calculated expected and observed heterozygosity (per ecotype), and tested for deviations from Hardy-Weinberg and linkage equilibrium. These analyses were carried out using the software program, GenePop v.4 (Raymond & Rousset 1995) followed by a sequential Bonferroni correction for multiple tests (Rice 1989). In addition, the allele frequencies observed from HRM genotyping of each locus were compared with the frequency expected from the pooled transcriptome read data. For consistency, the major-allele was defined as the allele with the highest frequency observed in the transcriptome data.

3.3 RESULTS AND DISCUSSION

3.3.1 Sequencing and assembly

We generated ~750 million base pairs of sequence data corresponding to $1.3x10^6$ and $1.4x10^6$ reads for the shore and stream-spawning cDNA libraries, respectively, with an average length of 271 bases per read (Table 3.2). A *de novo* assembly was carried out with the trimmed reads from both ecotype libraries in order to generate contigs to use as reference sequences; this assembly incorporated 87% of the transcriptome reads to produce 123,547 contigs (Table 3.3). We then mapped the raw reads back to these reference contigs separately for each ecotype and generated a refined data set consisting only of contigs that had a minimum average coverage of 5x for each ecotype and a minimum length of 200 bases. The resulting data set (hereafter referred to as the high coverage data set) was retained for subsequent analyses (Table 3.3; Figure 3.1).

3.3.2 Transcriptome analyses and annotation

A genome duplication event in the early evolution of salmonids and subsequent diploidization has resulted in a genome in which isoloci and paralogous gene copies are common (Allendorf & Thorgaard 1984). The presence of PSVs among contigs would impede our ability to identify sequence variants between the two ecotypes. To address this problem, we used conservatively high similarity parameters during the *de novo* assembly in order to minimize the incorporation of PSVs within each contig (Renaut *et al*. 2010; Seeb *et al*. 2011). In doing so, two separate but highly similar contigs may be created: one corresponding to functional coding sequence, the other to a PSV. To test for these redundancies in the data, a *de novo* assembly with a slightly reduced sequence-similarity value (0.90) was carried out on the high coverage contigs in order to identify any contig sequences that either partially or totally overlapped. Of the 13,593 contigs initially retained in the high coverage data set, 1,864 (13.7%) aligned with one other contig, and 643 (4.7%) aligned with two or more other contigs. These contigs were discarded from the high coverage data set. The remaining 11,085 contigs (81.5%) were unique and did not show similarity with any other contig.

The strict similarity value used to create the reference contigs may also explain the relatively short size of the contigs retained in the high coverage data set (average length = 595 bases). While the short contig size may limit some down-stream analyses, the minimization of PSVs in the data was of higher priority when optimizing the assembly parameters to ensure the accuracy of each inferred genotype.

A BLAST search of all contigs in the high coverage data set (n = 11,085) produced 4,410 positive matches to sequences in the NCBI database (minimum e-value cut-off = 10^{-6};

average e-value = 1.9×10^{-9}). Of those contigs with positive BLAST hits, 60% matched published salmonid sequences (contigs matching *Oncorhynchus sp*. = 628; *Salmo sp*. = 1993; all other salmonids = 12). In addition, eleven contigs were a positive match to the infectious hematopoietic necrosis virus (IHNV), a lethal virus that is enzootic in western North America and can have detrimental impacts on salmonid aquaculture (Nichol *et al*. 1995). These 11 contigs were subsequently removed from the high coverage data set for all further analyses. Of the high coverage contigs that had positive matches from the Blastx search, 2,285 were subsequently annotated with one or more gene ontology (GO) terms (Figure 3.2,).

3.3.3 Mitochondrial genome

There was coverage across all genes in the sockeye salmon mitochondrial genome [GENBANK: EF055889], representing 0.3% of the trimmed transcriptome reads (n = 6,824).

3.3.4 SNP detection

Within the high coverage data set, 8,339 contigs contained SNPs that fell within our detection parameters (minimum coverage 8x, minimum variant frequency 10%, minimum reads per allele = 2, minimum central quality 20). From these contigs, we identified 32,699 putative SNPs that may be used for population genetic analyses of kokanee. Although there has been much focus on marker development in anadromous sockeye salmon (Smith *et al*. 2005; Elfstrom *et al*. 2006; Habicht *et al*. 2010; Campbell & Narum 2011), to our knowledge, this is the first study that has identified SNPs specifically for reproductive ecotypes of kokanee.

Given that low levels of molecular divergence among ecotypes of Okanagan Lake kokanee have been observed in previous studies, we expected the majority of loci to display similar allele frequencies in both ecotypes. Of the putative SNPs identified in this study, 7,931 were polymorphic in both ecotypes, 12,835 were polymorphic in the stream ecotype but fixed in shore, and 11,933 were polymorphic in the shore ecotype but fixed in stream. There were no SNPs within our detection parameters that were fixed for alternate alleles in the two ecotypes.

In this study, the high frequency of loci that appear to be fixed in one ecotype may be artificially inflated as a result of the small sample size (only eight individuals per ecotype) in the pooled transcriptome libraries and/or due to the SNP detection parameters, which required a minimum of eight reads at a given site and at least two reads for each allele. If one ecotype had coverage below these cut-off values, then it would erroneously appear to be fixed at that site, even if there was some variation present. This could result if one ecotype under-expressed a given gene, preventing it from being detected at sufficiently high levels in the transcriptome data. These potential biases reflect the trade-off between avoiding false SNPs resulting from sequence error, while attempting to account for all possible variation in the data.

The difference in the frequency of the major allele between ecotypes ranged from <1-88% (mean = 19%). Ninety-five percent of all SNPs were contained within a divergence value of 44% or less, suggesting that the overall level of nucleotide divergence is low. The remaining 5% of loci (n = 1,493) represented the most divergent SNPs, with values between 45-88%. These highly divergent SNPs may represent promising targets for ecotype discrimination, potentially associated with genes underlying ecological diversification.

3.3.5 SNP validation

Primer pairs were designed for 36 loci such that they amplified a 60-200 base pair fragment containing a single SNP. Of these loci, 18 exhibited successful PCR amplification, were free of introns, and produced sufficiently clear high resolution melt (HRM) signal to attempt the subsequent genotyping validation. These loci were then used to genotype 32 stream-spawners, 32 shore-spawners, and 24 anadromous sockeye salmon using HRM analysis. From the panel of 18 SNPs for which broad-scale HRM genotyping and Sanger validation was attempted, nine loci produced consistently scorable HRM clusters and were retained for subsequent analyses (Table 3.4).

Sanger sequence data for the remaining loci confirmed that the expected SNP site was indeed polymorphic, however the resulting HRM clusters were not sufficiently discrete to enable accurate genotype assignment. For these loci, Sanger sequencing identified either multiple melt curves with the same genotype or multiple genotypes within the same melt curve. Some of the inconsistencies with the HRM data may be explained by the reduced accuracy of HRM compared with other conventional genotyping methods such as TaqMan® assays (Martino *et al*. 2010). Martino et al. (2010) found that rare genotypes and low minor allele frequencies decreased the accuracy of HRM analysis. Both of these factors are likely present in our data given that we preferentially chose divergent loci that were fixed for a single allele in one population. The presence of additional polymorphic sites within the amplicon (Seeb *et al*. 2011) and the use of loci containing Class 3 (C/G) or Class 4 (A/T) SNPs (Liew *et al*. 2004) may also have been factors resulting in weakly differentiated clusters. Given that Sanger sequencing confirmed the presence of the expected

polymorphism, we conclude that the failed assays were not due to errors with the initial SNP detection but rather reflect the limitations of the HRM assays at those loci.

There was no evidence of linkage disequilibrium among any of the nine loci for which genotypic data were obtained. One locus (*One74958*) showed a significant deviation from Hardy-Weinberg equilibrium (HWE) in all three populations tested (stream spawning kokanee, shore spawning kokanee and anadromous sockeye salmon). One additional locus (*One73476*) showed a significant deviation from HWE among the shore-spawning samples only (Table 3.4). Among the anadromous sockeye salmon samples used in this study, all SNPs showed similar patterns of allele frequencies to kokanee. These results indicate that SNP loci developed from the kokanee transcriptome data could also be useful for genotyping populations of anadromous sockeye salmon.

The frequency data obtained from HRM genotyping suggests that while the transcriptome data set provides a valuable tool for identifying novel SNPs, it is limited in its ability to infer population allele frequencies (Table 3.4). The disparity between the transcriptome sequences and SNP genotypes may be indicative of differences in gene expression in the transcriptome data obscuring the identity of genotypes present. Additionally, small sample size in the pooled transcriptome libraries may have contributed to the observed inconsistencies between HRM genotypes and the allele frequencies predicted from the transcriptome data. We initially hypothesized that the most divergent SNPs in the transcriptome sample could indicate high-confidence targets for genetically discriminating ecotypes. Based on these results, however, we feel that the relative frequency of alleles present in the pooled transcriptome reads is not an appropriate surrogate for large-scale SNP genotyping followed by statistical outlier tests. The use of greater sample sizes within the

pooled cDNA libraries may improve this outcome. Likewise, emerging protocols that utilize combinatorial labelling methods and RAD tags may provide more efficient and cost-effective alternatives for simultaneously discovering SNPs in non-model organisms and genotyping populations (Baird *et al*. 2008; Peterson *et al*. 2012).

3.3.6 Ecotype unique data

We also generated additional data sets with contigs that contained only reads from a single ecotype (Table 3.3). The presence/absence of ecotype-specific contigs may represent candidates for differences in gene expression (rather than sequence variation) between ecotypes. Interestingly, we observed twice as many ecotype-unique contigs from the stream-spawning DNA library, which may indicate a number of genes that are not expressed by the shore ecotype during spawning. However, this discrepancy in ecotype-specific contigs could partially be accounted for by genes with very low levels of expression that were observed in one library by chance alone. A study specifically designed to examine differences in levels of gene expression may be useful at identifying divergence between the two ecotypes.

BLAST searches (Blastn, NCBI, max e-value = 10^{-05}) of the ecotype-unique contigs produced 150 positive matches in the shore-unique data set (mean e-value = 1.1×10^{-07}) and 393 positive hits for contigs that are unique to the stream ecotype (mean e-value = 2.1×10^{-06}). Interestingly, 12 stream-unique contigs (2%, including the highest coverage contig) were a match to *Saprolegnia ferax* [GENBANK: AY534144], a pathogenic water mould that is associated with pre-spawning mortality in salmonids (Van West 2006). Contigs matching this pathogen were absent among the shore-unique contigs. Similarly, 92 stream-unique contigs (14%) were a match to *Flavobacterium psychrophilum* [GENBANK: AM398681], which is a bacterial infection associated with high levels of salmonid mortality (Duchaud *et*

al. 2007). Again, there were no matches to this pathogen among the shore-unique contigs. A similar pattern was also observed for contigs matching *Gyrodactylus salmonis* [GENBANK: JN230351], a pathogenic flatworm (Rubio-Godoy *et al*. 2012).

To test the possibility that there were reads matching these pathogens that had not been assembled into contigs, we mapped the raw transcriptome reads from each library to the NCBI reference sequence from each of these pathogens. This assembly supported the prevalence of pathogen sequences among the stream ecotype transcriptome data (Table 3.5). The exception was infectious hematopoietic necrosis virus (IHNV), which had been identified in the high coverage data set and was expected to be present in both ecotypes.

Each of these pathogens is associated with reduced fitness in salmonids, not only by killing adults before they are able to spawn (Van West 2006), but also by persisting in the spawning location and infecting emerging juveniles in the next season (Duchaud *et al*. 2007). A reduction in pathogen sequences present among kokanee collected in shore-spawning habitats suggests the possibility that shore-spawning behaviour may have evolved, in part, as a way to reduce the pathogen load. While highly speculative, this hypothesis is consistent with other studies (Frazer and Russello, in review), which detected outlier loci associated with immune response in kokanee. Further, the fact that only internal organs and subcutaneous muscle tissue were sampled increases the probability that the observed pathogen sequences were indeed present within the fish tissue, rather than being environmental contaminants that could occur if external fin-clips or operculum punches had been collected. While the present study was not explicitly designed to address this question and may be influenced by stochastic factors, our results warrant future research to

quantitatively compare the pathogen load among kokanee within each of the two spawning habitats.

Subsequent research examining *Flavobacterium spp*. diversity (Appendix A) has detected the presence of several *Flavobacterium species* present in the muscle tissue of kokanee from both ecotypes in Okanagan Lake. Quantitative PCR assays designed to test for differences in *Flavobacterium* abundance in kokanee operculum tissue did not find significant differences between ecotypes (Appendix A). However, these data were from operculum tissue collected from dead kokanee, and may not be an accurate representation of infection level. Future quantitative studies should examine *Flavobacterium* abundance in muscle tissue sampled from live kokanee in their spawning habitat, as well as the pathogen abundance in water samples taken from the spawning environments.

All contigs that matched pathogen sequences where removed from the ecotype-unique transcriptome data sets prior to down-stream analyses. Blast2Go was then used to assign GO annotations to Blastx matches (e-value threshold = 10^{-6}) among the remaining contigs that were unique to each ecotype. This analysis found sequence similarity matches for 160 and 76 contigs from the stream and shore data sets, respectively. From these contigs with positive BLAST matches, 64 stream contigs and 34 shore contigs were subsequently annotated with at least one GO term (Figure 3.3).

3.3.7 Conclusions

In this study, we used next-generation sequencing technology in order to compare transcriptome-wide patterns of sequence variation among divergent ecotypes of kokanee salmon. We identified 32,699 putative SNPs that could be used for population genetic and genomic studies of both kokanee and anadromous sockeye salmon. We further detected

contigs that were unique to each ecotype, which may be indicative of differential gene expression.

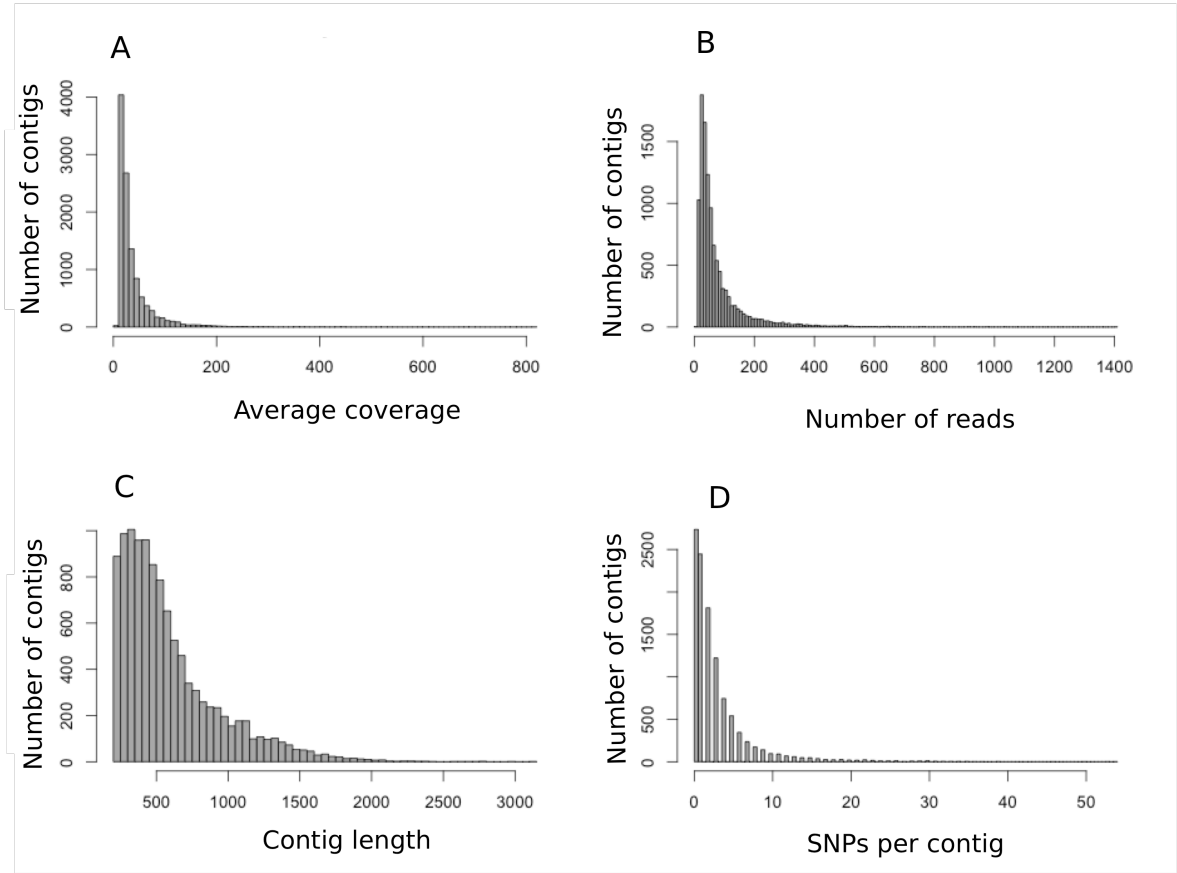

Figure 3.1: Characterization of the contigs present in the high coverage data set**.** Histograms represent (A) average coverage of each contig (mean = 37.0), (B) number of reads (mean = 77.9), (C) contig lengths (mean = 594.8 bases), and (D) the number of SNPs for each of the high coverage contigs.

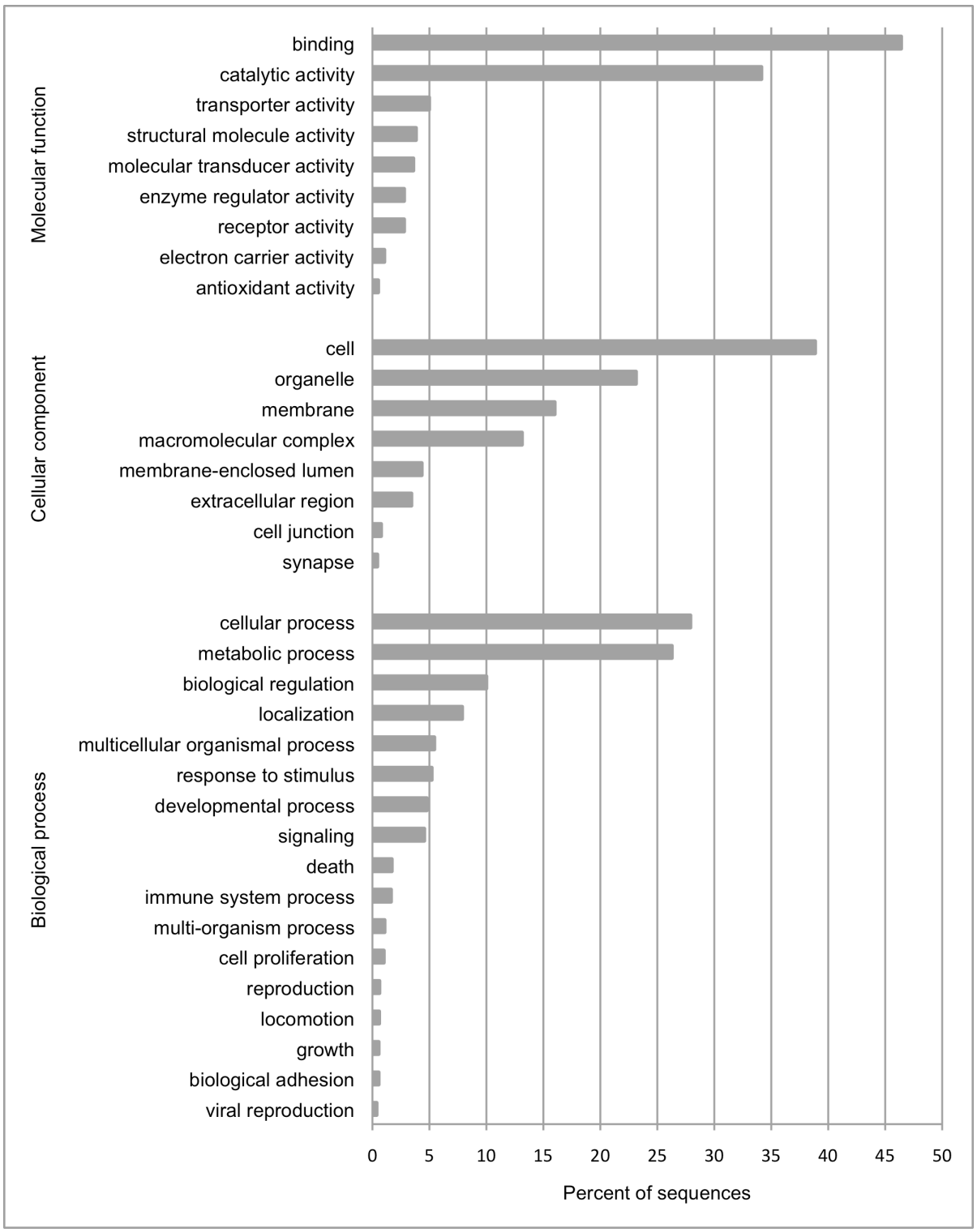

Figure 3.2: Functional annotation of the high coverage contigs. The frequency (%) of each observed gene ontology (GO) term is given for the three GO domains (biological process, cellular component, and molecular function).

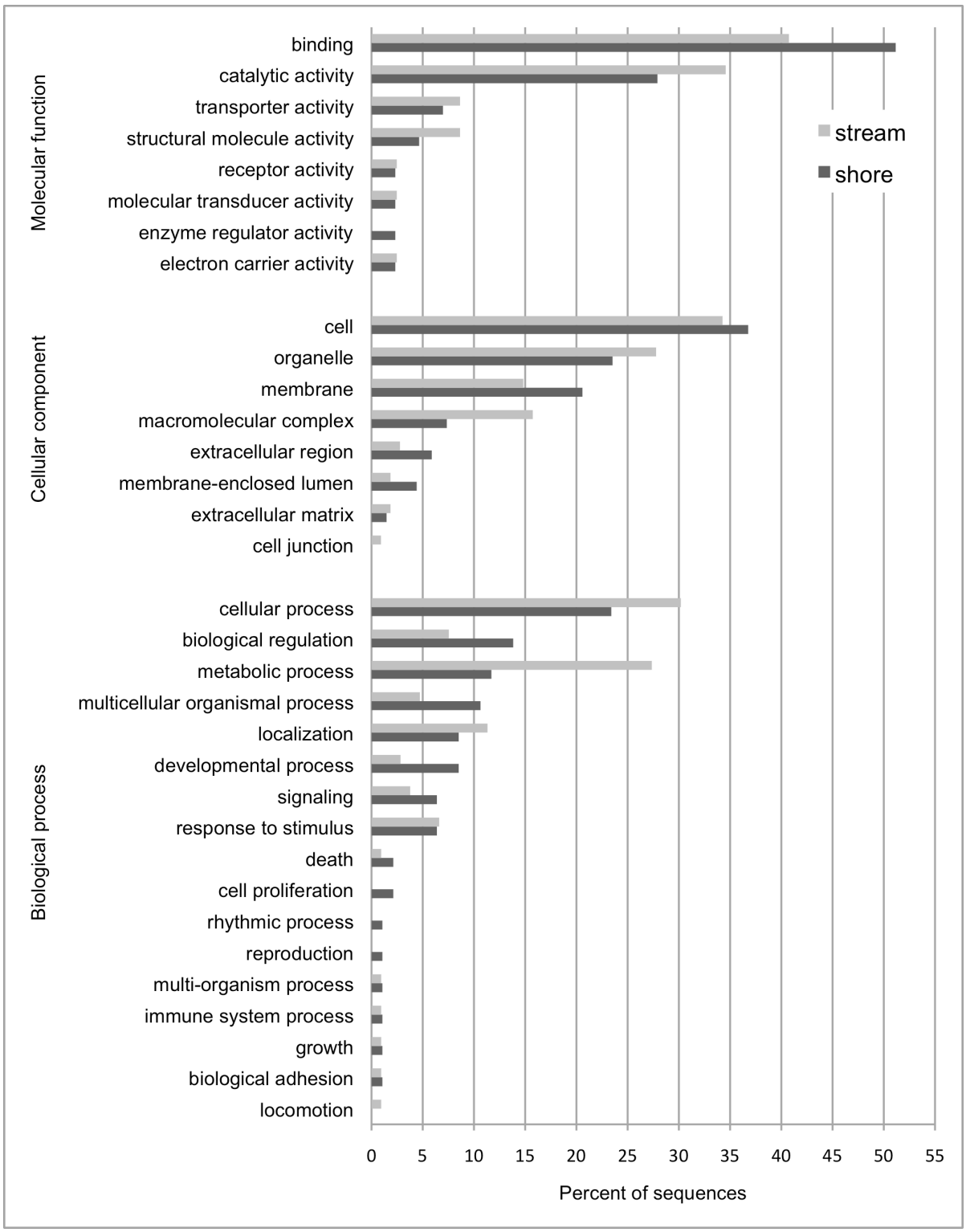

Figure 3.3: Functional annotation of contigs that were unique to each ecotype. The frequency (%) of each observed gene ontology (GO) term is presented for both ecotypes.

Table 3.1: Collection information of all individuals used for each pooled cDNA library.

Ecotype	Spawning Site	Sex
Stream-Spawner	Mission Creek	1. Female
		2. Male
	Penticton Creek	3. Male
		4. Female
	Peachland Creek	5. Male
		6. Male
	Powers Creek	7. Male
		8. Male
Shore-Spawner	Southeast Shore	1. Female
		2. Male
		3. Female
	Northeast Shore	4. Male
		5. Female
	Northwest Shore	6. Male
		7. Male
		8. Female

Table 3.2: Summary of next-generation sequence data obtained for each ecotype of Okanagan Lake kokanee.

	Shore-spawner library	Stream-spawner library
No. of bases	371,876,524	373,169,057
No. of reads	1,343,483	1,406,375
Mean read length	276.8 bases	265.3 bases

Table 3.3: Summary of the contigs present in each kokanee data set.

	Total contigs	High coverage data set[1]	Shore-unique contigs[2]	Stream-unique contigs[2]
No. of contigs	123,547	11,074	277	557
Mean coverage	7.5	37.0	6.8	8.4
Mean length	463.7	594.8	374.2	404.2
Mean no. of reads	14.2	77.9	8.5	12.7

[1] In the high coverage data set each contig has a minimum length of 200 bases and a minimum of 5x coverage for each ecotype. In addition, this data set has duplicate contigs and contigs that map to pathogen DNA sequences removed.

[2] Contigs composed of reads from a single ecotype. Minimum length = 200 bases; minimum coverage = 5x. Contigs that map to pathogen DNA sequences have been removed.

Table 3.4: Genetic diversity estimates from loci that were successfully genotyped using High Resolution Melt Analysis (HRMA).

Locus	H_E/H_O			Major allele[1] frequency from HRMA (frequency predicted from transcriptome data)		
	Stream	Shore	Sockeye	Stream	Shore	Sockeye
One74958	0.13 / 0.00 *	0.32 / 0.00 *	0.19 / 0.05 *	0.93 (1.00)	0.80 (0.48)	0.90
One81166	0.29 / 0.29	0.32 / 0.40	0.41 / 0.48	0.18 (1.00)	0.20 (0.46)	0.28
One81284+	0.47 / 0.57	0.43 / 0.61	0.47 / 0.57	0.74 (1.00)	0.70 (0.54)	0.63
One81385	0.20 / 0.22	0.38 / 0.36	0.19 / 0.22	0.79 (0.52)	0.75 (1.00)	0.90
One113434	0.32 / 0.40	0.32 / 0.40	0.00 / 0.00	0.80 (1.00)	0.80 (0.77)	1.00
One73115	0.43 / 0.52	0.44 / 0.47	0.16 / 0.17	0.78 (0.42)	0.67 (0.79)	0.91
One74836	0.49 / 0.37	0.50 / 0.53	0.43 / 0.46	0.58 (0.33)	0.58 (0.65)	0.32
One24190	0.22 / 0.25	0.35 / 0.45	0.44 / 0.48	0.88 (0.31)	0.77 (0.85)	0.67
One73476	0.50 / 0.61	0.50 / 0.25 *	0.50 / 0.57	0.57 (0.35)	0.56 (0.70)	0.52

* Denotes significant deviation from HWE following sequential Bonferroni correction.

[1] The identity of the major allele is defined as the allele with the highest frequency in the transcriptome data.

\+ The contig from which this locus was created had some overlap with one other contig (34452). Both contigs were subsequently removed from the high coverage data set. As the overlap did not impact the SNP site, this locus has been retained.

Table 3.5: Number of next-generation sequencing reads that aligned to reference sequences from four salmonid pathogens.

Pathogen [GenBank accession]	Number of reads aligned	
	Stream-spawner library	Shore-spawner library
Flavobacterium psychrophilum Complete genome [AM398681]	7,393	219
Gyrodactylus salmonis Partial ITS1, complete 5.8S rRNA gene, partial ITS2 [JN230351]	17	0
Saprolegnia ferax Mitochondrion, complete genome [AY534144]	393	1
Infectious hematopoietic necrosis virus[1] Glycoprotein (G) and non-virion protein (NV) genes [IHNGNVJ]	285	327

[1] IHNV was identified in the high coverage data set containing reads from both ecotypes and was not expected to show ecotype specificity.

Chapter 4: Genomic evidence for ecological divergence in kokanee

4.1 BACKGROUND

Divergent natural selection across a heterogeneous landscape has long been recognized as the dominant force shaping phenotypic diversity within species (Darwin 1859; Endler 1986), and can also be an important factor in the evolution of reproductive isolation between populations that inhabit different environments or niches (Schluter 2001; 2009).

At the molecular level, genome-wide signatures of divergent selection can be inferred through the identification of statistical outlier loci, which are genes that deviate from a neutral model of evolution when compared among different populations (Luikart *et al*. 2003; Storz 2005; Allendorf *et al*. 2010). Patterns of nucleotide divergence at these loci provide evidence that natural selection, rather than neutral processes, govern their distribution (Lewontin & Krakauer 1973; Nielsen 2001; Beaumont & Balding 2004; Nosil *et al*. 2009). Further, the rapidly expanding library of published of genomic data makes it increasingly possible to annotate outlier loci to genes of known function (Galindo *et al*. 2010; Prunier *et al*. 2012; Frazer & Russello 2013).

Fish inhabiting post-glacial lakes provide ideal systems for studying adaptive population divergence (Schluter 1996a; b; Hendry *et al*. 2000). Following the last glacial maximum, newly formed lakes were colonized by many species of fish that have since undergone rapid adaptive radiations to fill available niches (reviewed by Schluter 1996b). For example, lake whitefish (*Coregonus sp*.) and threespine stickleback (*Gasterosteus sp*.) have undergone trophic differentiation, in which both species have evolved a larger benthic ecotype and a small planktivorous limnetic ecotype, each with divergent

morphological phenotypes locally adapted to their respective ecological niches (Schluter & McPhail 1992; Schluter 1995; 1996b; Rundle *et al*. 2000; Bernatchez *et al*. 2010). The parallel evolution of these divergent ecotypes in lakes across their range provides strong evidence for natural selection (Hendry 2009). Further, the use of high-density genome scans of both species has provided key insights into the genomic architecture underlying the process of ecological speciation with gene flow (Peichel *et al*. 2001; Bernatchez *et al*. 2010).

Kokanee, the freshwater form of sockeye salmon (*Oncorhynchus nerka)*, have also undergone ecotype divergence following the colonization of post-glacial lakes in North America. However, unlike stickleback and whitefish, ecotype divergence in kokanee does not appear to be associated with trophic differentiation and doesn't involve obvious morphological divergence (Taylor *et al*. 1997). Rather kokanee ecotypes exhibit divergent reproductive behaviour in their choice of spawning habitat and in the timing of spawning. The 'stream-spawning' ecotype exhibits typical sockeye behaviour, migrating into tributaries to spawn in early autumn. Conversely, the 'shore-spawning' ecotype forgoes a tributary migration and instead spawns directly on the submerged shoreline of lakes in late fall or winter (Taylor *et al*. 2000; Winans *et al*. 2003). Despite the spatial and temporal segregation of their reproductive behaviour, outside of the spawning season both ecotypes co-occur in many lakes and are morphologically indistinguishable (Taylor *et al*. 1997).

During the 1900s, many lakes experienced declines in kokanee population sizes resulting from factors such as competition with invasive species, reduced nutrient input, and loss of spawning habitat (Thompson 1999; Askey & Johnston 2013). Conservation

strategies for kokanee tend to manage the ecotypes as distinct stocks (Askey & Johnston 2013), yet a lack of morphological differences precludes in-season monitoring of stock abundance. Kokanee management would be significantly enhanced by the identification of molecular markers capable of assigning individuals to the correct ecotype, however the identification of sufficiently fine-scale markers has been hindered by the recent divergence and correspondingly weak genetic structure of the two ecotypes (Lemay & Russello 2012; Russello *et al*. 2012; Frazer & Russello 2013).

Okanagan Lake, British Columbia, has emerged as a model location for investigating the genetic basis of ecotype divergence in kokanee and for testing a genetics-based approach to inform fisheries management. Early research found no significant differences between ecotypes of kokanee using allozyme, minisatellite, and microsatellite markers (Taylor *et al*. 1997; Taylor *et al*. 2000). However, Taylor *et al*. (1997) identified mitochondrial DNA restriction fragment length polymorphism (RFLP) haplotypes that were significantly differentiated between stream and shore-spawners, providing the first molecular evidence for restricted gene flow between kokanee ecotypes. More recently, Russello *et al*. (2012) carried out a population genomic assessment of ecotype differentiation in Okanagan Lake. Using a panel of 52 expressed sequence tag (EST)-linked and non-EST-linked microsatellites, they identified eight loci that were statistical outliers; while they found no evidence for genetic divergence at neutral loci, the outlier loci showed significant population genetic structuring between ecotypes. These studies provide preliminary evidence for restricted gene flow and adaptive population divergence, yet the limited genomic coverage and inconsistency among different marker types makes it difficult to draw definitive conclusions.

Here, we used restriction site associated DNA sequencing (RAD-seq) in order to test for signatures of divergent selection between reproductive ecotypes of kokanee. We identified and annotated a suite of outlier loci that are candidate regions under divergent selection and exhibit genetic structure beyond what is apparent at neutral loci. We further demonstrated the utility of these outlier loci for individual assignment and mixed composition analyses to inform stock assessment and guide fisheries management.

4.2 METHODS

4.2.1 Sample collections

This study focuses on several spawning populations of kokanee salmon from Okanagan Lake, British Columbia, which supports a kokanee fishery composed of both reproductive ecotypes (Taylor *et al*. 2000; Askey & Johnston 2013). Boat-based surveys of Okanagan Lake have observed shore-spawning behaviour in most undeveloped sections of the shoreline; stream-spawning is currently monitored in 18 streams, though the majority of stream-spawning (89% in 2013) occurs in four main tributaries (Paul Askey, unpublished data). We obtained tissue samples (either operculum punches or fin clips) from spawning adult kokanee at the four major stream-spawning tributaries and at three shore-spawning sites in Okanagan Lake (Figure 4.1, Table 4.1). In order to test for temporally divergent genetic patterns among cohorts, replicate samples were obtained from two different years (2007 and 2010) (n = 24 individuals per ecotype per year; 96 total).

4.2.2 Molecular methods

Restriction site-associated DNA sequencing provides a rapid, cost effective, method for

generating genome-wide sequence data in wild populations of non-model species (Baird *et al*. 2008; Etter *et al*. 2011). We generated RAD libraries using a modified version of the protocol described by Etter *et al*. (2011). Genomic DNA was extracted using the NucleoSpin Tissue Kit (Macherey-Nagel) following the manufacturer's suggested protocol with the addition of RNase A (Qiagen). For each individual sample, 500ng of DNA was digested using the *Sbf1* RE. Four RAD libraries were constructed using pools of 24 individually barcoded samples. The barcodes used were six nucleotides in length and each differed by at least two bases (Hohenlohe *et al*. 2010; Miller *et al*. 2012). During the library preparation, a Bioruptor® (Diagenode) was used to shear DNA strands to a mean length of ~500 base pairs and a targeted fragment-size selection device (Pippin Prep™, Sage Science) was used instead of gel extractions to isolate DNA fragments between 400-600 base pairs in length. One full lane of Illumina HiSeq 2000 sequencing was carried out for each of the four RAD libraries.

4.2.3 Assembly and SNP discovery

Raw Illumina reads from the ecotype libraries were processed into RAD-tags using the STACKS bioinformatic software pipeline (Catchen *et al*. 2011; Catchen *et al*. 2013). Initially, the PROCESS_RADTAGS program was used to separate reads by their barcode, remove low quality reads (Phred quality score < 10), trim all reads to 90 base pairs in length, and remove any reads that did not contain the *Sbf1* recognition sequence. Next, the USTACKS program was used for the *de novo* assembly of raw reads into RAD-tags. The minimum number of reads to create a stack was set at 2 (*m* parameter in USTACKS), and the maximum number of pairwise differences between stacks was 2 (*M* parameter in USTACKS). A catalogue of RAD-tags was then generated using all 96 kokanee in

CSTACKS. Several different assemblies were carried out with a range of parameter values in order to optimize the assembly for this system. Given the genome duplication event in the history of salmonid evolution, genomic samples are expected to contain a high proportion of paralogous sequence variants (PSVs); optimization of assembly parameters in salmonid systems is a fine balance between separating PSVs from their functional genes while not overly splitting informative variation. In general, we took an approach that was more likely to remove PSVs at the expense of potentially separating truly divergent sequence variants into separate loci. Given the large-scale of the data, we feel that the advantages of this approach (i.e. removing PSVs) outweigh the cost of potentially discarding informative loci.

Following assembly and genotyping, the data was further filtered in order to maximize data quality. Using the POPULATIONS module, we retained only those loci that were genotyped in ≥ 80% of individuals from each ecotype, had a minor allele frequency ≥ 0.05, and a minimum coverage of 10x per allele for each individual. Data was exported from STACKS in GENEPOP format (Raymond & Rousset 1995), and converted for subsequent analyses using PGD SPIDER v. 2 (Lischer & Excoffier 2012).

To further reduce the occurrence of PSVs in the data, we removed all loci that displayed significant deviation from Hardy-Weinberg equilibrium (HWE) as assessed using the method of Guo and Thompson (1992) implemented in GENEPOP 4.2 (Raymond & Rousset 1995; Rousset 2008). Loci were removed if they deviated from HWE in both ecotypes.

In order to annotate RAD-tags to a putative genetic function, each haplotype from the combined datasets of neutral and outlier RAD-tags were subject to a sequence

similarity search (BLASTN) of the NCBI non-redundant (nr) database (e-value ≤ 10^{-5}) as implemented in the CLC GENOMICS WORKBENCH v. 6.5 (CLC Bio). If there were several different results for a given RAD-tag, then the BLAST hit with the lowest e-value was retained.

4.2.4 Population genetic analyses

Polymorphic loci were screened for statistical outliers using the Bayesian simulation method of Beaumont and Balding (2004) as implemented in BAYESCAN 2.1 (Foll & Gaggiotti 2008). For this analysis all samples were coded as two ecotypes (stream vs. shore-spawners), we used a prior odds value of 10, with 100,000 iterations and a burn-in of 50,000 iterations. We identified loci that were significant outliers at false discovery rates (FDR) of 0.20 and 0.05. Loci that were identified as outliers were segregated into an outlier dataset; the remaining loci comprised the 'neutral' dataset that was used for population genetic analyses (see below).

Pairwise neutral and outlier F_{ST} was calculated between ecotypes for each sampling year (2 years x 2 ecotypes) using ARLEQUIN 3.5 (Excoffier & Lischer 2010). In order to test for differences in the distribution of genetic variation between ecotypes, an analysis of molecular variance (AMOVA) was carried out using both the neutral and outlier datasets; AMOVAs were implemented in ARLEQUIN 3.5 (Excoffier & Lischer 2010) using 1000 permutations.

We tested for population genetic structure using the Bayesian method of Pritchard *et al*. (2000) as implemented in STRUCTURE v.2.3. First the entire neutral dataset was analyzed with a run length of 100,000 Markov chain Monte Carlo steps following a burn-in of 50,000 steps. The analysis was carried out using the admixture model without

incorporating prior information. We varied the number of clusters (*K*) from 1 to 15, with 5 replicates for each value of *K*. The most likely number of clusters was determined by plotting the natural log probability of the data (ln Pr(X|K); Pritchard et al. 2007) across the range of *K* values tested and selecting the *K* where the value of *ln Pr(X|K)* plateaued as suggested in the STRUCTURE manual. We also employed the Δ*K* method of Evanno et al. (2005) as implemented in STRUCTURE HARVESTER (Earl & Vonholdt 2012). Inferred assignment of individuals to each cluster was averaged across replicate values of *K* using CLUMPP v.1.1.2 (Jakobsson & Rosenberg 2007).

Bayesian clustering analysis was also used to assess the level of genetic structure at outlier loci. This analysis used the genotypic data for all 96 kokanee of known ecotype at the outlier loci identified above. The STRUCTURE analysis was carried out with a run length of 100,000 Markov chain Monte Carlo steps following a burn-in of 50,000 steps using the admixture model without incorporating prior information; the number of clusters (*K*) from 1 to 15, with 5 replicates for each value of *K*.

4.2.5 Mixed stock analyses

We tested the effectiveness of the outlier loci identified above as mixed stock assignment tools using the program ONCOR (Kalinowski *et al*. 2007). First, we compared the assignment accuracy of 20 outlier loci against 20 randomly chosen neutral loci and 500 randomly chosen neutral loci using the 100% mixture simulation method of Anderson *et al*. (2008). This analysis works by simulating a sample in which all individuals are from the same population (i.e. the proportion of each stock is set to 1.0), and then determining the assignment accuracy to each stock. Each test was carried out using 100 simulations with 200 individuals simulated in the mixture.

Next, we performed realistic fisheries simulations using the same three datasets (20 outlier, 20 random neutral, 500 random neutral). This was accomplished by simulating a mixture of 200 individuals (based on allele frequencies of the stream- and shore-spawners of known ecotype) in which the mixture compositions of the two ecotypes (stream: shore) varied as follows: 0.05:0.95, 0.10:0.90, 0.25:0.75, 0.50:0.50, 0.75:0.25, 0.90:0.10, 0.95:0.05. The mixture analysis then estimated the proportion of each ecotype from each of the simulated datasets in order to assess the accuracy of each set of markers (20 outlier, 20 random neutral, 500 random neutral).

4.2.6 Genetic stock identification

We applied the RAD-seq approach to a realistic fisheries scenario in which the outlier markers identified from our baseline of known ecotype were used to estimate the proportion of each ecotype in a trawl sample (i.e. a mixed sample of individuals from both ecotypes). For this experiment, juvenile kokanee of unknown ecotype were collected by trawl in Okanagan Lake in September 2008. The trawl was carried out at night using a 3 x 6m net on a bearing due west for 60 minutes at 3.1 km/hr (see Figure 4.1). Individuals with fork length ≤ 60 mm were retained for sequencing; this size is indicative of 'age 0' kokanee, making them the likely offspring of 2007 spawners. Tissue sampling, DNA extraction, and RAD library preparation were carried out using the same protocols as listed above (Section 4.2.2). The 48 individuals were given unique barcodes in two pools of 24; one full lane of Illumina HiSeq 2000 was carried out for each pool of 24 individuals.

Assembly and genotyping of RAD data for the trawl samples were carried out using components of the STACKS pipeline as described above (Section 4.2.3), with the

exception that the catalogue of RAD-tags developed for the kokanee of known ecotype was used for genotyping. This approach allowed the trawl to be genotyped at the same loci that were used for the previous samples, facilitating direct comparisons of sequence variation across samples.

We used the program ONCOR to perform individual assignment tests for all samples in the trawl dataset at the 20-outlier loci identified above. This analysis used the kokanee of known ecotype (n = 48 stream, n = 48 shore) as the baseline, and then estimated the probability of assignment of each trawl individual to the stream or the shore ecotype.

4.3 RESULTS

4.3.1 Sequencing and assembly

Illumina sequencing produced ~350,000,000 reads per lane. Assembly and quality filtering resulted in the identification of 6,877 RAD-tags that contained a total of 7,765 SNPs. We further filtered the dataset by retaining only a single SNP per RAD-tag, and by removing SNPs that significantly deviated from HWE. This produced a final dataset of 5,976 high quality SNPs (mean F_{ST} = 0.005; Figure 4.2) for subsequent analyses (Table 4.2).

4.3.2 Outlier loci

Twenty high F_{ST} outliers were identified at FDR = 0.20, with an average F_{ST} of 0.13 between ecotypes (neutral mean F_{ST} was 0.005). Of these loci, 11 were also detected as outliers when the FDR was reduced to a highly conservative value of 0.05 (Table 4.3).

Within the outlier dataset, 14 RAD-tags matched sequences from the NCBI nr

database (Table 4.3; e-value ≤ 10^{-5}). Of particular interest, five outlier loci were annotated to genes that are associated with sex determination and early sexual development: Two outliers (14305_21 and 110128_19) mapped to a *Foxl2*-like protein identified in Atlantic salmon, *Salmo salar* [NCBI HM159472]. This gene is involved in sex-determination and gonad development, with gene expression observed early in development prior to the formation of gonads (von Schalburg *et al*. 2011). Locus 14305_21 was retained at the more conservative FDR = 0.05.

Similarly, one outlier (9426_88) matched a follicle stimulating hormone beta subunit (FSHbeta) identified in Chinook salmon, *Oncorhynchus tshawytscha* [NCBI AY493564]. Follicle–stimulating hormones are gonadotropins involved in the growth and development of gonads (Chong *et al*. 2004). This outlier was retained at FDR = 0.05 and was fixed for a single allele among the shore-spawners. Two additional outliers (29236_34 and 88470_57) mapped to a *DOT1*-like histone H3 methyltransferase identified in Atlantic salmon [NCBI HM159472]; *DOT1* is a 'Disruptor of telomeric silencing' gene involved in the inhibition of gene transcription through methylation (Singer *et al*. 1998; Ng *et al*. 2002). *DOT1* gene disruption has been linked to several processes including cell cycling, embryonic development and cardiac function (Nguyen & Zhang 2011). Further, *DOT1* is associated with *MIS* (Mullerian-inhibiting substance), which plays a critical role in gonad development (von Schalburg *et al*. 2011).

4.3.3 Population genetic analyses

Pairwise analysis of F_{ST}, suggest a lack of neutral genetic structure between years and ecotypes (Table 4.4). Similarly, the AMOVA used to examine structure between ecotypes indicated that divergence at neutral loci was non-significant (Table 4.5), with

100% of variation occurring within treatments; these results contrast with those using the outlier data, which estimated that a significant proportion of variation (24%) occurs among ecotypes but not between years.

Bayesian clustering analysis did not find evidence for neutral genetic structure among ecotypes, however the results were variable depending on the method of interpretation. When all individuals of known ecotype were analyzed at 5,956 neutral loci, a visual assessment of the change in *Ln P(K)* suggests that all values of *K* from 1-5 were equally probable (Figure 4.3a), which is indicative of a lack of genetic structure (i.e. K = 1; Pritchard *et al*. 2007). However, the method of Evanno et al. (2005) estimated that the optimal structure was $K = 3$, but at a very low ΔK value of 20.4. It is important to note that, due to the manner in which it is calculated, the ΔK method of Evanno et al. (2005) is incapable of inferring a $K = 1$. In contrast to the results based on neutral loci, Bayesian clustering analyses carried out using the 20 outlier loci on same 96 individuals revealed strong evidence for genetic structuring among ecotypes ($K = 2$, $\Delta K = 1385.4$; Figure 4.3b&c).

4.3.4 Mixed composition analyses

Using the 100% simulation method of Anderson *et al*. (2008), the assignment accuracy of outlier loci was estimated to be > 0.99 for each ecotype (Table 4.6). Outlier loci significantly outperformed an identical number of neutral loci (n = 20) as well as 500 neutral loci (Table 4.6). These results were further corroborated by the realistic fisheries simulation tests in which outlier loci outperformed 20 neutral and 500 neutral loci across every simulated stock composition (Table 4.7); for each simulated stock proportion, the panel of 20 outlier loci were not significantly different from the true proportion.

Conversely, there were significant deviations between estimated and true proportions when the neutral data were used. These deviations were most pronounced in skewed mixture compositions when the simulated proportion of the rare ecotype was ≤ 0.10.

4.3.5 Genetic stock identification

Following the *de novo* assembly, RAD-tags from trawl samples were successfully mapped to a catalogue of tags generated for the kokanee of known ecotype. This approach allowed us to determine multi-locus genotypes for all trawl samples at the same genetic markers as the kokanee of known ecotype, with an average of only 1% missing data across all trawl individuals at outlier loci.

Using an individual assignment analysis implemented in ONCOR (Kalinowski *et al*. 2007), we inferred the most likely origin (stream or shore) for each of the trawl individuals. Among these 48 trawl samples, five (10.4%) were predicted to be stream-spawners.

4.4. DISCUSSION

4.4.1 Ecological divergence

We used restriction site associated DNA sequencing to identify ~6000 SNPs in a population-level sample of kokanee from Okanagan Lake, British Columbia. From this large baseline of SNPs, we then identified a suite of 20 outlier loci, which are candidate regions under divergent selection. These 20 loci demonstrate significant genetic structure with respect to ecotype and have very high assignment accuracy in mixed composition simulations, providing evidence that ecotype divergences in kokanee has a genetic basis.

The absence of genetic structure at neutral loci combined with a small number of highly informative outlier loci (Russello *et al*. 2012; and this study) matches theoretical predictions for the earliest stage of divergence with gene flow (Box 2 from Feder *et al*. 2012) in which only loci experiencing divergent selection will show evidence for restricted gene flow (Wolf *et al*. 2010). Indeed, the distribution of F_{ST} values across all loci in this study was similar to the distribution observed between parapatric races of *Heliconius* butterflies (Figure 1Ba from Seehausen *et al*. 2014), which are at an early stage of ecological divergence (Martin *et al*. 2013).These patterns suggest that kokanee may be an informative natural system for investigating genomic mechanisms underlying early stages of adaptive divergence.

Based on the annotation of outlier loci detected in this study, we found evidence that genes associated with early juvenile development may be candidate regions of elevated genomic divergence between reproductive ecotypes of kokanee. While the lack of a fully annotated sockeye genome precludes a structural investigation of the size and frequency of putative islands of divergence, the occurrence of multiple outlier loci that annotate to the same gene or functional class suggests that these may be genomic regions of divergent selection; however, further research is needed to test this hypothesis. These findings are consistent with those of previous studies focused on kokanee spawning behavior and development. For example, Taylor *et al*. (2000) compared several factors associated with early development in stream and shore-spawning kokanee. They found that sub-gravel temperature during spawning and incubation was higher in the shore-spawning sites, resulting in a greater number of accumulated thermal units during their development. Research aimed at investigating differences in early development may be a

promising avenue for identifying traits subject to divergent selective pressures between spawning environments.

While the outlier loci identified in this study constitute promising candidates putatively involved in adaptive evolution, an important caveat is that these statistical patterns only provide correlative evidence for the role of natural selection (Kawecki & Ebert 2004). A direct causal relationship between genotype, phenotype, and fitness is essential before adaptation can be invoked as the driver of these patterns (Gould & Lewontin 1979; Barrett & Hoekstra 2011). Yet, the identification of several similar genes that demonstrate a significant correlation with environmental variation provides a promising avenue for future experimentation (Endler 1986). Further research is needed in order to experimentally link genetic differences observed at these outlier loci with differential fitness between spawning environments

4.4.2 Fisheries management

The use of outlier loci can provide additional insights into the genetic structure of natural populations, offering a powerful tool for conservation and management (Ouborg *et al*. 2010; Funk *et al*. 2012; Russello *et al*. 2012; Bradbury *et al*. 2013). We found that the panel of 20 outlier loci had very high (>99%) accuracy at assigning individuals to the correct ecotype in realistic fisheries simulations. These outlier loci easily out-performed an identical number of randomly chosen neutral loci as well as a large panel of 500 randomly chosen neutral loci. In addition, the loci identified in this study outperformed previous assignment tests carried out with neutral microsatellites (Taylor *et al*. 2000) and outlier EST-linked microsatellite loci (Russello *et al*. 2012).

We also observed that each of the non-outlier datasets tended to over-estimate the

abundance of rare samples when the simulated mixture compositions were skewed towards one ecotype (table 4.7). This outcome was expected given that mixture analyses tend to be biased towards values of $1/k$, where k is the number of stocks in the baseline (i.e. 50:50 in this case; Kalinowski *et al*. 2007). An overestimation of rare stocks would have negative consequences for kokanee fisheries management in Okanagan Lake, which continually fluctuates in stock composition and where the once abundant stream-spawning population is in decline (Askey & Johnston 2013). Current estimates have the stream-spawning ecotype representing only 5% of the total Okanagan Lake kokanee population (Paul Askey, unpublished data). Given these significant differences in the relative abundance of each ecotype, markers that are sufficiently robust to handle drastically skewed stock proportions are essential for guiding effective fisheries management.

As proof of concept, we applied a RAD-seq approach for genotyping a mixed kokanee sample of unknown composition obtained from a trawl. We found the STACKS pipeline to be ideal for this application as it provides a catalogue of loci that can be used to genotype new samples (i.e. samples that were not used to generate the initial base line), making it straightforward to calibrate genotyping results across different projects. Using a Bayesian clustering algorithm (Pritchard *et al*. 2000), we then assigned individuals of unknown origin to their most likely ecotype. Our results suggested a very low proportion of stream-spawners in the trawl sample as expected given that: (1) the overall proportion of stream-spawners in Okanagan Lake is very low (~5%; Paul Askey, unpublished data); and (2) that the trawl took place in the northern section of the lake (see Figure 4.1), far from the four major tributaries where stream-spawning occurs.

Overall, we found RAD-seq to be effective in providing markers for fisheries management. While this approach would produce much more data than is typically needed, it has the advantage that genome-wide data are already collected if additional loci from the baseline become useful. In addition, this approach avoids the large upfront costs of locus-specific 5'exonuclease assays (e.g. TaqMan), which may eventually prove uninformative in other populations across a species' range.

4.4.3 Summary

The use of modern 'genotyping by sequencing' methods provides an abundance of genome-wide SNP data that can be used to inform studies in ecology, evolution, and conservation biology (Narum *et al*. 2013). We identified >5000 SNPs derived from RAD-tags in two reproductive ecotypes of kokanee salmon. Outlier tests revealed 20 loci that are putatively under divergent selection between stream and shore-spawners, many of which annotated to genes associated with early development. These outliers had very high (>99%) assignment accuracy to ecotype, suggesting that they may have utility as genetic tools for stock identification. Overall, our results provide strong evidence that reproductive ecotypes of kokanee are in the early stages of ecological differentiation, making them an informative system for investigating the genomic architecture underlying adaptive divergence. Future research should test whether there are parallel patterns of genomic divergence at these outlier loci in lakes across their range.

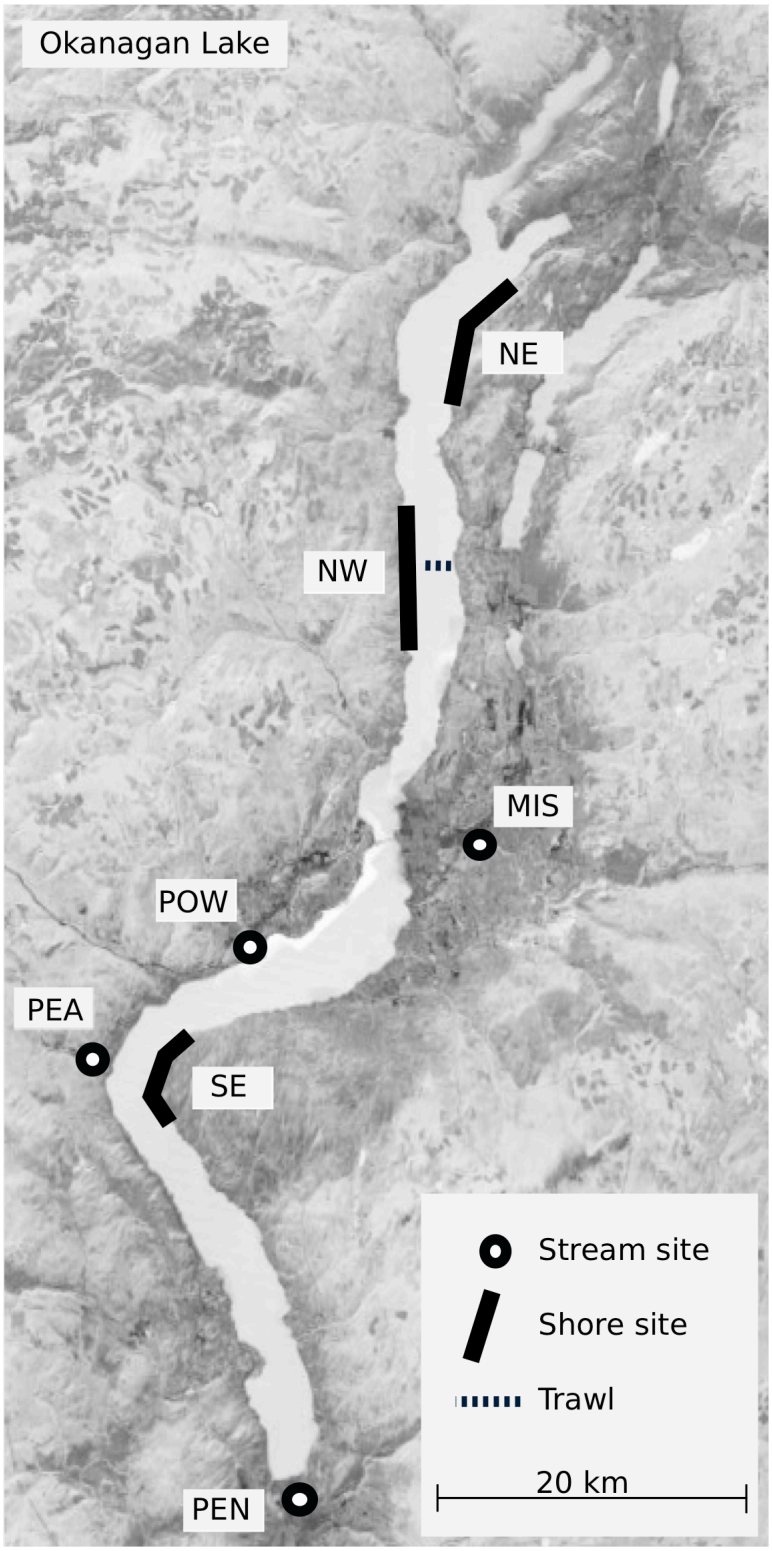

Figure 4.1: A map of Okanagan Lake identifying each of the seven sampling locations used in this study. Detailed information about each site is included in Table 1.

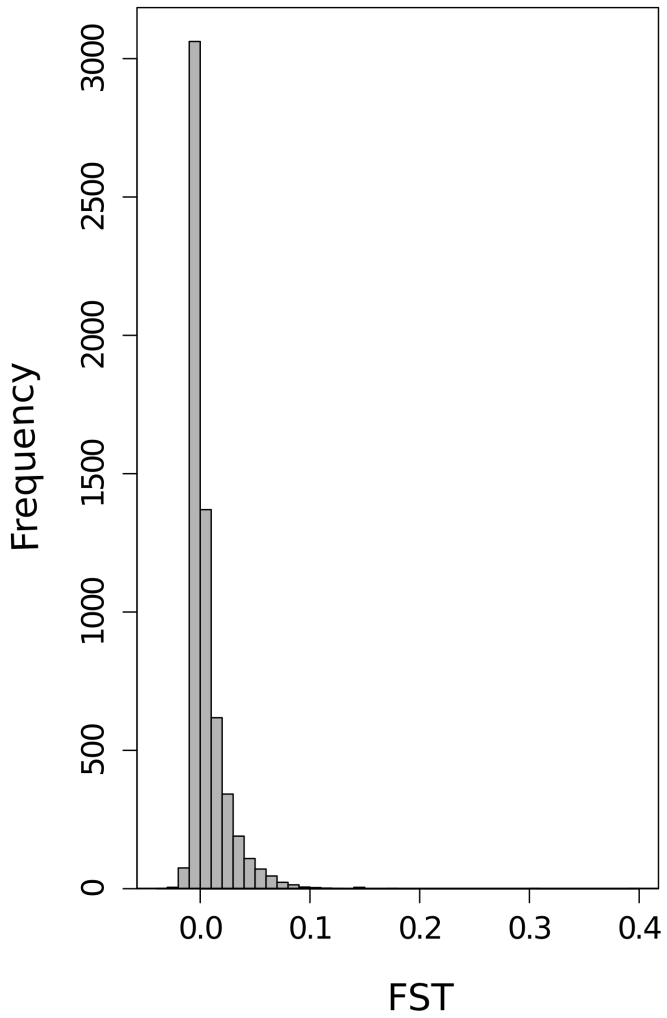

Figure 4.2: Frequency distribution of F_{ST} values obtained across all loci (neutral and outlier; n = 5976) using the method of Weir and Cockerham (1984) followed by Fisher's exact tests for statistical significance.

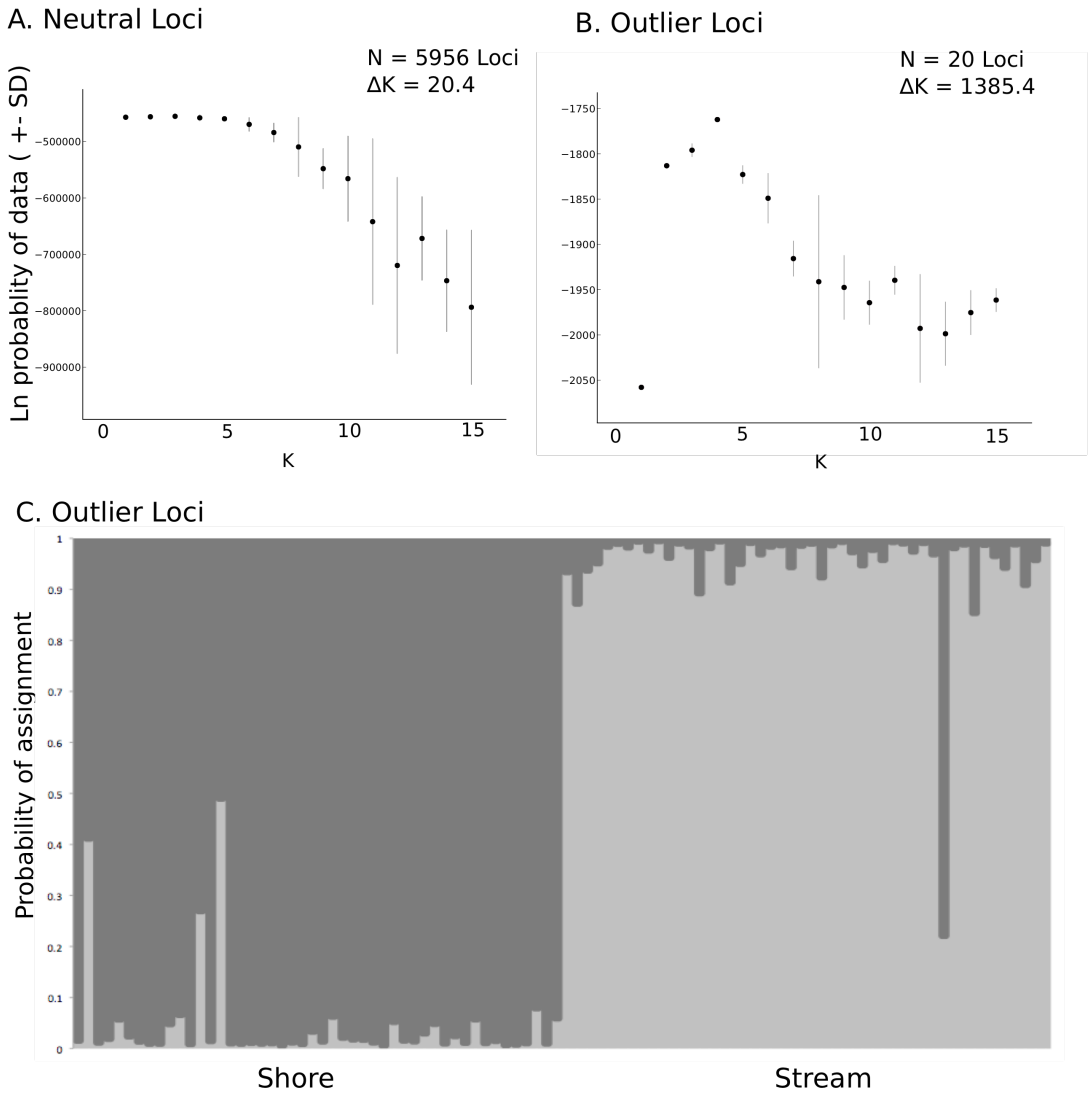

Figure 4.3: Results of the Bayesian clustering analysis carried out for all individuals (n = 96) of known ecotype as implemented in the program STRUCTURE: (A) visual representation of *Ln P(K)* for all neutral loci (5956 loci); (B) visual representation of *Ln P(K)* for all outlier loci (20 loci). Both analyses were run with 100,000 steps and a burn in of 50,000, with five replicates of each value of *K* ranging from 1-15. Based on the delta-*K* method of Evanno et al. 2005 an optimal *K* = 2 was retained (*ΔK* = 1385.4); (C) A barplot displaying the percent membership of each individual to the inferred number of clusters (*K* = 2) using outlier loci.

Table 4.1: Distribution of samples used to generate RAD-libraries for Illumina sequencing. Samples were collected from seven locations in Okanagan Lake, encompassing two ecotypes (stream & shore) and two spawning years (2007 & 2010).

Treatment	Location	Sample size
Stream 2007 (n = 24)	Mission Creek (MIS)	8
	Peachland Creek (PEA)	6
	Penticton Creek (PEN)	5
	Powers Creek (POW)	5
Stream 2010 (n = 24)	Mission Creek (MIS)	6
	Peachland Creek (PEA)	6
	Penticton Creek (PEN)	6
	Powers Creek (POW)	6
Shore 2007 (n = 24)	Northwest (NW)	10
	Northeast (NE)	4
	Southeast (SE)	10
Shore (2010) (n = 24)	Northwest (NW)	10
	Northeast (NE)	4
	Southeast (SE)	10

Table 4.2: Single nucleotide polymorphisms (SNPs) retained at each step of the quality filtering process.

Data	Number of SNPs
SNPs that were present in both ecotypes [a]	158 070
SNPs passing additional quality filters [b]	7 765
Retain single SNP per RAD-tag [c]	6 877
SNPs in HWE [d]	5 976

[a] SNPs were required to be present in both ecotypes.

[b] To be retained, SNPs had to be present in ≥ 80% of individuals from each ecotype, have a minor allele frequency ≥ 0.05, and have a minimum coverage of 10 reads per allele.

[c] To reduce linkage in subsequent analyses, only a single SNP was retained from each RAD-tag

[d] HWE = Hardy-Weinburg Equilibrium within treatments (steam 2007, stream 2010, shore 207, shore 2010).

Table 4.3: Summary of outlier loci identified using BAYESCAN at false discovery rate (FDR) = 0.2; loci that were identified as outliers at FDR=0.05 are indicated with an asterisk.

Locus [a]	SNP	F_{ST} [b]	H_O / H_E Stream	H_O / H_E Shore	Top Blast hit [c] (Accession)	Abbreviated description
8199_87	G / T	0.08	0.37 / 0.36	0.45 / 0.50	FJ969488	*Salmo salar* clone BAC CHORI214-114L13 von Willebrand factor A domain containing 5A (VWA5A) gene.
9426_88*	A / T	0.40	0.41 / 0.50	Fixed (T)	AY493564	*Oncorhynchus tshawytscha* follicle stimulating hormone beta subunit (FSHbeta) gene.
12888_79	A / T	0.09	0.17 / 0.23	0.37 / 0.49	NA	
13802_89	A / G	0.09	0.20 / 0.25	Fixed (G)	NM_001140328	*Salmo salar* Zinc finger HIT domain-containing protein 4 (znhi4).
14305_21*	A / G	0.30	0.26 / 0.37	0.34 / 0.34	HM159472	*Salmo salar* clone BAC 261D01 Foxl2-like protein (Foxl2) gene.
29236_34	A / C	0.08	0.05 / 0.04	0.30 / 0.36	HM159473	*Salmo salar* clone BAC 19C14 DOT1-like histone H3 methyltransferase pseudogene.
33487_15*	A / C	0.11	0.53 / 0.51	0.38 / 0.31	XM_005803003	PREDICTED: *Xiphophorus maculatus* protein DVR-1-like (LOC102233502), mRNA.
37568_54*	C / T	0.15	0.32 / 0.40	0.39 / 0.46	NA	

Locus [a]	SNP	F_{ST} [b]	H_O / H_E Stream	H_O / H_E Shore	Top Blast hit [c] (Accession)	Abbreviated description
38617_87*	G / T	0.08	0.41 / 0.39	0.50 / 0.49	EU481821	*Salmo salar* physical map contig 483, genomic sequence.
55106_87	G / T	0.10	0.39 / 0.35	0.50 / 0.50	XM_003446219	PREDICTED: *Oreochromis niloticus* glutaminyl-peptide cyclotransferase-like (LOC100700099), mRNA.
59501_10*	A / C	0.14	0.24 / 0.40‡	0.50 / 45	NA	
62310_56*	C / T	0.12	0.26 / 0.29	Fixed (C)	NA	
66759_73*	C / T	0.15	0.14 / 0.24‡	0.49 / 0.51	BT049187	*Salmo salar* clone ssal-evd-554-269 Transposable element Tc1 transposase putative mRNA.
85651_23	A / T	0.10	0.35 / 0.29	0.51 / 0.50	NA	
88470_57*	C / T	0.15	0.30 / 0.40	0.45 / 0.46	HM159473	*Salmo salar* clone BAC 19C14 DOT1-like histone H3 methyltransferase pseudogene…
92099_85*	G / T	0.12	0.38 / 0.44	0.52 / 0.44	GQ505859	*Salmo salar* clone BAC CHORI214-439H13 genomic sequence.
107674_24	C / G	0.08	0.02 / 0.02	0.36 / 0.30	NA	

Locus [a]	SNP	F_{ST} [b]	H_O / H_E Stream	H_O / H_E Shore	Top Blast hit [c] (Accession)	Abbreviated description
108521_45*	C / T	0.15	0.30 / 0.34	0.32 / 0.48[‡]	EU481821	*Salmo salar* physical map contig 483, genomic sequence.
110128_19	G / T	0.10	0.53 / 0.51	0.31 / 0.32	HM159472	*Salmo salar* clone BAC 261D01 Foxl2-like protein (Foxl2) gene.
121265_54	C / G	0.06	0.13 / 0.13	0.52 / 0.43	XM_003967613	PREDICTED*: Takifugu rubripes* tyrosine--tRNA ligase, mitochondrial-like (LOC101062180), mRNA.

[a] Loci are given in the format "RAD-tag ID _ SNP site"

[b] F_{ST} as calculated in the POPULATIONS module of STACKS, following the method of Weir and Cockerham(1984). All are significantly different from zero (Fishers exact test, $p < 0.0001$). Average F_{ST} across all neutral loci (n = 5,956) was 0.005.

[c] BLASTN of NCBI nr database as implemented in CLC GENOMICS WORKBENCH v. 6.5. Top blast hit was based on having the lowest e-value.

* Detected as outliers at FDR= 0.05

[‡] Significant deviate from HWE ($p \leq 0.05$)

Table 4.4: Pairwise F_{ST} for each comparison between sampling locations and years. Neutral loci (n = 5956) are above the diagonal, outlier loci (n = 20) are below the diagonal. Significant deviations from zero are indicated with an asterisk.

	Stream 2007 (n = 24)	Stream 2010 (n = 24)	Shore 2007 (n = 24)	Shore 2010 (n = 24)
Stream 2007 (n = 24)		-0.011	0.008*	0.012*
Stream 2010 (n = 24)	-0.015		-0.001	-0.006
Shore 2007 (n = 24)	0.220*	0.238*		0.003*
Shore 2010 (n = 24)	0.240*	0.266*	0.006	

Table 4.5: Analysis of molecular variance (AMOVA) results showing the percentage of variation at each hierarchical level of organization for neutral and outlier loci. For this analysis, 'group' = ecotype (stream and shore), and 'population' is each sampling year for each ecotype (n = 4).

	Percentage of variation		
Data	Among groups (Ecotypes)	Among populations within groups	Within populations
Neutral loci (n = 5956)	0.67	-0.31	99.65*
Outlier loci (n = 20)	24.32*	-0.23	75.91*

* $p < 0.05$

Table 4.6: Assignment accuracy for the 100% mixture simulations following the method of Anderson *et al*. (2008) as implemented in ONCOR.

Data	Ecotype	Accuracy (± SD)
20 outlier loci	Stream	0.9999 (± 0.0007)
	Shore	0.9999 (± 0.0006)
20 neutral loci	Stream	0.8277 (± 0.0517)
	Shore	0.8400 (± 0.0592)
500 neutral loci	Stream	0.9413 (± 0.0168)
	Shore	0.9598 (± 0.0158)

Table 4.7: Estimated proportion of stream and shore-spawners from each of seven simulated mixture compositions carried out using realistic fisheries simulations implemented in ONCOR.

Ecotype	Simulated proportion	Estimated proportion (± SD)		
		20 outlier loci	20 neutral loci	500 neutral loci
Stream	0.05	0.05 (±0.02)	0.19 (±0.07)	0.09 (±0.02)
Shore	0.95	0.95 (±0.02)	0.81 (±0.07)	0.91 (±0.02)
Stream	0.10	0.10 (±0.02)	0.23 (±0.06)	0.14 (±0.03)
Shore	0.90	0.90 (±0.02)	0.77 (±0.06)	0.86 (±0.03)
Stream	0.25	0.25 (±0.03)	0.32 (±0.07)	0.26 (±0.04)
Shore	0.75	0.75 (±0.03)	0.68 (±0.07)	0.74 (±0.04)
Stream	0.50	0.50 (±0.04)	0.49 (±0.07)	0.48 (±0.04)
Shore	0.50	0.50 (±0.04)	0.51 (±0.07)	0.52 (±0.04)
Stream	0.75	0.75 (±0.03)	0.66 (±0.06)	0.71 (±0.04)
Shore	0.25	0.25 (±0.03)	0.34 (±0.06)	0.29 (±0.04)
Stream	0.90	0.90 (±0.02)	0.76 (±0.06)	0.84 (±0.03)
Shore	0.10	0.10 (±0.02)	0.24 (±0.06)	0.16 (±0.03)
Stream	0.95	0.95 (±0.02)	0.80 (±0.06)	0.88 (±0.03
Shore	0.05	0.05 (±0.02)	0.20 (±0.02)	0.12 (±0.03)

Chapter 5: Conclusions

5.1 RESEARCH FINDINGS

Pacific salmonids are an ideal system for studying natural selection in the wild. Their broad geographical distribution combined with a unique combination of life history characteristics, which may include anadromy, semelparity, and philopatry, can result in complex patterns of population structure, providing ideal conditions for the evolution of adaptation (Taylor 1991; Fraser *et al*. 2011).

In this dissertation, I used three complimentary molecular approaches in order to investigate the genetic basis of adaptation in natural populations of kokanee, the freshwater form of sockeye salmon, *Oncorhynchus nerka*. In Chapter 2, I found that genetic diversity at the salmonid circadian rhythm gene, *OtsClock1b*, is significantly correlated with photoperiod in kokanee populations sampled across a latitudinal transect from lower BC to Alaska. This pattern was not present among the neutral loci that were used as controls, suggesting that the association between *OtsClock1b* length variation and latitude may be the result of divergent natural selection across differing photoperiod regimes. Future research is needed in order to identify the *Clock*-mediated phenotype that is putatively under selection, and to experimentally link the phenotype(s) with differential fitness at across latitudes.

In Chapter 3, I used next-generation DNA sequencing technology to obtain pooled transcriptome sequences from the two reproductive ecotypes of kokanee in Okanagan Lake. Transcriptome data was used to identify SNPs that exhibited divergent allele frequencies between ecotypes and to identify genomic regions of differential gene expression between ecotypes from Okanagan Lake, BC. Sequence data from this Chapter represents a novel

genomic resource that can be used to design genotyping assays and as a scaffold for subsequent genomic research. Further, this study has identified a suite of genomic regions that may be differentially expressed between stream and shore-spawners, warranting future research.

In Chapter 4, I again used next-generation sequencing to obtain genomic data for 96 individually barcoded kokanee samples from Okanagan Lake, BC. The use of individual barcodes facilitates a 'genotyping-by-sequencing' approach in which genotypic data was collected for all individuals at ~6,000 SNPs, providing a fine-scale genomic tool for quantifying levels of genetic differentiation. Outlier detection revealed a suite of 20 loci that are putatively under divergent selection; these loci demonstrated significant genetic structure and had very high assignment accuracy to ecotype, making them promising candidates for the genetic basis of ecotype divergence in kokanee. Of particular interest, the results of this study suggest that divergent selection between reproductive ecotypes may be important at very early developmental stages before fry migrate away from their natal spawning grounds.

5.2 LIMITATIONS AND FUTURE DIRECTIONS

The absence of overt morphological differentiation between reproductive ecotypes of kokanee limits a top-down investigation of adaptation. Rather, this dissertation largely employs bottom-up genome scans in order to infer divergent selection through the identification of statistical outlier loci, which show locus-specific signatures of increased structure relative to the rest of the genome. The identification of a correlation between gene frequency and an environmental variable provides indirect evidence for natural selection, warranting future study (Endler 1986). I recommend further research aimed at testing for

phenotypic differences associated with early development based in the outlier loci identified in Chapter 4; the identification of these phenotypes would allow for experimental studies to directly link genotypic and phenotypic variation with fitness differences between spawning environments.

In addition, future research is planned that will use the genomic approach outlined in Chapter 4 to test for parallel patterns of divergence on samples from a variety of lakes across British Columbia. The identification of parallel patterns of evolution would provide further evidence for the presence of natural selection in this system. It is important for these studies to keep in mind that parallel evolution may affect the same genes, but in different ways. For example, it is possible that replicated lakes will also show a correlation between kokanee ecotype and genetic diversity in the *FOXl2*-like protein (see Chapter 4), yet differences may not be associated at the same SNP position as in Okanagan Lake; i.e. lack of parallel pattern at the same SNP locus does not negate the possibility that the gene region is subject to divergent selection. For example, three species of lizard (*Aspidoscelis inornata*, *Sceloporus undulates*, and *Holbrookia macultata*) demonstrate convergent evolution of a white phenotype in populations inhabiting White Sands, New Mexico; yet parallel evolution of the white phenotype among these species is the result of different mutations affecting the same gene (Rosenblum 2006; Manceau *et al*. 2010; Rosenblum *et al*. 2010; Rosenblum & Harmon 2011). Given that kokanee populations are polyphyletic, having evolved through multiple freshwater colonization events (Taylor *et al*. 1997), it is possible that convergent evolution in this system may also evolve different mutations that produce convergent phenotypes. Conversely, the occasional presence of shore-spawning behaviour in anadromous populations of sockeye, suggests that the basis of ecotype divergence in kokanee may be

associated with standing genetic variation in the ancestral populations (rather than novel mutations), and would therefore be similar among replicate lakes. Landscape-scale research designed to test these alternate hypotheses will provide key insight into ecotype divergence in kokanee.

5.3 SIGNIFICANCE

This study significantly advances our understanding of ecotype divergence in kokanee by providing evidence for a molecular basis underlying the observed differences in reproductive behavior. Further, the lack of genetic differentiation observed at neutral loci suggests that divergence is very shallow, only affecting a small number of functionally important loci. Given these patterns, I suggest that kokanee may be an underappreciated system for examining the genomic architecture underlying very early ecological divergence.

This study also provides data of significant utility to the conservation and management of kokanee stocks. The outliers identified in Chapter 4 outperform all previous loci used for genetic stock identification in Okanagan Lake (Taylor *et al*. 1997; Russello *et al*. 2012; Frazer & Russello 2013), providing a powerful tool for in-season monitoring of relative ecotype abundance. Multidisciplinary research that bridges the gap between genomics research and fisheries management is a promising avenue for the future of kokanee conservation in British Columbia.

Bibliography

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

Allendorf FW, Thorgaard G (1984) Polyploidy and the evolution of salmonid fishes. In: *The Evolutionary Genetics of Fishes* (ed. Turner BJ), pp. 1-53. Plenum Press, New York.

Anderson EC, Waples RS, Kalinowski ST (2008) An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* **65**, 1475-1486.

André C, Larsson LC, Laikre L *et al*. (2011) Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity* **106**, 270-280.

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomics regions early in speciation: incipient speciation in sunflower ecotypes. *Evolution* **67**, 2468-2482.

Askey PJ, Johnston NT (2013) Self-regulation of the Okanagan Lake kokanee recreational fishery: Dynamic angler effort response to varying fish abundance and productivity. *North American Journal of Fisheries Management* **33**, 926-939.

Baird NA, Etter PD, Atwood TS *et al*. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS One* **3**, e3376.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics* **12**, 767-780.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science* **322**, 255-257.

Beacham TD, Murray CB, Barner LW (1994) Influence of photoperiod on the timing of reproductive maturation in pink salmon (*Oncorhynchus gorbuscha*) and its application to genetic transfers between odd-year and even-year spawning populations. *Canadian Journal of Zoology* **72**, 826-833.

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

Bernatchez L, Renaut S, Whiteley AR *et al*. (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 1783-1800.

Bonin A, Nicole F, Pompanon F, Miaud C, Taberlet P (2007) Population adaptive index: A new method to help measure intraspecific genetic diversity and prioritize populations for conservation. *Conservation Biology* **21**, 697-708.

Boughman JW, Rundle HD, Schluter D (2005) Parallel evolution of sexual isolation in sticklebacks. *Evolution* **59**, 361-373.

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* **6**, 450-461.

Bradshaw WE, Holzapfel CM (2007) Evolution of animal photoperiodism. *Annual Review of Ecology Evolution and Systematics* **38**, 1-25.

Brown LL, Cox WT, Levine RP (1997) Evidence that the causal agent of bacterial cold-water disease *Flavobacterium psychrophilum* is transmitted within salmonid eggs. *Diseases of Aquatic Organisms* **29**, 213-218.

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* **20**, 1004-1010.

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.

Campbell NR, Narum SR (2011) Development of 54 novel single-nucleotide polymorphism (SNP) assays for sockeye and coho salmon and assessment of available SNPs to differentiate stocks within the Columbia River. *Molecular Ecology Resources* **11**, 20-30.

Caprioli M, Ambrosini R, Boncoraglio G *et al*. (2012) Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. *PLoS One* **7**, e35140.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

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.

Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **164**, 362-379.

Chong KL, Wang S, Melamed P (2004) Isolation and characterization of the follicle-stimulating hormone beta subunit gene and 5' flanking region of the Chinook salmon. *Neuroendocrinology* **80**, 158-170.

Clarke WC, Withler RE, Shelbourn JE (1994) Inheritance of smolting phenotypes in backcrosses of hybrid stream-type x ocean type Chinook salmon (*Oncorhynchus tshawytscha*). *Estuaries* **17**, 13-25.

Colosimo PF, Hosemann KE, Balabhadra S *et al*. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**, 1928-1933.

Conesa A, Gotz S, Garcia-Gomez JM *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676.

Coyne JA (1992) Genetics and speciation. *Nature* **355**, 511-515.

Coyne JA (1994) Ernst Mayr and the origin of species. *Evolution* **48**, 19-30.

Coyne JA, Orr HA (2004) *Speciation* Sinauer Associates, Sunderland, MA.

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.

Darwin C (1859) *The origin of species* Oxford University Press.

Davey JW, Cezard T, Fuentes-Utrilla P *et al*. (2012) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology* **22**, 3151-3164.

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.

Dill P (1996) A study of shore-spawning kokanee salmon (*Oncorhynchus nerka*) at Bertram Creek Park, Okanagan Lake, BC 1992-1996. *Report prepared for the BC Ministry of Environment, Penticton, BC*.

Dionne M, Miller KM, Dodson JJ, Caron F, Bernatchez L (2007) Clinal variation in MHC diversity with temperature: Evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution* **61**, 2154-2164.

Dor R, Cooper CB, Lovette IJ *et al*. (2011) Clock gene variation in *Tachycineta* swallows. *Ecology and Evolution* **2**, 95-105.

Duchaud E, Boussaha M, Loux V *et al*. (2007) Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. *Nature Biotechnology* **25**, 763-769.

Earl DA, Vonholdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Egan SP, Nosil P, Funk DJ (2008) Selection and genomic differentiation during ecological speciation: Isolating the contributions of host association via a comparative genome scan of *Neochlamisus bebbianae* leaf beetles. *Evolution* **62**, 1162-1181.

Elfstrom CM, Smith CT, Seeb JE (2006) Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. *Molecular Ecology Notes* **6**, 1255.

Endler JA (1986) *Natural Selection in the Wild* Princeton University Press, Princeton, New Jersey.

Etter P, Bassham S, Hohenlohe P, Johnson E, Cresko W (2011) SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds. Orgogozo V, andRockman MV), pp. 157-178.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

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.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics* **28**, 342-350.

Fidler A, Johnsen A, Carter K, Kempenaers B (2006) Avian clock gene polymorphism: Evidence for occurrence and latitudinal clines in allele frequencies. *Journal of Ornithology* **147**, 166-166.

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.

Frazer KK, Russello MA (2013) Lack of parallel genetic patterns underlying the repeated ecological divergence of beach and stream spawning kokanee salmon. *Journal of Evolutionary Biology* **12**, 2606-2621.

Freamo H, O'Reilly P, Berg PR, Lien S, Boulding EG (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources* **11**, 254-267.

Freeland JR, Biss P, Conrad KF, Silvertown J (2010) Selection pressures have caused genome-wide population differentiation of *Anthoxanthum odoratum* despite the potential for high gene flow. *Journal of Evolutionary Biology* **23**, 776-782.

Funk CW, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**, 489-496.

Galindo J, Grahame JW, Butlin RK (2010) An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. *Journal of Evolutionary Biology* **23**, 2004-2016.

Garcia de Leaniz C, Fleming IA, Einum S *et al*. (2007) A critical review of adaptive genetic variation in Atlantic salmon: implications for conservation. *Biological Reviews* **82**, 173-211.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**, 915-934.

Godbout L, Wood CC, Withler RE *et al*. (2011) Sockeye salmon (*Oncorhynchus nerka*) return after an absence of nearly 90 years: a case of reversion to anadromy. *Canadian Journal of Fisheries and Aquatic Sciences* **68**, 1590-1602.

Gotz S, Garcia-Gomez JM, Terol J *et al*. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**, 3420-3435.

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* **86**, 485-486.

Gould SJ, Lewontin RC (1979) Spandrels of San-Marco and the Panglossian paradigm: A critique of the adaptionist program. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **205**, 581-598.

Groot C, Margolis L (1991) *Pacific salmon life histories* University of British Columbia Press, Vancouver, BC.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.

Habicht C, Seeb LW, Myers KW, Farley EV, Seeb JE (2010) Summer-Fall Distribution of Stocks of Immature Sockeye Salmon in the Bering Sea as Revealed by Single-Nucleotide Polymorphisms. *Transactions of the American Fisheries Society* **139**, 1171-1191.

Harmer SL, Panda S, Kay SA (2001) Molecular bases of circadian rhythms. *Annual Review of Cell and Developmental Biology* **17**, 215-253.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al*. (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 123-136.

Hendry AP (2009) Ecological speciation! Or the lack thereof? *Canadian Journal of Fisheries and Aquatic Sciences* **66**, 1383-1398.

Hendry AP, Wenburg JK, Bentzen P, Volk EC, Quinn TP (2000) Rapid evolution of reproductive isolation in the wild: Evidence from introduced salmon. *Science* **290**, 516-518.

Hohenlohe PA, Bassham S, Etter PD *et al*. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6**, e1000862.

Hut RA, Paolucci S, Dor R, Kyriacou CP, Daan S (in press) Latitudinal clines: an evolutionary view on biological rhythms. *Proceedings of the Royal Society Series B-Biological Sciences*.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis spp*., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology* **19**, 5389-5403.

Johnsen A, Fidler AE, Kuhn S *et al*. (2007) Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. *Molecular Ecology* **16**, 4867-4880.

Kalinowski ST, Manlove KR, Taper ML (2007) ONCOR: a computer program for genetic stock identification, Department of Ecology, Montana State University, Bozeman, MT. Available at http://www.montana.edu/kalinowski/kalinowski_software.htm.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* **7**, 1225-1241.

Kitano T, Matsuoka N, Saitou N (1997) Phylogenetic relationship of the genus *Oncorhynchus* species inferred from nuclear and mitochindrial markers. *Genes & Genetic Systems* **72**, 25-34.

Koop BF, von Schalburg KR, Leong J *et al*. (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* **9**, 545.

Kyriacou CP, Peixoto AA, Sandrelli F, Costa R, Tauber E (2008) Clines in clock genes: fine-tuning circadian rhythms to the environment. *Trends in Genetics* **24**, 124-132.

Leder EH, Danzmann RG, Ferguson MM (2006) The candidate gene, Clock, localizes to a strong spawning time quantitative trait locus region in rainbow trout. *Journal of Heredity* **97**, 74-80.

Lemay MA, Donnelly DJ, Russello MA (2013) Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon. *BMC Genomics* **14**, 308.

Lemay MA, Russello MA (2012) Neutral loci reveal structure by geography, not ecotype, in Kootenay Lake kokanee. *North American Journal of Fisheries Management* **32**, 282-291.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175-195.

Liedvogel M, Szulkin M, Knowles SCL, Wood MJ, Sheldon BC (2009) Phenotypic correlates of Clock gene variation in a wild blue tit population: evidence for a role in seasonal timing of reproduction. *Molecular Ecology* **18**, 2444-2456.

Liew M, Pryor R, Palais R *et al*. (2004) Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clinical Chemistry* **50**, 1156-1164.

Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Madetoja J, Nyman P, Wiklund T (2000) *Flavobacterium psychrophilum*, invasion into and shedding by rainbow trout *Oncorhynchus mykiss*. *Diseases of Aquatic Organisms* **43**, 27-38.

Madetoja J, Nystedt S, Wiklund T (2003) Survival and virulence of *Flavobacterium psychrophilum* in water microcosms. *FEMS Microbiology Ecology* **43**, 217-223.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 2439-2450.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al*. (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research* **23**, 1817-1828.

Martino A, Mancuso T, Rossi AM (2010) Application of High-Resolution Melting to Large-Scale, High-Throughput SNP Genotyping: A Comparison with the TaqMan(R) Method. *Journal of Biomolecular Screening* **15**, 623-629.

Mayr E (1942) *Systematics and the origin of species*. Columbia University Press, New York.

Mayr E (1947) Ecological factors in speciation. *Evolution* **1**, 263-288.

McGurk MD (2000) Comparison of fecundity-length-latitude relationships between nonanadromous (kokanee) and anadromous sockeye salmon (*Oncorhynchus nerka*). *Canadian Journal of Zoology* **78**, 1791-1805.

McKinnon JS, Mori S, Blackman BK *et al*. (2004) Evidence for ecology's role in speciation. *Nature* **429**, 294-298.

Miller KM, Kaukinen KH, Beacham TD, Withler RE (2001) Geographical heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica* **111**, 237-257.

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.

Murata S, Takasaki N, Saitoh M, Okada N (1993) Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 6995-6999.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841-2847.

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

Nematollahi A, Decostere A, Pasmans F, Haesebrouck F (2003) *Flavobacterium psychrophilum* infections in salmonid fish. *Journal of Fish Diseases* **26**, 563-574.

Ng HH, Feng Q, Wang HB *et al*. (2002) Lysine methylation within the globular domain of histone H3 by *Dot1* is important for telomeric silencing and *Sir* protein association. *Genes & Development* **16**, 1518-1527.

Nguyen AT, Zhang Y (2011) The diverse functions of *Dot1* and *H3K79* methylation. *Genes & Development* **25**, 1345-1358.

Nichol ST, Rowe JE, Winton JR (1995) Molecular epizootiology and evolution of the glycoprotein and non-virion protein genes of the infectious hematopoietic necrosis virus, a fish rhabdovirus. *Virus Research* **38**, 159-173.

Nicolas P, Mondot S, Achaz G *et al*. (2008) Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. *Applied and Environmental Microbiology* **74**, 3702-3709.

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641-647.

Nosil P (2008) Ernst Mayr and the integration of geographic and ecological factors in speciation. *Biological Journal of the Linnean Society* **95**, 26-46.

Nosil P (2012) *Ecological Speciation* Oxford University Press.

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

O'Brien C, Bradshaw WE, Holzapfel CM (2010) Testing for causality in covarying traits: genes and latitude in a molecular world. *Molecular Ecology* **20**, 2471-2476.

O'Malley KG, Banks MA (2008) A latitudinal cline in the Chinook salmon (*Oncorhynchus tshawytscha*) Clock gene: evidence for selection on PolyQ length variants. *Proceedings of the Royal Society Series B-Biological Sciences* **275**, 2813-2821.

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.

O'Malley KG, Ford MJ, Hard JJ (2010a) Clock polymorphism in Pacific salmon: evidence for variable selection along a latitudinal gradient. *Proceedings of the Royal Society Series B-Biological Sciences* **277**, 3703-3714.

O'Malley KG, McClelland EK, Naish KA (2010b) Clock genes localize to quantitative trait loci for stage-specific growth in juvenile coho salmon, *Oncorhynchus kisutch*. *Journal of Heredity* **101**, 628-632.

Olsen JB, Wenburg JK, Bentzen J (1996) Semi-automated multilocus genotyping of Pacific salmon (*Oncorhynchus spp*.) using microsatellites. *Molecular Marine Biology and Biotechnology* **5**, 259-272.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**, 177-187.

Pearcy WG (1992) *Ocean ecology of North Pacific salmonids* University of Washington Press, Seattle, WA.

Peichel CL, Nereng KS, Ohgi KA *et al*. (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901-905.

Percell MK, Getchell RG, McClure CA, Garver KA (2011) Quantitative polymerase chain reaction (PCR) for detection of aquatic animal pathogens in a diagnostic laboratory setting. *Journal of Aquatic Animal Health* **23**, 148-161.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* **7**, e37135.

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

Pritchard JK, Wen X, Falush D (2007) Documentation for *structure* software: Version 2.3. *Available from: http://pritch.bsd.uchicago.edu/software*.

Prunier J, Gerardi S, Laroche J, Beaulieu J, Bousquet J (2012) Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. *Molecular Ecology* **21**, 4270-4286.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Renaut S, Grassa CJ, Yeaman S *et al*. (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications* **4**, 8.

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.

Rice WR (1989) Analyzing tables of statistical tests. *Evolution* **43**, 223-225.

Rosenblum EB (2006) Convergent evolution and divergent selection: Lizards at the White Sands ecotone. *American Naturalist* **167**, 1-15.

Rosenblum EB, Harmon LJ (2011) "Same same but different": Replicated ecological speciation at White Sands. *Evolution* **65**, 946-960.

Rosenblum EB, Rompler H, Schoneberg T, Hoekstra HE (2010) Molecular and functional basis of phenotypic convergence in white lizards at White Sands. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2113-2117.

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

Royce WF, Smith LS, Hartt AC (1968) Models of oceanic migrations of Pacific salmon and comments on guidance mechanisms. *United States Fish and Wildlife Service Fishery Bulletin* **66**, 441-462.

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods for Molecular Biology* (eds. Krawetz S, andMisener S), pp. 365-386. Humana Press, Totowa, NJ.

Rubio-Godoy M, Paladini G, Freeman MA, Garcia-Vasquez A, Shinn AP (2012) Morphological and molecular characterisation of *Gyrodactylus salmonis* (Platyhelminthes, Monogenea) isolates collected in Mexico from rainbow trout (*Oncorhynchus mykiss* Walbaum). *Veterinary Parasitology* **186**, 289-300.

Rundle HD, Nagel L, Boughman JW, Schluter D (2000) Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306-308.

Russello MA, Kirk SL, Frazer K, Askey P (2012) Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications* **5**, 39-52.

Schluter D (1995) Adaptive radiation in stickelbacks: Trade-offs in feeding performance and growth. *Ecology* **76**, 82-90.

Schluter D (1996a) Ecological causes of adaptive radiation. *American Naturalist* **148**, S40-S64.

Schluter D (1996b) Ecological speciation in postglacial fishes. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **351**, 807-814.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution* **16**, 372-380.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science* **323**, 737-741.

Schluter D, Marchinko KB, Barrett RDH, Rogers SM (2010) Natural selection and the genetics of adaptation in threespine stickleback. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 2479-2486.

Schluter D, McPhail JD (1992) Ecological character displacement and speciation in stickelbacks. *American Naturalist* **140**, 85-108.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233-234.

Scribner KT, Gust JR, Fields RL (1996) Isolation and characterization of novel salmon microsatellite loci: Cross-species amplification and population genetic applications. *Canadian Journal of Fisheries and Aquatic Sciences* **53**, 833-841.

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.

Seehausen O, Butlin RK, Keller I *et al*. (2014) Genomic and the origin of species. *Nature Reviews Genetics* **15**, 176-192.

Shepherd BG (2000) A case history: the kokanee stocks of Okanagan Lake, 609-616.

Singer MS, Kahana A, Wolf AJ *et al*. (1998) Identification of high-copy disrupters of telomeric silencing in *Saccharomyces cerevisiae*. *Genetics* **150**, 613-632.

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

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**, 671-688.

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

Taylor EB, Foote CJ, Wood CC (1996) Molecular genetic evidence for parallel life-history evolution within a Pacific salmon (sockeye salmon and kokanee, *Oncorhynchus nerka*). *Evolution* **50**, 401-416.

Taylor EB, Harvey S, Pollard S, Volpe J (1997) Postglacial genetic differentiation of reproductive ecotypes of kokanee *Oncorhynchus nerka* in Okanagan Lake, British Columbia. *Molecular Ecology* **6**, 503-517.

Taylor EB, Kuiper A, Troffe PM, Hoysak D, Pollard S (2000) Variation in developmental biology and microsatellite DNA in reproductive ecotypes of kokanee, *Oncorhynchus nerka*: implications for declining populations in a large British Columbia lake. *Conservation Genetics* **1**, 231-249.

Taylor JF, Migaud H, Porter MJR, Bromage NR (2005) Photoperiod influences growth rate and plasma insulin-like growth factor-I levels in juvenile rainbow trout, *Oncorhynchus mykiss*. *General and Comparative Endocrinology* **142**, 169-185.

Thompson LC (1999) *Abundance and production of zooplankton and kokanee salmon (Oncorhynchus nerka) in Kootenay Lake, British Columbia during artificial fertilization* PhD Thesis, University of British Columbia.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gamiae*. *PLoS Biology* **3**, e285.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Van West P (2006) *Saprolegnia parasitica*, an oomycete pathogen with a fishy appetite: new challenges for an old problem. *Mycologist* **20**, 99-104.

von Schalburg KR, Yasuike M, Yazawa R *et al*. (2011) Regulation and expression of sexual differentiation factors in embryonic and extragonadal tissues of Atlantic salmon. *BMC Genomics* **12**, 31.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63.

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

Westgaard JI, Fevolden SE (2007) Atlantic cod (*Gadus morhua L.*) in inner and outer coastal zones of northern Norway display divergent genetic signature at non-neutral loci. *Fisheries Research* **85**, 306-315.

Winans GA, Pollard S, Kuligowski DR (2003) Two reproductive life history types of kokanee, *Onchorynchus nerka*, exhibit multivariate morphometric and protein genetic differentiation. *Environmental Biology of Fishes* **67**, 87-100.

Wolf JBW, Lindell J, Backstrom N (2010) Speciation genetics: current status and evolving approaches. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 1717-1733.

Wolfe KH, Li WH (2003) Molecular evolution meets the genomics revolution. *Nature Genetics* **33**, 255-265.

Wood CC, Bickham JW, Nelson RJ, Foote CJ, Patton JC (2008) Recurrent evolution of life history ecotypes in sockeye salmon: implications for conservation and future evolution. *Evolutionary Applications* **1**, 207-221.

Wright JJ, Lubieniecki KP, Park JW *et al*. (2008) Sixteen Type 1 polymorphic microsatellite markers from Chinook salmon (*Oncorhynchus tshawytscha*) expressed sequence tags. *Animal Genetics* **39**, 84-85.

Appendices

APPENDIX A: Diversity of the bacterial pathogen, *Flavobacterium sp*., infecting reproductive ecotypes of kokanee salmon.

Pathogen resistance can be a powerful driver of adaptation. This process occurs when variation in pathogen diversity over small spatial or temporal scales imposes divergent selection on populations of their host species (Fraser et al. 2011). In salmonids, for example, genetic diversity associated with the major histocompatability complex can vary at micro-geographical scales (Miller *et al*. 2001; Fraser *et al*. 2011), reflecting adaptation in response to heterogeneous pathogen regimes (Dionne et al. 2007).

Kokanee, the freshwater form of sockeye salmon, *Oncorhynchus nerka* (Walbaum), can occur as two reproductive ecotypes, which differ in their spawning habitats (streams vs. beaches). Previous research found that the abundance of cDNA from several pathogens (bacterial, fungal, and parasitic) was greater in transcriptome-sequenced samples from the stream ecotype compared with the beach ecotype in Okanagan Lake, British Columbia (Lemay *et al*. 2013). It was hypothesized that asymmetrical pathogen infection between spawning environments could confer divergent selection between ecotypes. However, the lack of information on pathogen diversity between spawning habitats has precluded a direct test of this hypothesis.

In this study we measured the diversity and abundance of *Flavobacteriu*m *sp*. infecting kokanee from the two spawning habitats (streams and beaches) in Okanagan Lake, BC. *Flavobacterium* was chosen because its abundance was highly correlated with ecotype

in Lemay *et al*. (2013), and because several *Flavobacterium* species are associated with high levels of salmonid mortality (Nematollahi et al. 2003).

To examine the diversity of *Flavobacterium* present in Okanagan Lake kokanee, DNA was extracted from muscle tissue of one adult from each of four stream and three beach-spawning locations (see Lemay *et al*. 2013). *Flavobacterium sp*. DNA was isolated using primers designed to amplify a ~200 nucleotide region of the gene encoding the ATP synthase alpha subunit (*AtpA*) of all species within the genus *Flavobacterium* [*Fspp1_F*: 5'-TTRTTAAGAAGACCACCRGG-3', *Fspp1_R*: 5'- GGRATATATGCAGAAACGTCACC-3']. This region was chosen because *AtpA* it is part of a panel of genes used for strain identification in the species *Flavobacterium psychrophilum* (Bernardet & Grimont) Bernardet (Nicolas et al. 2008), allowing comparisons with previously published sequence data.

Each PCR contained 2μl DNA, 2.5μl 10X PCR buffer, 2.5μl dNTPs (2mM), 1.0μl forward primer (10μM), 1.0μl reverse primer (10μM), and Ampli*Taq* Gold polymerase (0.5 units, Applied Biosystems) in a 25μl total volume. Touchdown PCR was used with initial denaturation of 94˚C for 10 minutes, then 10 cycles at 94˚C for 30 seconds, 60˚C for 30 seconds, 72˚C for 60 seconds, with the annealing step decreasing by 1˚C per cycle to 50˚C. The annealing temperature was maintained at 50˚C for an additional 30 cycles, followed by extension at 72˚C for 2 minutes. All PCR products were purified using a Qiagen MinElute kit, diluted to 4ng/μl, and ligated overnight at 4˚C using the pGEM®-T Easy Vector System (Promega). Transformed cells were added to plates containing 100µg/ml ampicillin, 0.5mM IPTG, and 80µg/ml X-Gal, and incubated for 16-20 hours at 37˚C. White colonies were then boiled at 100˚C for 10 minutes in 100μl of TE buffer. We amplified cloned inserts from 100

colonies (50 from each ecotype) using *T7* and *Sp6* primers (Promega). Each PCR contained 1μl colony boil, 1.25μl 10X PCR buffer, 1.25μl dNTPs (2mM), 0.5μl each primer (10μM), and KAPA*Taq* polymerase (0.5 units; KAPA Biosystems) in a 13.5μl total volume. Each PCR had an initial denaturation of 94˚C for 2 minutes, followed by 35 cycles at 94˚C for 30 seconds, 50˚C for 30 seconds, 72˚C for 30 seconds, with a final extension at 72˚C for 2 minutes. Sequencing was carried out in one direction using *Sp6* on an Applied Biosystems 3130XL.

A phylogenetic approach was used to infer species of origin among the 100 colonies that were sequenced. Unique haplotypes from each ecotype were aligned with the *AtpA* region from the published genomes of nine *Flavobacterium* species and used to generate a neighbor-joining tree in Geneious v.6.1 (Biomaters); 100 Bootstrap replicates were performed with a 50% support threshold (Figure A1).

We identified 13 unique haplotypes among the stream samples and nine unique haplotypes among beach samples, with only two haplotypes shared between spawning habitats. Haplotypes occurring in the same clade as *F. psychrophilum* were the most prevalent, occurring at all sampling locations and accounting for 64% of retained haplotypes. This analysis provides evidence that multiple *Flavobacterium* species may be present in Okanagan Lake, however the small size of the amplicon precludes determination of species identity.

Quantitative PCR (qPCR) was then used to determine the relative abundance of *Flavobacterium sp*. in kokanee from the same seven locations. DNA was extracted from kokanee operculum tissue sampled in 2007 (n = 48) and 2010 (n = 48) (see Russello et al. 2012). The total quantity of extracted DNA (fish and pathogen) was determined for each

sample using the Quant-iT™ Pico Green ds DNA Assay Kit (Invitrogen) run on a ViiA7 qPCR machine (Life Technologies). Quantitative PCR was carried out using *AtpA* primers described above to quantify the pathogen component of each DNA sample. For the standard curve, we used *F. psycrophilum* DNA of known strain and concentration [strain: CIP103534(T)]; three replicates of each concentration in the standard curve were used. Each qPCR contained 1μl DNA template, 0.5μl *Fspp1_F* forward primer (1μM), 0.5μl *Fspp2_R* reverse primer (10μM), and 5.0μl Fast SYBR® Green Master Mix (Applied Biosystems) in a 10μl total volume. The two-step PCR protocol had an initial denaturation of 94˚C for 2 minutes, followed by 55 cycles at 94˚C for 30 seconds and 60˚C for 30 seconds, followed by a melt curve beginning at 60˚C and increasing to 98˚C.

For each individual, the inferred quantity of *Flavobacterium* amplicons was normalized to the DNA template concentration in order to derive a measure of pathogen infection per unit of kokanee DNA (Percell et al. 2011). Statistical comparisons using non-parametric Kruskal-Wallis tests found no difference in *Flavobacterium* abundance between kokanee collected from the two different spawning habitats (2010 $p = 0.211$; 2007 $p = 0.204$; Figure A2).

In summary, we found that *Flavobacterium sp*. was present at all seven sites tested in Okanagan Lake. We observed slightly greater diversity of *Flavobacterium AtpA* haplotypes in the stream-spawning habitat, but no significant differences in *Flavobacterium* abundance between habitats.

While these data provide preliminary diversity estimates for *Flavobacterium*, it would be informative for future research to sequence additional genes in order to determine the identity of species present. It may also be useful for future research to examine

Flavobacterium diversity and abundance from environmental samples in the two habitats; while *F. psychrophilum* is both horizontally and vertically transmitted to new hosts (Brown *et al*. 1997; Madetoja *et al*. 2000), it can also persist for long periods of time (300 days) without a host (Madetoja *et al*. 2003). Therefore, the compliment of pathogens infecting fish tissues may not be an accurate estimate of the total diversity and abundance present in each habitat. Despite the limitations of this study, the identification of *Flavobacterium sp*. in Okanagan Lake suggests that hatchery managers should be cautious when collecting wild kokanee for captive spawning programs.

Figure A:1 Unrooted neighbor-joining tree showing the relationship between the unique *Flavobacterium* haplotypes amplified from each ecotype (n = 13 stream, n = 8 beach) and nine previously published reference sequences (indicated with bold font). Genbank accession numbers are included next to species names. Branch labels are bootstrap support values (%).

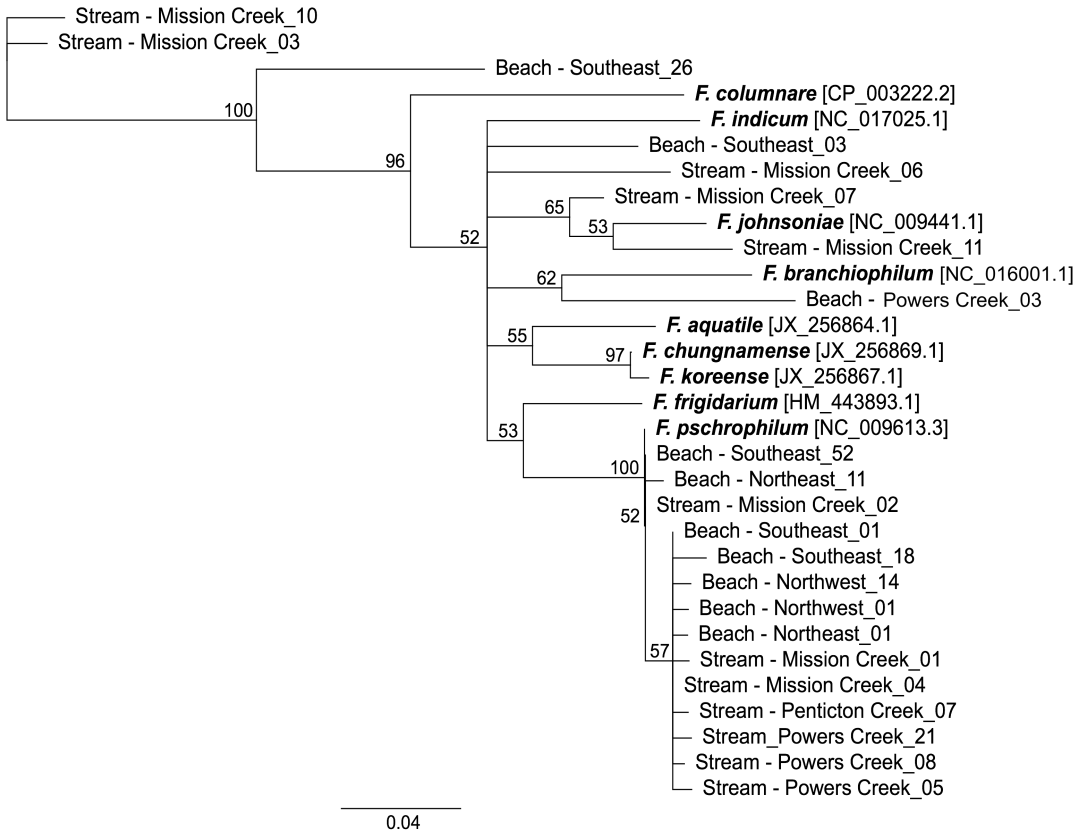
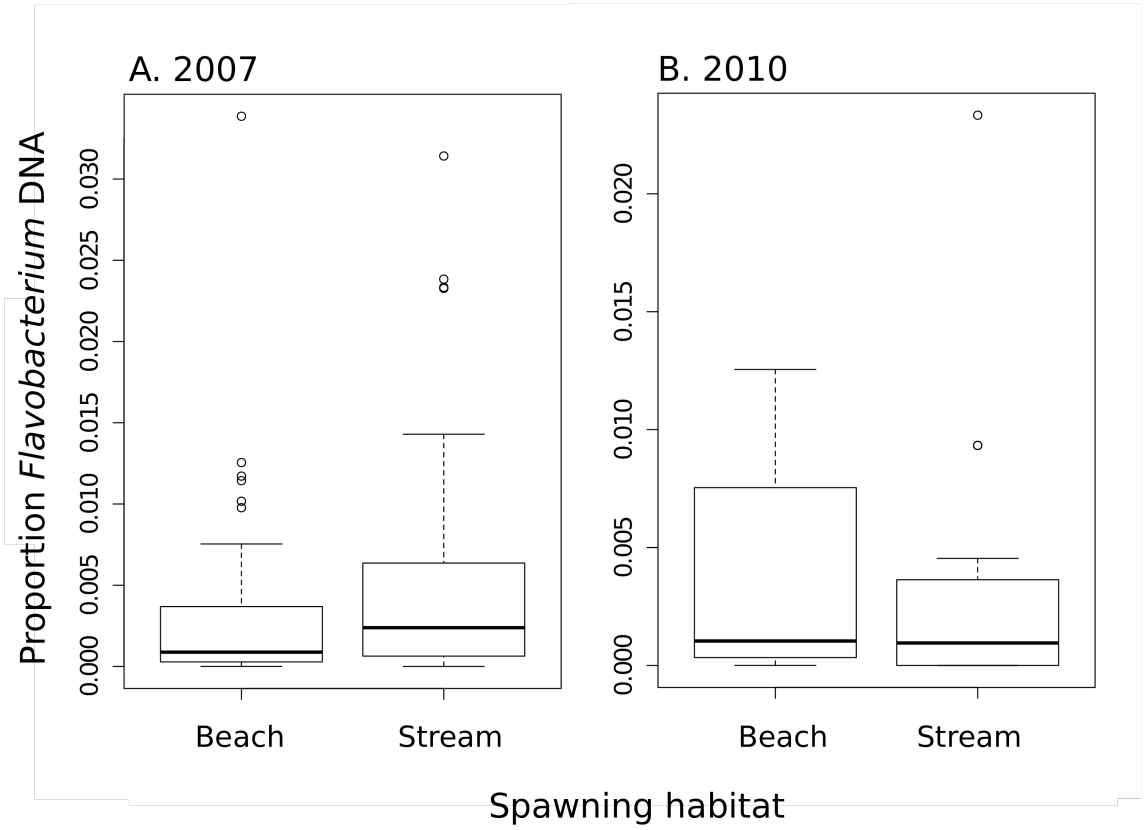
Figure A:2 Relative proportion of *Flavobacterium sp*. DNA amplified from the operculum tissue of kokanee sampled at stream and beach spawning sites in Okanagan Lake, BC. Samples are separated by year: (A) 2007 (n = 48) and (B) 2010 (n = 48).