Genomic and Behavioral Diversity in Rainbow Trout

By

KRISTEN A. CHRISTENSEN

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Molecular Biosciences

December 2013
To the Faculty of Washington State University:

The members of the committee appointed to examine the dissertation of KRISTEN A. CHRISTENSEN find it satisfactory and recommend that it be accepted.

______________________________
Gary H. Thorgaard, Ph.D., Chair

______________________________
Ruth C. Newberry, Ph.D.

______________________________
Eric A. Shelden, Ph.D.

______________________________
John J. Wyrick, Ph.D.

______________________________
B. Dan Rodgers, Ph.D.
ACKNOWLEDGEMENTS

Gary, you are an amazing advisor and after all your help I wish you my deepest thanks. I really enjoyed our fishing trips and I appreciated your nudges in the right direction. Bobbi and Anna, you are my best friends and your encouragement helped me through graduate school. Stacy and Adam, thanks for introducing me to climbing, it helped me stay relaxed in stressful times. I would like to thank Jennifer Geiger for her guidance and help getting me into graduate school. Dr. Ping Ye introduced me to the wonderful world of programming and I still love exploring the use of global variables. To my committee members, thanks for listening to my seminars and reading all these documents. Other notable figures in the progress along my graduate school path include: Kelly McGovern, Paul Wheeler, Joe Brunelli, Julie Stanton, Patty Brandt, and Lisa Gloss. Thank you all for your efforts to help me and for being the people that you are.
Genomic and Behavioral Diversity in Rainbow Trout

ABSTRACT
By Kristen A. Christensen Ph.D.
Washington State University
December 2013

Chair: Gary H. Thorgaard

Genomic diversity between hatchery and wild salmonids is thought to underlie differential survival rates in the wild. Decreased survival may result from a hatchery fish's inability to properly respond to a predator. In this dissertation, behavioral and physiological responses to a predator are reviewed, followed by an analysis of behavioral differences between clonal rainbow trout with wild and hatchery origins. Genomic diversity of single nucleotide polymorphisms (SNPs) in rainbow trout and evidence at the sequence level for ancestral genome duplications are analyzed and discussed.

Behavioral differences were observed between hatchery and wild-derived clonal lines of rainbow trout at multiple time points after swim-up. Genetic associations with behavioral characteristics were determined through Quantitative Trait Loci (QTL) analysis and loci associated with the behaviors were mapped. Some of the identified loci had similar map locations and potentially may influence boldness in rainbow trout.

Thousands of possible single nucleotide polymorphisms (SNPs) were identified among five different clonal lines of rainbow trout using a novel transcriptome analysis. This method was developed for organisms with duplicated genomes to avoid mistaking paralogous sequence variants for allelic variation (i.e. SNPs) and was demonstrated to be superior to previously published strategies of SNP discovery in polyploids.
In a final study, gene copy number variance was examined in rainbow trout and other organisms.

Evidence for previously reported genome duplications in a variety of organisms was found using a new method of transcriptome analysis. This strategy offers advantages in cost and efficiency over existing methods of identifying whole genome duplications.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................................................................................. iii

ABSTRACT ........................................................................................................................................ iv

LIST OF TABLES ................................................................................................................................. ix

LIST OF FIGURES .............................................................................................................................. x

DEDICATION ........................................................................................................................................ xi

CHAPTER ONE .................................................................................................................................... 1

INTRODUCTION ................................................................................................................................. 1

Hatchery influences on behavior .................................................................................................... 1

Genome duplication in salmonids .................................................................................................... 5

References .......................................................................................................................................... 9

CHAPTER TWO ................................................................................................................................... 13

ANTIPREDATOR BEHAVIOR QTL: DIFFERENCES IN RAINBOW TROUT CLONAL LINES DERIVED
FROM WILD AND HATCHERY POPULATIONS ............................................................................. 13

Abstract ........................................................................................................................................ 13

Introduction .................................................................................................................................... 13

Materials and Methods .................................................................................................................. 16

*Experimental Fish* ....................................................................................................................... 16

*Filming strategy and processing (Phenotyping)* ......................................................................... 17

*Genotyping* ................................................................................................................................. 19

Results ............................................................................................................................................ 21

*Clonal Line Behavior* .................................................................................................................. 21

*Doubled Haploid Panels (QTL results)* .................................................................................... 23

Discussion ...................................................................................................................................... 24

Acknowledgements ....................................................................................................................... 26

References ...................................................................................................................................... 27

Tables ............................................................................................................................................. 31

Figures ........................................................................................................................................... 32

CHAPTER THREE .............................................................................................................................. 40

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS FROM THE TRANSCRIPTOME OF AN
ORGANISM WITH A WHOLE GENOME DUPLICATION .................................................................. 40
Acknowledgements ........................................................................................................... 80
References .......................................................................................................................... 81
Figures ................................................................................................................................ 84

CHAPTER 5 .......................................................................................................................... 89
CONCLUSIONS AND FUTURE DIRECTIONS ................................................................ 89
References .......................................................................................................................... 95
LIST OF TABLES

Table 2.1 Dates of Fertilization and Trials of Rainbow Trout Clonal Lines and Doubled Haploids............ 31

Table 2.2 Trial 1 Lengths and Weights of Clonal Lines of Rainbow Trout .................................................. 31

Table 3.1 Counts of cDNA reads from embryo and head kidney tissues from five rainbow trout lines. ... 68

Table 3.2 Gene duplication counts and SNPs for rainbow trout embryo and head kidney tissues .......... 68

Table 3.3 A comparison of SNPs between clonal rainbow trout lines for the combined tissues.......... 68

Table 3.4 SNP validation comparison using a transcriptome dataset from polyploid wheat. ................. 69
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Camera and Tank Setup</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Heatmap of a Single Fish (clonal Arlee - YY) During the Four Different Stimuli</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparing Experiments from Different Times</td>
<td>35</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparing Clonal Lines for Three Different Trials</td>
<td>36</td>
</tr>
<tr>
<td>2.5</td>
<td>A Histogram of the Doubled Haploid Panels Percent Cover Use</td>
<td>37</td>
</tr>
<tr>
<td>2.6</td>
<td>QTL Results Whale Rock x Arlee Panel</td>
<td>38</td>
</tr>
<tr>
<td>2.7</td>
<td>QTL Results OSU x Swanson Panel</td>
<td>39</td>
</tr>
<tr>
<td>3.1</td>
<td>A Comparison Between Traditional Approaches and Our Strategy of SNP Identification</td>
<td>64</td>
</tr>
<tr>
<td>3.2</td>
<td>UniGene Copy Number Distribution in Rainbow Trout Embryo and Head Kidney Tissues</td>
<td>65</td>
</tr>
<tr>
<td>3.3</td>
<td>Alignment Comparison</td>
<td>66</td>
</tr>
<tr>
<td>3.4</td>
<td>Effect of Window Size on Distinguishing Paralogs and the Amount of Useable Data</td>
<td>67</td>
</tr>
<tr>
<td>4.1</td>
<td>Illustration of Methodology</td>
<td>85</td>
</tr>
<tr>
<td>4.2</td>
<td>Relationship Between Alignment Number and Genes used in Analysis</td>
<td>86</td>
</tr>
<tr>
<td>4.3</td>
<td>Gene Duplications in a Variety of Organisms</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Distribution of Sequence Identity for Select Organisms</td>
<td>88</td>
</tr>
</tbody>
</table>
DEDICATION

To Lorene and Marty Christensen
CHAPTER ONE

INTRODUCTION

This dissertation addresses two major areas that are of high current relevance to the genetics of trout and salmon: the genetics of the domestication process, and accurate interpretation of variation in expressed genes (transcriptomes) originating from both allelic variation and from ancestral gene duplication. Chapter Two relates to domestication genetics, Chapter Three deals with detection of allelic variation in transcriptome datasets and Chapter Four deals with detection of gene duplications in transcriptome datasets.

Hatchery influences on behavior

Since the early 1900s, when land-use and over-fishing depleted once large fisheries (Meengs and Lackey 2005), there has been concern for the future of Pacific salmon and sea-run trout (WDFW). During the 1930s, hatchery supplementation and technical innovation were seen as viable options for maintaining healthy salmon populations (Nehlsen et al. 1991). In the 1990s, a particularly poor time for ocean-going fishes due to unfavorable ocean conditions (as low as 3-6% of the historical run of coho salmon, Meengs and Lackey 2005), the impact of dams, domestication from hatcheries, and deteriorating habitat conditions was appreciated, with the majority of Endangered Species Act listings being filed during this period for Pacific salmon and trout (NOAA, BPA). Nehlsen et al. (1991) concluded that over 200 Pacific salmon and trout stocks were susceptible to extinction.

Artificial propagation of Pacific salmon and trout began in the 1900s (National Archives, WDFW) to remedy these problems and continues on a large scale today (WDFW). The economic benefit of these supplementation programs may reach up to a billion dollars for a state (WDFW) and likely makes this investment in hatchery programs worthwhile. Negative impacts on wild salmon and trout populations may be the cost of this course.
As early as 1949, there has been evidence that wild fish survive better in the wild than hatchery-reared fish (Needham 1949). Chilcote et al. (2011) found that as the levels of hatchery fish (steelhead, coho, and Chinook) increased in a wild population, the level of reproductive success decreased. These findings are in line with another study showing that the number of generations a fish population spent in a hatchery correlated to a decrease in reproductive ability in the wild (Araki et al. 2007).

Why the survival rate appears to be lower in hatchery fish has been the topic of extensive research. For example, hatchery fishes seem unable to feed properly after their release into the wild (Johnson et al. 1996, Johnsen and Ugedal 2008, Bachman 1984). Bachman (1984) reported that stream brown trout spent the majority of their foraging time (86%) in a sit-and-wait search state and spent 17-43% under cover depending on their age (Bachman 1984). Hatchery fish in the same environment fed less, and used more energy in their pursuits (Bachman 1984). A similar conclusion was reached while observing out-migrating Atlantic salmon. The hatchery fish tended to have less food in their stomachs (Johnson et al. 1996). Johnsen and Ugedal (2008) found that after release into the wild, hatchery fish (brown trout) needed a week before they had similar eating habits as wild fish.

The survival rate of hatchery fish may also be affected by how they respond to a predator and novel environments. Antipredator behavior may have the greatest effect on survival rates and appears to be perfected only with experience (reviewed in Bori et al. 1998). Berejikian (1995) found that experience decreased the level of predation on wild and hatchery rainbow trout, but the wild population survived better than the hatchery fish regardless of treatment. Wright et al. (2006) found that genetic variation underlies the phenotypic variation of antipredator behavior in zebrafish. Lucas et al. (2004) found that genetic variation contributed to differences in antipredator behavior in clonal rainbow trout lines as well.
The genetic underpinnings of antipredator behaviors are likely complex in nature and involve many interacting components. Early research on mammals implicated amygdala function in the response of animals to conditioned stimuli after a stressful event (reviewed in LeDoux 2003). Typically, the unconditioned response of an organism to such an event would be to freeze or try to escape (LeDoux 2003). The medial pallia of the telencephalon act in a similar manner in ray-finned fishes as the amygdala in tetrapods (Broglio 2005). A conserved response to fear conditioning between mammals and other vertebrates (e.g. bony fishes) correlates with similar amygdala structures seen in a diversity of vertebrate organisms (reviewed in Pedro et al. 2012). In the previous example of experience contributing to survival success (Berejikian 1995), the unconditioned response came from a fish viewing conspecifics being preyed on, and the conditioned stimuli were the predators. This response was likely mediated by the telencephalon.

When rainbow trout are exposed to damaged conspecific tissue and pike odor, the fish will exhibit characteristic behaviors of fear after being exposed to pike odor alone in further experiments (Brown and Smith 1998). Chinook salmon show similar behavior when conditioned to the smell of cutthroat trout (Berejikian et al. 1999). The unconditioned stimulus in these experiments was the alarm substance found in the skin of the conspecifics and the responses were likely genetically driven because they require no learning and are stereotypic (LeDoux 2003, Pedro et al. 2012). In these cases a part of the telencephalon (homologous to the amygdala) mediates the response to the conditioned response (Pedro et al. 2012).

The brain structures effected downstream of the amygdala are likely important in behavioral responses in both mammals and fishes. Rats unconditionally fear cat odor. When they are exposed to the odor, neurons in the medial amygdala are activated followed by activation of structures in the hypothalamus (reviewed in Dielenberg 2006). In fishes, the amygdala-like medial pallia of the telencephalon likely
affect many other brains structures as well, with downstream effects on respiration, the parasymathetic nervous system, reflexes, jaw movements, and other systems (reviewed in Davis 1992).

In fishes, the primary physiological response to a major stressor is the activation of the hypothalamus and pituitary gland with the release of catecholamines (epinephrine, norepinephrine, and dopamine) and corticosteroids (cortisol, aldosterone, and corticosterone) from the chromaffin and interrenal cells in the head kidney (corresponding to the mammalian adrenal medulla, reviewed in Gallo and Civinini 2003) respectively (reviewed in Bonga 1997). Both types of pathways (hypothalamic-sympathetic-chromaffin cell axis and hypothalamic-pituitary-interrenal axis) affect energy utilization (Bonga 1997). Cortisol levels are elevated minutes after an acute stressor and may not return to standard levels for an hour (Bonga 1997). The level of this primary stress response and the rate at which homeostasis returns may be critical to the variation seen between hatchery and wild antipredator behavior in fishes.

Modern analyses have frequently turned to personality traits as a model for interpreting antipredator behavior in fishes (reviewed in Schjolden and Winberg 2007). In such a model, populations are composed of proactive (bold) and reactive (timid) individuals. During a stressful event, reactive individuals may show higher hypothalamic-pituitary-interrenal axis activity and release of more catecholamines (Schjolden and Winberg 2007).

In rainbow trout, a single allelic variation of the MHC 1a locus correlated with proactive and reactive behavior (Azuma et al. 2004). In poeciliids, offspring showed similar behavior phenotypes as parental fish (Brown et al. 2007). These results suggest that personality traits have a heritable component in ray-finned fishes. They also suggest that single allelic differences may affect whether an individual is categorized as bold or shy.

Evidence from Wright et al. (2006) suggests that boldness is polygenic (i.e. influenced by multiple loci). These researchers found multiple quantitative trait loci (QTL) associated with variations in boldness.
measured between laboratory and wild zebrafish. Similar QTL experiments in other fish models would help in our understanding of the genetic architecture of personality in fishes.

**Genome duplication in salmonids**

Two genome duplications are suspected to have occurred in the ancestors of all vertebrates leading to an increase in gene number and diversity (reviewed in Meyer and Schartl 1999; Meyer and Van de Peer 2005). A genome duplication also transpired early in teleost evolution (Jaillon et al. 2004, Taylor et al. 2003, Christoffels et al. 2004) after the split between ray-finned and lobe-finned fish (Meyer and Schartl 1999, Meyer and Van de Peer 2005). Another duplication specific to salmonids occurred long before the speciation of modern salmonids (Allendorf and Thorgaard 1984). These duplications created multiple copies of every gene, though many copies are expected to have been lost through time (reviewed in Wolfe 2001).

Gene duplication can complicate analyses of data derived from organisms with ancestral genome duplications. At QTL (quantitative trait loci), genetic variation at a locus correlates with phenotypic variation of a trait. When comparing QTL between experiments, variation at one duplicated gene locus may be responsible for phenotypic variation in one experiment and the other gene copy may be responsible for phenotypic variation in the other experiment. Without knowledge of gene duplication, interpretation of such a dataset would incorrectly lead to the conclusion that the phenotypic variation was caused by different mechanisms in the different experiments. Correctly identifying duplicated genes may prevent such misinterpretations.

At a genomic level, whole genome duplications may be identified by chromosome sequence alignments or through synteny searches, a.k.a. map-based approaches (Wolfe 2001, The Arabidopsis Genome Initiative 2000). Typically, at a single gene level, sequencing or phylogenetic analyses (Taylor et al. 2003)
are used to identify duplications, a.k.a. tree-based approaches (Wolfe 2001). Map-based approaches require a sequenced genome, while tree-based approaches require long sequences of a gene.

Chromosomal alignments and synteny searches are dependent on the quality of the sequenced genome. Unfortunately, assembly quality will often be low for duplicated regions of the genome. In the mouse, segmental duplications, with greater than 90 percent sequence identity, are often missassembled (Bailey et al. 2004). Depending on when a segment or whole genome was duplicated, inappropriate assemblies may mask recent duplications. Conversely, genomic rearrangements may interfere with detection of ancient genome duplications (Wolfe 2001).

Tree-based approaches require long sequences and may give mixed results depending on which genes are analyzed. In order to identify a common duplication, multiple genes are analyzed for the level of sequence-divergence (Wang and Gu 2000). Diploidization (elimination of duplicated copies of genes from the genome), gene conversion (homogenization through recombination between the two duplicated loci), and allopolyploidy may make identification of a genome duplication difficult using the tree-based approach (Wolfe 2001). If nucleotide variations are not consistent among duplicated gene pairs, a single genome duplication will not be detected because the variation may not be attributed to a single evolutionary event.

Duplications of a single gene may be found through sequence alignments, where the gene aligns to multiple locations on a chromosome (Suyama et al. 2004, Suyama et al. 2005). Another method is to align all gene sequences of an organism against all of the same sequences (all-against-all) and create a phylogenetic tree with multiple species (Chen et al. 2000, Han et al. 2009). In this case each of the duplicated genes will form separate clusters. An all-against-all search may also be used in conjunction with alignment length and sequence similarity criteria to identify duplicated genes (Ouedraogo et al. 2001).
2012). These methods require full gene sequences, and potentially the complete genome sequence of an organism.

With the growing abundance of transcriptomic data available from high-throughput sequencing, a method to identify gene duplications that could utilize this information would be advantageous for organisms without sequenced genomes or large gene databases. Also, using the transcriptome means that genome rearrangements would no longer affect genome duplication analyses. A potential complication from using the transcriptome is that duplicated genes may not be expressed at the same time. Using multiple tissues at different times may help avoid this complication as long as the gene is at some time expressed. If a gene is not expressed, gene sequence and function would degrade and would be unlikely to interfere with other types of research (e.g. QTL analysis).

**SNP Detection in Polyploids**

Another complication of research with tetraploid organisms arises when searching for allelic variation. The variation between duplicated loci (paralogous sequence variation), is often mistaken as allelic variation (Marth et al. 1999, Eichler et al. 2004). Eichler et al. (2004) suggested that even in the finished human genome, segmental duplications could cause paralogous sequence variation to be mistaken for allelic variation. They also show that many of the gaps in the euchromatin portion of the human genome remain because of duplicated genomic regions.

Several strategies have been developed to address paralogous sequence variants (Hand et al. 2008, Marth et al. 1999, Miller et al. 2012). Hand et al. (2008), were able to compare sequences of an allopolyploid to its extant ancestral genomes to find allelic variation; the ancestral genome would have greater sequence similarity to the alleles of a gene than to a paralogous sequence. Marth et al. (1999) eliminated possible sequences that may have been duplicated (i.e. if they suspected a sequence was paralogous, they would eliminate it from further analysis), and Miller et al. (2012) eliminated sequence
variants if they had an excess of heterozygosity in a panel or if they were unable to genetically map the variant. Trick et al. (2009) suggested that paralogous sequence variants may be distinguished from allelic variants when comparing sequences from inbred lines.

Without extant ancestral genomes (or gene sequences) or with autopolyploidy, the first method would not be possible. Eliminating duplicated genes means allelic variation would be missing for many genes or sequences depending on when a duplication event occurred. Heterozygosity checks and genetic mapping offer a practical approach for SNP evaluation, but require sequencing of many individuals.

All of the methods described, except the use of ancestral genomes, interrogate every nucleotide position individually to identify allelic variation and distinguish between paralogous sequence variants. Nucleotide information around each potential SNP is ignored, but could be used to distinguish between allelic and paralogous sequence variants. An approach for using such information in identifying allelic variation was developed in the current study.
References


Araki, H, B Cooper, and MS Blouin. 2007. Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. Science 318(5847):100-103.


Berejikian, BA. 1995. The effects of hatchery and wild ancestry and experience on the relative ability of steelhead trout fry (Oncorhynchus mykiss) to avoid a benthic predator. Canadian Journal of Fisheries and Aquatic Sciences 52:2476-2482.


CHAPTER TWO

ANTIPREDATOR BEHAVIOR QTL: DIFFERENCES IN RAINBOW TROUT CLONAL LINES DERIVED FROM WILD AND HATCHERY POPULATIONS

Abstract

Variation in antipredator behavior may partially explain the survival differences seen between wild and hatchery trout and salmon. Antipredator behavior is thought to change during the domestication process, along with other traits. Investigations of antipredator behavior could benefit conservation efforts and supplementation programs. Our goal was to characterize the antipredator behavior in clonal rainbow trout lines derived from either wild or hatchery populations and identify genetic loci associated with variation between lines. We identified several behaviors that varied between clonal lines and QTL for several behavioral and size traits. Characterizing genetic variation underlying these behaviors may prove valuable in future conservation efforts by enabling monitoring of allele frequencies of loci affecting predation in wild populations.

Introduction

Behavioral and physiological traits in hatchery-reared fishes have often been altered in a characteristic manner associated with domestication. These studies have been most widely conducted in trout and salmon. Alterations have been observed in size (Devlin et al. 2001), aggression (Lucas et al. 2004, Swain and Riddell 1990), antipredator behavior (Biro et al. 2004, Johnsson and Abrahams 1991), and reproductive behavior (reviewed in Huntingford 2004). Understanding how these traits are modified under artificial propagation is important for both conservation and aquaculture. Pleiotropic genes may play a significant role in the conversion of wild to domesticated phenotypes.

When Johnsson et al. (1996) injected growth hormone into wild brown trout, they found that the trout displayed size and antipredator behavior similar to hatchery-reared brown trout. In the chicken, the
arginine vasopressin receptor 1a gene affects size and behavior (Wirén et al. 2009). These examples are of genes affecting both size and behavior though it is expected many genes would be involved.

Hatchery managers often inadvertently selected for large size in the past (Brannon et al. 2004) and may have consequently selected for changes in growth, antipredator behavior, and other characteristics associated with domestication. Any alteration in growth or antipredator behavior may affect how an organism interacts with its environment (Tymchuk et al. 2009) and their fitness.

Araki et al. (2007) found that the fitness of steelhead (sea-run rainbow trout, *Oncorhynchus mykiss*) declined by 40 percent with each generation spent in a hatchery. This implies that if the change in fitness is genetically driven, there must be a strong selective pressure against hatchery-reared fish in the wild. At least a portion of such a drastic decline in fitness might result if a fish is unable to properly respond to a predator.

Several studies suggest that antipredator behavior is altered when fish are raised in artificial environments. Sundström et al. (2004) and Johnsson et al. (1996) found that sea-ranched brown trout were bolder on average than wild brown trout. Fleming and Einum (1997) found that juvenile, farmed Atlantic salmon left cover after contact with a predator sooner than wild salmon. Wright et al. (2006) found that domesticated zebrafish took less time to enter a zone with a novel object present than wild zebrafish. Presumably, an alteration of antipredator behavior would be beneficial in an environment with no predators and large quantities of food. A bolder fish could obtain more food, grow faster, and reach maturity quicker. However, in a wild setting, bolder fish may be more susceptible to predators (Biro and Post 2008).

Antipredator behavior includes how fish recognize, process, and physically respond to predators. Recognition and processing of threats (Flood et al. 1976) occurs in the telencephalon (Overmier and Papini 1986) where the medial pallium, similar to the amygdala in mammals, is located (Broglio et al.
The amygdala plays an important role in emotional memory such as fear conditioning (Adolphs et al. 1995, Broglio et al. 2005).

The hypothalamus mediates the stress response following a stressful event (reviewed in Bonga 1997 and Martínez-Porchas et al. 2009). The hypothalamic-pituitary-interrenal (HPI) axis regulates the release of potential stress-related hormones/molecules (Bonga 1997). These hormones affect the response of a fish to a predator or its behavior after encountering a predator. Changes caused by selection or environmental factors that alter the medial pallium, hypothalamus, HPI axis, or the concentrations of various hormones may all affect antipredator behavior.

Researchers have used various assays to measure antipredator behaviors including: time to leave cover (Harcourt et al. 2009, Fleming and Einum 1997, Einum and Fleming 1997), novel object avoidance (Sundström et al. 2004), open field tests (Brown et al. 2005), foraging in the presence of a predator (Magnhagen 2006, Johnsson et al. 1996, Johnsson and Abrahams 1991), and predator inspection (Godin and Dugatkin 1996). All of these assays estimate risk of predation and should correlate with survival if a population is under strong predation. The anadromous life history of salmonids incurs an increased risk of predation as they travel through marine and freshwater ecosystems with numerous predators.

A population may be comprised of proactive (bolder) fish and reactive (timid) fish (reviewed in Huntingford 2004, Brown et al. 2005, Brelin 2008). Brelin et al. (2005) found that plasma noradrenalin levels correlated strongly with these two types of responses, with the proactive fish showing a significant increase in noradrenalin. Brelin et al. (2005) observed sea-ranched brown trout after transferring them to observation tanks, after an intruder was introduced, and under stressful events. They discovered that the trout could be assigned as proactive (n = 6) or reactive (n = 17) based on noradrenalin levels and clustering of fish responses to the different stimuli. Domestication may shift a population toward a more proactive mean if reactive fish are selected against in a hatchery.
environment, or as Brown et al (2007b) and others (Brelin et al. 2008) suggest, underlying personality traits are shifted from individuals with shy to bold phenotypes. While it is possible to select for these behavioral types in rainbow trout (Pottinger and Carrick 1999), the consistency and heritability of behavioral type remains uncertain (reviewed in Sih et al. 2004).

The first attempt to identify the genetic architecture underlying an antipredator behavior in fishes was conducted with zebrafish (Wright et al. 2006). The researchers found three consistent QTLs which underlie differences in boldness between wild and domesticated zebrafish but did not identify candidate genes for these QTLs. Identifying such genes and the genetic architecture of antipredator behavior in other fish species could help to identify general mechanisms governing antipredator behavior.

Our research goal was to identify QTL(s) underlying several antipredator behaviors in rainbow trout in response to four different stimuli. Two different panels of fish from crosses between clonal lines (Young et al. 1996, Robison et al. 1999) differing in domestication level (number of generations spent in a hatchery) were phenotyped for the same behavioral traits and for size and genotyped for DNA markers.

**Materials and Methods**

*Experimental Fish*

Two crosses were produced from four clonal lines of rainbow trout. The first cross was between the Whale Rock (XX female) and Arlee (YY male) clonal lines, both from California. The Arlee line has been under artificial propagation for over 50 generations, while the Whale Rock clonal line was produced from gametes taken from the wild. After these clonal lines were produced they have been propagated at Washington State University for a number of generations. Fish that are homozygous for all loci remain genetically uniform and are no longer affected by selection or domestication.
The other cross was produced from the OSU (XX female) and Swanson (YY male) clonal lines. The OSU line was originally derived from the Mt. Shasta, California hatchery and has been under propagation for over 50 generations. The Swanson line was established from broodstock from the Alaska Department of Fish and Game’s Fort Richardson hatchery and had been under propagation for 2-3 generations when sperm was collected to found the line.

Recombinant doubled haploid progeny were produced through androgenesis (Young et al. 1996, Young et al. 1998, Robison et al. 2001) from the hybrid Whale Rock x Arlee (n = 154 doubled haploids) and OSU x Swanson (n = 152 double haploids) male parents. Three clonal lines (OSU, Swanson, and Arlee) and the two doubled haploid crosses (also referred to as a panel later) were raised at the Washington State University hatchery until one month after swim-up. The Whale Rock (XX female) line was unavailable for observations, but we were able to study the Whale Rock (YY male) line, genetically a full-sib of the female line.

**Filming strategy and processing (Phenotyping)**

Fry reared under standard conditions (10°C, at approximately the same densities n = 20, using natural light-dark cycles and commercial feed once a day to satiation) for one month after swim-up (Table 1) were taken from their rearing tanks and placed into experimental chambers in a 10°C cold room (Figure 1). Each of the observation tanks had a segment of cover over more than half of the bottom of the tank that the fish could hide under. The clonal lines were observed another two times in order to understand the effects of age and habituation for these experiments: two months after swim-up and the day after the two month observation. To understand experimental repeatability, naive fish from the Whale Rock (YY) and Arlee (YY) clonal lines were measured on two separate occasions one month after swim-up (Washington State University Institutional Animal Care and Use Committee approved protocol #02991-005 and #02991-06).
In every trial, the fish were filmed for twenty-five minutes after being placed into the experimental tanks, similar to an open field test (Brown et al. 2007a, Kilgour 1975)(called After Introduction below). The fry were then allowed to acclimatize overnight, and in the morning were filmed for another thirty minutes. Following this Prestimulus phase, an artificial predator (a Rapala Ireland Floating Lure, 5 inches without hooks mounted on a metal rod) was lowered into the tank four times in a row, hitting the cover in case the fry was unable to see the predator and the fry was again filmed for thirty minutes (After Scare). Finally, feed was placed into the tanks and filming lasted another thirty-five minutes (After Feed). When filming the clonal lines, the tanks were alternated between the lines to reduce any effect of tank bias.

Videos were then processed using SwisTrack software (Lochmatter et al. 2008) to identify the position of the fry in every frame. Missing data (caused by background noise interference) were inferred from later frames using Perl scripts. The x and y coordinates were averaged for three frames to reduce the effect of noise (e.g. identifying the x and y position for the head and then for the tail in the next frame). The following metrics were calculated using the averaged positional information with Perl scripts: Average swim level, total movement, number of moves to the top, percent cover use, percent spatial use, percent top use, and bottom use.

The average swim level was calculated by averaging the y coordinates of a film and then subtracting the top value from this average value and dividing by the number of pixels between the top and bottom of the tank (e.g. top = 10 pixels, bottom = 400 pixels, average = 15 pixels, and the normalized value = 0.005: low values are near the top, and high values are near the bottom with a range from 0-1). This normalized the results for tank size, which varied due to camera positioning. The total movement was calculated using the Pythagorean theorem to find how much a fish moved between two time points and then adding this value to all other lengths from all time intervals. The number of moves to the top (feed
attempts) was counted if a fish moved from the bottom of the tank to the top at any point during filming (the top of the tank was defined as the top 100 pixels or ~1/4).

The percent cover used was calculated as the percent of the frames where the fish was within a defined boundary of the cover. Spatial use was calculated as the percent of the pixels the fish occupied during filming relative to the total pixels of a tank (note: this value will never reach 100 percent because there are areas of the tank the fish may not use and the size of a fish was reduced to only one pixel—the x and y coordinate). The percent top and bottom use were calculated in a similar manner as the cover use (the top was defined as the top ~1/4 and the bottom was the bottom ~1/4 of the tank).

A "bold" binary trait was also calculated from the average swim level of the After Introduction, Prestimulus, and After Feed time intervals. The fish was counted as bold (value of one) if the average level was above the bottom of the tank by 30 percent for the three time intervals. A heat map was created using custom Perl scripts and the R statistical package to visualize where the fry tended to swim (Figure 2). Perl scripts are available on request.

**Genotyping**

After filming, fish were anesthetized using MS-222 and rinsed fin clips were taken for DNA extraction after wet weight and length were measured. The condition factor was calculated from the weight and length information (Barnham and Baxter 1998). A phenol-chloroform protocol (Sambrook et al. 1989) was used for DNA extraction and an AFLP protocol (Vos et al. 1995) or an AFLP-style high-throughput sequencing protocol that is similar to RAD sequencing (Parchman et al. 2012) was used to genotype each fry. For AFLP genotyping, polyacrylamide gels were visualized on a Typhoon 9400 variable mode imager (Amersham Biosciences) and were manually scored for polymorphic markers using various restriction enzyme combinations (similar to Robison et al. 2001).
For the high-throughput sequencing, SbfI (New England BioLabs - NEB, as in Miller et al. 2012) was used instead of EcoRI as in Parchman et al. (2012) and AccuPrime Pfx DNA Polymerase was used. The DNA restriction digest was performed simultaneously with adaptor ligation using the following reagents: 1.2 µl 10x T4 buffer (NEB), 0.65 µl 1M NaCl, 0.65 µl 1mg/ml BSA, 1 µl 100µM MseI adaptor, 1µl 50nM barcoded SbfI adaptor, 1.45 µl water, 0.28 µl SbfI, 0.27 µl MseI (NEB), 5.5 µl DNA (20 - 100ng total). These reagents were combined and incubated at 37°C for 2 hours in a thermocycler. Diluted (1:100 in 1x T4 Buffer), high concentration T4 Ligase (NEB) was then added (0.5 µl) to the reaction and was incubated for another 3 hours followed by an overnight incubation at room temperature. The restriction/ligation mixture was diluted with 189 µl 0.1 x TE the following morning and PCR was then performed.

The following reagents were used for PCR: 2 µl 10x AccuPrime Pfx reaction mix, 1 µl each of 10 µM primer (see Parchman et al. 2012 for primer sequence, with a modification on the MseI-Reverse primer - an N instead of a G on the 3’ end), 4 µl of diluted restriction/ligation mixture, 0.4 µl Accuprime Pfx DNA Polymerase, 11.6 µl water. The following conditions were used in a thermocycler: 2 min at 95°C, then 16 cycles at 95°C for 20 sec, 66°C for 30 sec, and 68°C for 40 sec.

Gel purification was similar to Parchman et al. (2012) except the pooled PCR product was speedvac concentrated, and an agarose gel of 1.5% was used instead of 2.5%. The use of the speedvac made it necessary to run the gel longer than recommended because the gel was overloaded. The gel-purified product was quantified using a Hoefer DyNA Quant 200 Fluorometer.

The quantified product was then sent to the IBEST Genomics Core at the University of Idaho, where the sample was sized on an Agilent 2100 Bioanalyzer. Finally, the sample was sent to the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley where it was sequenced on an
Illumina HiSeq2000 sequencer. High-throughput genotyping was performed in two separate runs on the OSU x Swanson panel n = 152, OSU n = 3, and Swanson clonal lines n = 3.

Sequences were processed using the protocol and Perl scripts (with lower quality stringencies used during the read filtering step) from Miller et al. (2012) to identify SNPs in the parental clonal lines and then score the SNPs in the individuals of the doubled haploid panel. A genetic map was created using R/qtl (Broman et al. 2003) from the AFLP data (Whale Rock x Arlee panel) and the high-throughput data (OSU x Swanson panel). Microsatellites (Rexroad et al. 2008) were used to assign linkage groups from the AFLP dataset to their respective chromosomes. Custom Perl scripts, and Novoalign (Novocraft) were used to link the OSU x Swanson panel map to that from Miller et al. (2012) in order to identify the correspondence between chromosomes and linkage groups. Quantitative trait loci analysis using interval mapping was performed using R/qtl. A permutation test using 1,000 permutations was used to set the significance threshold of a QTL.

Results

Clonal Line Behavior

The average length and weights of the four clonal lines differed significantly based on a one-way anova, with the hatchery lines being larger than those derived from the wild (Table 2, p < 0.001). To understand how the behaviors under investigation might vary between experiments, naive Whale Rock (YY, trial 1a n = 12 and trial 1b n = 16) and Arlee (YY, Trial 1a n=12 and Trial 1b n=16) clonal lines were filmed at two different times (Trial 1a in 2010 and Trial 1b in 2012). Rearing, filming, and environmental conditions were similar and all fish were aged one-month after swim-up. Although there was considerable variance between the same clonal lines at different times (i.e. Arlee vs. Arlee), the pattern between the Arlee and Whale Rock clonal lines remained fairly similar for the different time points (Figure 3). In both cases the Arlee tended to use the cover more and moved to the top less often. Both
lines tended to use the cover less during trial 1b than trial 1a (Figure 3). The variation seen between times may reflect the effect of uncontrolled variation in the environment or maternal effect.

Observations from the three trials (one month after swim-up, two months after swim-up and two months and a day after swim-up) again show large variation, but consistent patterns between the different clonal lines (Figure 4). When most fish were first placed into the observation tank they tended to explore the tank, and rarely utilized the cover (Figure 4, personal observation). After an overnight acclimatization, many of the fish utilized the cover more, and after being scared used the cover even more (Figure 4). After being fed, the fry would leave the cover and make more moves to the top (Figure 4) in order to feed on flakes on the surface of the water.

These trends could basically be seen in all four clonal lines during all of the trials. For the first trial (one-month after swim-up), the fish tended to hide less often than in later trials (Figure 4). It is difficult to say whether this is an effect of habituation or age. Between the second and third trials, the behaviors varied much less than between the first and the second or third, suggesting that habituation played only a minor role. The interval of only a day between the second and third trials would likely have a greater habituation effect than the month between the first and second trial (during the first and second trials the fish would have more time to forget their previous experiences).

Since there was a greater difference between the first and second trials, age may have played a larger role in the observed variation. Interestingly, between trials the relative trait values of the different clonal lines remained consistent. Perhaps the age of the fish altered how the fry responded to the different stimuli, but the clonal line differences remained similar in all trials. These behaviors show a norm of reaction whereby the two genotypes (wild vs. hatchery-derived clonal line) have parallel slopes for the different times, the relative rankings remain similar, and the absolute trait values change depending on environment (in this case different times). The clonal line origin’s effect on behavior
seemed consistent between trials, with Arlee and OSU (both with over 50 generations of artificial propagation) hiding more than the other two lines for at least part of the time during a trial (Figure 4).

*Doubled Haploid Panels (QTL results)*

Both doubled haploid panels showed similar trends to their parental clonal lines during the four time intervals for percent cover use (Figure 5). The largest variation between the two panels occurred during the *After Scare* and *After Feed* intervals between the individuals that spent all of their time under cover and those that spent less time under cover. Why this variation exists and its significance remain unclear.

Significant quantitative trait loci (QTL) were seen for various behaviors and physical traits (Figures 5, 6, and 7). Figure 5 shows the phenotypic average for the Whale Rock and Arlee genotypes in the doubled haploid panel for the percent cover use during the *After Introduction* phase. Figure 6 illustrates the QTL from the Whale Rock x Arlee panel. Several behavioral QTL overlap on different chromosomes. Interestingly, many QTL occur on the X-chromosome (close to the sex marker), which may indicate a correlation between sex and these behaviors (Figure 6). The only trait that had a significant correlation between sex (genetic marker) was the *After Feed* percent bottom use ($r = 0.23$), though this QTL was only suggestive and not statistically significant. Both the X and Y rainbow trout chromosomes are large (Thorgaard 1977, Phillips et al. 2009) and the sex marker may not actually be physically close to these QTL. All of the behavioral traits producing QTL in this panel were significantly correlated with weight (from $r = 0.19$ to $r = -0.44$).

Fewer QTL overlap with each other in the OSU x Swanson panel (Figure 7), and fewer QTL were observed. None of the QTL from the Whale Rock x Arlee panel overlie the QTL from the OSU x Swanson panel unless the linkage groups that we were unable to assign back to the physical chromosomes match. These results suggest that multiple genetic mechanisms can alter the behavioral traits under
investigation. The *Prestimulus* total movement trait has a significant correlation with sex \((r = 0.18)\), and the *Prestimulus* percent top use trait has as a significant correlation with weight \((r = 0.19)\).

Based on the comparative map presented in Rexroad et al. (2008), two QTLs for zebrafish boldness (Wright et al. 2006) may be found on the same chromosomes as QTL from this study. In the zebrafish a boldness trait QTL, measured by the number of times a fish inspected a novel object, was found on chromosome 23 and this may correspond with chromosome 9 in rainbow trout where a QTL was found (Figure 6). Another zebrafish QTL for shoaling tendency was found on chromosome 8, and this may correspond to the QTL found on chromosome 8 (Figure 7).

**Discussion**

The clonal rainbow trout lines significantly differed from each other at several time points for different behaviors, but the large variation between the naive fish of the same clonal line means that our interpretations must be made with caution. It is important to remember that each of the clonal lines represents only a single haploid genotype and may not be broadly representative of either hatchery or wild populations.

Although we may not be able to extrapolate the data to make broad statements about wild and domesticated populations, the fish from two clonal lines with ~50 generations of hatchery rearing did show differences from two lines with 0-3 generations of domestication. These differences were relatively stable between the first and second months after swim-up and between experiments. Based on other studies (Johnsson and Abrahams 1991, Johnsson et al. 1996, Lucas et al. 2004, Reinhardt et al. 2001), we expected the Whale Rock and Swanson lines to be on average less bold than the OSU and Arlee, but the reverse was observed.
If we examine our results in the context of predation instead of in terms of domestication, our observations fit into place with existing literature. In Brown et al. (2005) and Brown et al. (2007b) researchers observed that a poeciliid population under heavy predation had bolder phenotypes than a population without predation. Brown trout survived more often in the wild if they were more active, a trait correlated with boldness (Adriaenssens and Johnsson 2013) and possibly territoriality. These studies match our observations of the Whale Rock and Swanson wild lines utilizing the cover less and moving to the top of the tank more.

Almost all fish responded to our artificial predator in a characteristic manner, using the cover more and moving less. There was a clonal line difference between these two metrics during the After Scare time interval, but all lines would hide under the cover immediately after they were exposed to the predator (similar to Ryer (2004)). The lines with lower cover usage tended to leave the cover sooner and stay out of the cover once they left (data not shown). In Ryer (2004), significant differences were not found between wild and domesticated lines of fish for activity after the researcher observed fish for 10 minutes after a predator was introduced. Our clonal lines did not show significant movement differences for 10-15 minutes for one line comparison and 15-20 minutes for another line comparison (data not shown).

The behavioral traits we examined tend to vary considerably. This variation may be influenced by many small-effect genes. If each genetic component varied with environmental influences and was small, the total variation would be expected to be large. With QTLs identified in zebrafish for boldness and now in rainbow trout for antipredator behavior potentially being on the same chromosome, it is plausible that one of these loci commonly affects behavior type (i.e. affects whether an individual is proactive or reactive). Of course, much more research must be undertaken to make this assertion. Better comparative maps and a sequenced rainbow trout genome would facilitate these comparisons.
Our main goal was to identify traits that varied between rainbow trout clonal lines and identify possible regions of different chromosomes that are responsible for behavioral variation to improve our fundamental understanding of the genetics of domestication. In this regard, we were able to detect genetic variation underlying many of the behavioral differences we observed in the parental clonal lines. Characterizing genetic variation underlying these behaviors may prove valuable in future conservation efforts, especially when predators play a large role in a fish's life history and survival. Such a fundamental understanding of domestication genetics might allow markers to be developed that researchers could use to monitor allele frequencies of loci greatly affecting boldness in wild populations. This information, ideally, could be valuable when evaluating the effect of supplementation programs, invasive species, or habitat alteration.

The large investment (Oregon $22.9 million - Oregon Department of Fish and Wildlife, Washington $63.9 million - Washington Department of Fish and Wildlife) that governments are placing in hatcheries could have a far greater return if hatchery survival rates increased. If we have a better understanding of both the scope of antipredator behaviors and the genetics that underlie variations of these behaviors, we could potentially be more successful at raising fry that are better able to respond to predators in the wild. To this end our future directions could include testing more lines and crosses that vary in antipredator behavior. We would also like to characterize identified QTL and the candidate genes responsible for the QTL, including conducting fine-mapping with populations of fish. With a sequenced genome, identifying candidate genes should become increasingly feasible and affordable.

Acknowledgements

Support for KAC came from the National Institute of General Medical Sciences, Award Number T32GM083864 and Washington State University. Thanks to Ruth C. Newberry for insightful and valuable comments.
References


Araki, H, B Cooper, and MS Blouin. 2007. Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. Science 318(5847):100-103.


Oregon Department of Fish and Wildlife. www.dfw.state.or.us/fish/hatchery/, viewed 11/24/2013.

Overmier, JB, and MR Papini. 1986. Factors modulating the effects of teleost telencephalon ablation on retention, relearning, and extinction of instrumental avoidance behavior. Behavioral Neuroscience 100:190-199.


### Table 2.1. Dates of Fertilization and Trials of Rainbow Trout Clonal Lines and Doubled Haploids

<table>
<thead>
<tr>
<th>Clonal Line</th>
<th>Fertilization</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlee (YY)</td>
<td>03/18/2010</td>
<td>06/25 - 6/28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whale Rock (YY)</td>
<td>03/18/2010</td>
<td>06/25 - 6/28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arlee (YY)</td>
<td>04/19/2012</td>
<td>08/02 - 08/05</td>
<td>09/03 - 09/07</td>
<td>09/04 - 09/08</td>
</tr>
<tr>
<td>Whale Rock (YY)</td>
<td>04/19/2012</td>
<td>08/02 - 08/05</td>
<td>09/03 - 09/07</td>
<td>09/04 - 09/08</td>
</tr>
<tr>
<td>Swanson (YY)</td>
<td>05/04/2011</td>
<td>08/03 - 08/05</td>
<td>09/22 - 09/26</td>
<td>09/23 - 09/27</td>
</tr>
<tr>
<td>OSU (XX)</td>
<td>05/04/2011</td>
<td>08/03 - 08/05</td>
<td>09/22 - 09/26</td>
<td>09/23 - 09/27</td>
</tr>
<tr>
<td>Whale Rock x Arlee</td>
<td>03/11/2010</td>
<td>06/12 - 07/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSU x Swanson</td>
<td>05/04/2011</td>
<td>08/08 - 08/10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>05/30/2011</td>
<td>08/16 - 09/08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2.2. Trial 1 Lengths and Weights (mean ± SE) of Clonal Lines of Rainbow Trout

<table>
<thead>
<tr>
<th>Clonal Line</th>
<th>Length (cm)</th>
<th>Weight (g)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swanson</td>
<td>2.8 (0.04)</td>
<td>0.31 (0.13)</td>
<td>9</td>
</tr>
<tr>
<td>Whale Rock</td>
<td>2.92 (0.06)</td>
<td>0.42 (0.03)</td>
<td>16</td>
</tr>
<tr>
<td>Arlee</td>
<td>3.175 (0.04)</td>
<td>0.59 (0.03)</td>
<td>16</td>
</tr>
<tr>
<td>OSU</td>
<td>3.48 (0.03)</td>
<td>0.55 (0.03)</td>
<td>12</td>
</tr>
</tbody>
</table>
Figures

**Figure 2.1.** Camera and Tank Setup. Four tanks and two cameras were used to record the various behaviors analyzed in this study. Fluorescent lights were directly above tanks to reduce the background lighting noise (i.e. making the fish more distinguishable from the background). The tank setup was surrounded by curtains that restricted the fish’s view. In each tank a small cover was at the bottom where a fish could hide. The dimensions of a tank were: 38 cm height, 26 cm length, and 10 cm width.

**Figure 2.2.** Heatmap of a Single Fish (clonal Arlee - YY). Data of the four different stimuli are represented: AI stands for *After Introduction* (right after the fish was transferred to the tank), PS stands for *Prestimulus* (after an overnight acclimatization), AS stands for *After Scare*, and AF stands for *After Feed*. This heatmap shows, for each pixel column, where the fish was most of the time (with yellow representing most time and white representing no time). A scale model of the tank and fish are on the side for comparison.

**Figure 2.3.** Comparing Experiments from Different Times. Panel A illustrates the results from two experiments (trial 1a and trial 1b) performed two years apart between the Arlee and Whale Rock clonal lines for the percent cover use (*AI*-After *Introduction*, *PS*-Prestimulus, *AS*-After *Scare*, AF-*After Feed*). For both experiments the fish were naive and performed one-month after swim-up. Panel B illustrates the number of times the fish went to the surface.

**Figure 2.4.** Comparing Clonal Lines for Three Different Trials. Panel A shows the percent cover use for the four different clonal lines for trial 1 (one-month after swim-up), trial 2 (two-months after swim-up), and trial 3 (two-months and one day after swim-up). The same fish from trial 1 were used in subsequent trials. The different portions of each trial are: AI-*After Introduction*, PS-*Prestimulus*, AS-*After Scare*, AF-*After Feed*. Panel B represents the data on how many times the fish moved from the bottom of the tank to the top.

**Figure 2.5.** A Histogram of the Doubled Haploid Panels Percent Cover Use. On the bottom is the average score for the different genotypes (Arlee or Whale Rock) for a percent cover use QTL. The error bars are the standard error of mean.

**Figure 2.6.** QTL Results Whale Rock x Arlee Panel. The distance of a QTL was determined by a drop in the LOD score of 0.9. LOD scores for the various traits were: *Condition_Factor* 3.19, *After_Introduction_Total_Movement* 4.71, *After_Introduction_Moves_to_top* 2.81, *Bold_Binary* 3.17, *After_Introduction_Percent_Top/Bot* 7.22 and 3.01, *After_Introduction_Percent_Cover* 7.56, *After_Introduction_Level* 7.34. Gray QTL are suggestive, but not statistically significant. Image created using MapChart (Voorrips 2002).

**Figure 2.7.** QTL Results OSU x Swanson Panel. The distance of a QTL was determined by a drop in the LOD score of 0.9. LOD scores for the various traits were: *Length* 3.81, *Weight* 2.83, *Prestimulus_Percent_Top/Bot_Use* 3.21, *Prestimulus_Total_Movement* 3.10, *Prestimulus_Moves_to_top* 3.49. Gray QTL are suggestive, but not statistically significant. Image created using MapChart.
Figure 2.1
Figure 2.2
Figure 2.3
Figure 2.5
Figure 2.7
CHAPTER THREE

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS FROM THE TRANSCRIPTOME OF AN ORGANISM WITH A WHOLE GENOME DUPLICATION

Kris A. Christensen 1,4,5, Joseph P. Brunelli 2,4, Matthew J. Lambert 3,4, Jenefer DeKoning 3,4, Ruth B. Phillips 3,4, and Gary H. Thorgaard 2,4

1 School of Molecular Biosciences, Washington State University, Pullman, WA 99164-4660

2 School of Biological Sciences, Washington State University, Pullman WA 99164-4236

3 School of Biological Sciences, Washington State University, Vancouver, 14204 NE Salmon Creek Ave. Vancouver, WA 98686-9600

4 Center for Reproductive Biology, Washington State University, Pullman WA, 99164-7520

5 Corresponding author

Email addresses:

KAC: kris.christensen@email.wsu.edu

JPB: brunelli@wsu.edu

MJL: matthew.lambert@wsu.edu

JD: dekoning@vancouver.wsu.edu

RBP: phillipsr@vancouver.wsu.edu

GHT: gary.thorgaard@wsu.edu
Abstract

Background

The common ancestor of salmonid fishes, including rainbow trout (*Oncorhynchus mykiss*), experienced a whole genome duplication between 20 and 100 million years ago, and many of the duplicated genes have been retained in the trout genome. This retention complicates efforts to detect allelic variation in salmonid fishes. Specifically, single nucleotide polymorphism (SNP) detection is problematic because nucleotide variation can be found between the duplicate copies (paralogs) of a gene as well as between alleles.

Results

We present a method of differentiating between allelic and paralogous (gene copy) sequence variants, allowing identification of SNPs in organisms with multiple copies of a gene or set of genes. The basic strategy is to: 1) identify windows of unique cDNA sequences with homology to each other, 2) compare these unique cDNAs if they are not shared between individuals (i.e. the cDNA is homozygous in one individual and homozygous for another cDNA in the other individual), and 3) give a "SNP score" value between zero and one to each candidate sequence variant based on six criteria. Using this strategy we were able to detect about seven thousand potential SNPs from the transcriptomes of several clonal lines of rainbow trout. When directly compared to a pre-validated set of SNPs in polyploid wheat, we were also able to estimate the false-positive rate of this strategy as 0 to 28% depending on parameters used.

Conclusions

This strategy has an advantage over traditional techniques of SNP identification because another dimension of sequencing information is utilized. This method is especially well suited for identifying SNPs in polyploids, both outbred and inbred, but would tend to be conservative for diploid organisms.
Keywords
SNP, polyploid, rainbow trout, genome duplication

Background

Whole genome duplications followed by differentiation of many gene duplicates in the ancestors of rainbow trout and other salmonid fishes [1, 2] have created one of the most complex animal genomes [3-5]. Historical polyploid events have also been important in the ancestry of all vertebrates [6, 7], and are of broad significance in plant evolution [8, 9]. Genome duplications increase the complexity of genomic studies [10] and result in a need to distinguish between paralogs [11, 12]. Population and quantitative genetics fundamentally rely on the accurate identification of the alleles of a gene or locus [13].

Identification of alleles is complicated when paralogous genes or sequences are easily mistaken for alleles [14]. Recent analytical methods have increasingly turned to single nucleotide polymorphisms (SNPs) as the preferred genetic marker to distinguish between alleles because they are ubiquitous and can be readily scored [15]. SNPs are easily mistaken for paralogous sequence variants because many paralogs are similar in sequence and may contain common variants that were present before genome duplication.

Traditional SNP calling programs use sequence alignments of cDNA or genomic sequence and identify any variant position as a possible SNP [16]. If a paralog is included in the alignment, traditional strategies will incorrectly identify paralogous sequence variants as possible SNPs (Figure 1). Traditional approaches of transcriptome data analysis for SNP identification have been utilized for both diploid and polyploid organisms [17, 18].

In the diploid carrot, 18% of the potential SNPs were found to be false-positives (excluding technical failures) using a traditional method of analyzing transcriptomic data [17], and in some fish species the
false-positive rate was very high (e.g. catfish 41.4%) [18]. In fishes with polyploid ancestries, the false-positive rate was 19.4% for whitefish [19], and 15.9% for chum salmon [20].

Two main non-traditional approaches have been used to distinguish SNPs from paralogous sequence variants. The most common strategy has been to remove all polymorphisms associated with genes that have paralogs [21-25]. Such filtering drastically reduces the total number of genes in which polymorphisms can be detected in species with recent polyploidy ancestries and may still lead to a relatively high false-positive rate. A filtering strategy was used for a diploid coral transcriptomic dataset, which resulted in a 30% false-positive rate [26] and a 12% false-positive rate was found for ryegrass using a similar strategy [27]. In polyploid alfalfa, a window approach was used to eliminate SNPs if the potential alleles differed by 10% or more in the window surrounding the SNP [28]. This resulted in a false-positive rate of 37.5%, a very positive outcome considering that the alfalfa was both polyploid and having heterozygous genotypes. In salmonids, filtering resulted in a false-positive rate of 68.6% [29].

Other researchers have proposed strategies to detect polymorphisms without filtering paralogs [30-32]. These methods allow identification of differences between alleles in genes with paralogs, but alleles will still be indistinguishable from paralogs in future studies if the sequence is not reported with the SNP. Knowledge of a SNP at a given location is not fully informative as paralogs may share common sequence with an allele.

The SNP calling method described here involves the analysis of windows of sequence (similar to [28]) and only comparing homozygous alleles differing between individuals (Figure 1). We utilized homozygous clonal lines of rainbow trout because they ensure homozygous alleles at every locus and any difference within an individual likely represents a paralogous sequence variant. This strategy may also be used for SNP identification in outbred organisms, but only SNPs from homozygous loci will be identified.
Results

cDNA Output, SNP Calling, and Paralog Counting

Embryo and head kidney cDNA were sequenced for five homozygous clonal lines of rainbow trout. The average number of reads per clonal line was 114,096 (SD = 67k) for the embryo and 105,464 (SD = 71K) for the head kidney (Table 1). cDNA reads were aligned to the unique rainbow trout UniGene database (build #34) with Megablast version 2.2.21 [33] and with Novoalign version 3.00.05 [34] to compare these two alignment programs.

For each UniGene sequence, the aligned cDNA reads were used to estimate the number of paralogs and identify potential SNPs for both tissues (Table 2). Figure 2 shows the distribution of different categories of gene duplication for the different tissues and the combined result. SNPs were identified in 4,533 of the UniGene reference sequences and 7,412 (Additional file 1) total SNPs were found when both tissues and all the clonal lines were combined (SNP score > 0.25, Indels not included, Megablast alignments). For Novoalign alignments, SNPs were found in 2,979 references, with 5,052 total SNPs identified. The number of SNPs detected between clonal lines of rainbow trout ranged from 792 to 3451 (Table 3).

The number of SNPs between lines appears to be correlated with number of reads, with the lines having higher read counts also having more SNPs (Tables 2 and 3). A value of 0.25 for the SNP score was used as a threshold for reporting based on Table 4 (discussed below), and because an allele only represented by one read would score below this value (discussed in Methods section). Removing these alleles should remove most sequencing errors. The embryo and head kidney tissues had the same gene counts for roughly half of the overlapping UniGenes (2326 of 4782) when using Megablast alignments and 72% (1340 of 1858) using Novoalign alignments.

The different SNP numbers and percent gene duplication (~32% Megablast and ~14% Novoalign), identified while using different alignment programs resulted from how alignments were produced.
Novoalign uses a global alignment algorithm and produced fewer alignments between diverged paralogs (Figure 3). The Novoalign program produces relatively few alignments with lower sequence identity compared to the Megablast program. The Megablast result is similar to an all-against-all EST comparison from Atlantic salmon with greater divergence between ancestral genomes in the rainbow trout [35].

In a general sense, having few low sequence identity alignments indicates that the Novoalign program is more accurate because diverged paralogs are not aligning to the reference sequences. However, our strategy requires the alignment of paralogs so we can distinguish between these sequences and identify SNPs in paralogs that are not represented in the reference sequences.

**SNP validation**

We downloaded cDNA and reference sequences used by another research group to identify SNPs in polyploid wheat [36]. We applied our strategy for SNP discovery to these previously validated UniGenes, allowing for direct comparison of validation rates (Table 4). Variation in window size can produce differing results (Figure 4). Larger window sizes produced fewer false-positives (Table 4), and adopting a SNP score threshold of 0.25 eliminated all false-positives. The false-positive rate varied by alignment program, but in all instances the false-positive rate was lower using the strategy described here than was observed in the original study (Table 4).

There was little variance in SNP validation when comparing different alignment programs. During SNP identification in rainbow trout, nearly twice as many SNPs were identified when Megablast was used. In wheat, SNP identification is virtually the same between alignment programs. It appears from the 125 gene dataset that the genome duplication event was much more recent in wheat than in rainbow trout. For this reason both alignment programs were able to align the highly similar paralogs with similar
performance (Figure 3). It is believed that there was an allopolyploid event in tetraploid wheat less than 0.5 MYA [37], which corresponds to highly similar paralogous sequence seen in this dataset.

**Sequencing validation**

For SNP validation ten sets of primers were also created to amplify genomic DNA of the different trout lines of ten predicted SNPs in sequences downloaded from NCBI. Five of these pairs amplified well and were polymorphic at the predicted site when sequenced. The low amplification rate is attributed to the lack of information related to repetitive sequences and introns surrounding the window in which a SNP was predicted (technical errors). The five that amplified well had scores of: 0.35, 0.35, 0.41, 0.42, and 0.8. The others had scores of: 0.04, 0.37, 0.42, 0.57, and 0.76.

**Discussion**

Traditional methods of SNP identification are effective for standard diploid organisms, but are not appropriate for organisms that have had a genome duplication. Other methods have been employed to overcome traditional limitations [30-32], but they may be of limited use for future studies if they cannot effectively integrate the occurrence of paralogs into SNP identification. Our strategy incorporates such information during SNP calling and may be used for both inbred lines and outbred populations. The importance of alignment program selection increases as paralogs diverge. Early after a genome duplication, it is likely that most alignment programs could effectively align paralogous sequences together. As mutations alter the paralog sequences, high stringency alignment programs will no longer align paralogs together. SNP discovery in diploids benefits from accurate alignments, but information about the paralogs is much more valuable in polyploids than highly precise alignments.

Dependence on a nucleotide window is both a weakness and strength of this strategy. While modifications may be made to increase or decrease the window size, doing so may also influence the results. For example, a decreased window size may limit the ability to distinguish similar paralogs
(Figure 4). Alternatively, an increased window size would allow the identification of all non-identical paralogs, but would decrease the likelihood that a cDNA sequence would span the window completely and would decrease the amount of usable data. Traditional strategies have an effective window size of one nucleotide and for this reason are unable to differentiate between paralogs.

One distinctive component of our strategy is the stringent use of only homozygous alleles in SNP identification. This eliminates many false-positive SNPs that are the result of paralogous sequence variants. At the same time, fewer SNPs will be found because heterozygous alleles are eliminated during SNP identification. To maximize SNP identification, it may be more appropriate to use a low stringency method and use allele frequency as a filter instead of the homozygous selection method described here.

A complementary application of this method might be in the identification of gene families or paralogs. Our method tries to identify paralogs from short cDNA sequences in order to score SNPs, but the estimation may be useful in other ways. For example, the described strategy may be used as an initial method for identifying genes that may be duplicated. Identification and enumeration of paralogs would benefit gene expression, phylogenetic, and gene ontology studies.

Conclusions

The SNP validation evidence suggests that our method offers an advantage over existing strategies of SNP identification for polyploid organisms. This strategy might also be used for analyzing data derived from diploid organisms, but would likely be overly conservative for SNP identification in those organisms, resulting in identification of fewer SNPs.

Methods

mRNA isolation, cDNA synthesis, and 454 pyrosequencing
Total RNA was isolated from tissues of five different clonal lines of doubled haploid rainbow trout. Arlee, Clearwater, and Swanson are YY male lines [38], Skamania is apparently an XX male line with a female-to-male sex reversal mutation (Paul Wheeler, personal communication), and the Whale Rock line is an XX female line [30]. RNA was extracted from approximately 30 eyed-stage embryos (~15 days post fertilization at 11ºC) from each clonal line. The yolk sacs were removed from the embryos in a chilled 0.9% saline solution before RNA extraction. Head kidney RNA was also isolated from an ~one-year-old fish for all five clonal lines.

Tissues were suspended in TRIzol (Invitrogen), flash frozen in liquid nitrogen, and stored at -80°C until RNA extraction. Tissue samples were homogenized using a glass Dounce tissue grinder as recommended in the TRIzol RNA extraction protocol [39]. After TRIzol phase separation, two additional rounds of RNA extraction were performed using a four to one ratio of acidic phenol (pH ~4.5):chloroform, followed by a final chloroform extraction. Total RNA was precipitated with isopropyl alcohol, washed with 70% ethanol, and suspended in water. cDNA was synthesized from each total RNA source using the SMARTer™ PCR cDNA synthesis protocol (Clontech) and was amplified with Advantage PCR according to the manufacturer’s instructions. Amplified cDNAs were purified by phenol-chloroform extraction [40] and precipitated with an equal volume of isopropyl alcohol, washed with 70% ethanol, and suspended in water.

The samples were then sent to the Genomics core at Washington State University where the cDNA quality was assessed using the Agilent DNA 1000 kit (Agilent Technologies). Nebulization of the cDNA followed, using the 454 Rapid DNA Library protocol (Roche). Each cDNA sample was then barcoded and each cDNA library was quantified using a PerkinElmer VICTOR™ X. Libraries were titrated by small volume emulsion PCR and then pooled into two libraries for large volume emulsion PCR. Finally, both libraries were sequenced using a Roche 454 FLX Titanium Genome Sequencer on a single picotiter plate.
Sequence alignment

Barcoded cDNA sequences were separated based on their clonal line of origin and tissue using Roche’s SFF Tools [41] and modified to conform to standard FASTA using custom scripts. They were deposited to the NCBI’s sequence read archive (study accession: SRP028233). The cDNA dataset for each clonal line and each tissue was then aligned to the rainbow trout unique UniGene database build #34 [42] using default settings of standalone Megablast (version 2.2.21). Novoalign (version 3.00.05) was used to create the same alignments with default settings.

The default setting for the number of database sequence alignments to report for Megablast is 250, and the number of high-scoring segment pairs (HSPs) to save per database sequence is unlimited. These settings may produce pseudo-homologous alignments, but hypothetically may be beneficial in SNP discovery because they help us identify regions of sequence where we have little confidence. Sequence regions with low complexity (if not filtered) or high genomic copy number will produce many false alignments, and these regions will appear to have many extra copies when compared to the full reference UniGene sequence. A method for handling these circumstances is described later.

Novoalign uses a global alignment algorithm, while Megablast uses a local alignment algorithm. These two types of alignment strategies were used to determine how much SNP discovery and gene copy estimation relied on the type of alignment program used. It was expected that global alignment algorithms would be better suited for aligning similar sequence, and as a result fewer gene copies would be estimated using this strategy because dissimilar paralogs would not align.

The unique rainbow trout UniGene database contains 142,505 sequences and was generated by clustering publicly available sequences of high similarity together [42]. From each cluster the longest sequence was taken as a representative for that cluster [42]. If copies of a gene are represented in the databases and are of sufficient similarity, they would be clustered together and only one representative
would remain. This is important for estimating the number of genes that have been copied in the reference database. If there were genes with extra copies represented twice, then the estimate would be biased.

**Paralog counting**

Homozygous organisms are expected to have a single copy of every allele, if there is evidence for more than one allele, it likely indicates the presence of a paralog with similar sequence (Figure 4). To estimate the number of copies a gene has, each UniGene reference sequence alignment was examined for unique cDNA types using custom Perl scripts (Additional File 2). Estimating gene copy number was done to improve SNP identification, as described below.

First, a reference sequence was split into overlapping windows, and for each of these windows all unique cDNA types with at least two copies were counted. The database sequence was split into windows because often cDNA sequence reads are short and in order to identify variant cDNA types, they need to cover the same distance. This means that only those portions of cDNA that completely span a window will be used in paralog counting (and in SNP calling later). A window size of 80 nucleotides was used for all rainbow trout analyses. cDNAs that have a sequence similarity of ~99% would on average have one polymorphism in each 80 nucleotide window. cDNA types were identified by looking for differences between the cDNA read and the reference sequence in the alignments. The differences were recorded relative to the reference sequence, making divergent cDNA types comparable. Each cDNA type had a unique pattern of differences from the reference sequence.

Next, multiple windows, representing at least half of the reference sequence, with the same copy number were used as a threshold to reduce the misleading effect of repetitive motifs on the paralog count estimation. If the highest count of cDNA types did not cover 50% of the database sequence (e.g. one window had 10 different cDNA types and all the other windows had 2 different cDNA types), then
the count of cDNA types in those windows was reduced by one and the new number of cDNA types was checked for the threshold. This was done until the count was found in 50 percent of the windows or until the count reached zero.

Fifty percent was used instead of 100% because it is expected that even with perfect coverage, there will be some windows with fewer cDNA types than the true paralog number because paralogs may have complete sequence similarity in those windows.

**SNP calling**

For each of the UniGene sequence windows (described in the previous section), the cDNA types within those windows were compared if all the clonal lines were homozygous for either cDNA type (Figure 1). cDNA types found within a homozygous individual are not alleles and were not compared. The comparisons were then given a SNP score based on several criteria: sequence similarity, number of nearly identical paralogs, number of cDNA types in a window compared to the full UniGene sequence, coverage, even representation of alleles, and the phred quality score at the SNP position. These scores are not probabilities; they reflect the level of confidence that a variant represents a SNP (and not a paralog or one of several classes of errors) based on the previously mentioned criteria. The SNP score representational formulation is:  

\[ \text{SNP Score} = (\text{sequence similarity}^{10}) \times \text{Nearly identical paralogs} \times \text{paralog information} \times \text{sequencing error} \times \text{unequal representation and coverage} \times \text{sequencing quality score}. \]

Each of the elements from this equation are discussed below. The score will range from zero to one.

**Sequence similarity**

The initial SNP score was determined by taking the sequence identity and raising it to the tenth power (Additional file 3). For example, if two cDNA types had 0.99 sequence similarity, they would be given a score of 0.904. A variable threshold was then used to eliminate unlikely comparisons (e.g. < ~0.99 similarity). The sequence similarity was raised to the tenth power because comparisons with modest
sequence similarity (e.g. 0.95, score = 0.599) are assumed to be much more likely due to paralog differences than to allelic differences. The window size will greatly affect this score (Additional file 3).

**Nearly identical paralogs**

If a cDNA type was compared to multiple cDNA types, all comparisons with that cDNA type had their scores lowered. Multiple comparisons may be expected if there is expression of a gene and its paralog in one of the individuals being compared, but only expression of the gene in the other individual. The SNP scores were raised to the power of $(a + 1) \times 10$, where “a” is the number of comparisons with the cDNA type above 1 (Additional file 3). This expression was chosen to effectively eliminate known paralog differences from further consideration if a SNP score of 0.25 was used as a threshold (Additional file 3). It is possible that one of the comparisons would be between alleles, but identifying which would require more information.

**Paralog information**

To further distinguish between SNP and paralog differences, any window with a low or high paralog count relative to the rest of the UniGene reference sequence (estimate was found in the previous section), was scored lower (Additional file 3). If a particular window had more cDNA types than the paralog count estimated over the whole sequence, then it may contain a repetitive motif and any SNPs identified in these windows may be less reliable because of the sheer number of comparisons being made. If there were fewer cDNA types in a window, it may mean there is less information or that it was a conserved sequence and any SNPs found in this window may be shared among paralogs (and thus are not the locus-specific SNPs we are seeking). The modification was different depending on the type of comparison, as explained below.

**Insufficient data for a window**
If the cDNA count for a window was estimated to be zero, then the SNP score was multiplied by the following expression: $1/c$ (Additional file 3). The 'c' was the estimated paralog count for the whole database sequence. We expect to see windows with estimated cDNA counts of zero when there is insufficient data and therefore we are less confident about any SNPs found in these windows. As the number of estimated paralogs increases, our confidence for which paralog is represented in the window, with insufficient data, decreases (e.g. 1/1, 1/2, 1/3 using the equation above).

**Insufficient data for the entire database sequence**

If there were few representations of cDNA types along half of the database sequence (i.e. the paralog count was zero), then the SNP scores were multiplied by $(0.5/b)^2$ (Additional file 3). The 'b' in the equation was the number of cDNA types for the window with the SNP. If the length of a database sequence has poor sequence coverage, but a single window contains ample evidence, we assumed that the window with strong evidence is likely a repetitive motif. The more cDNA types found in one of these windows, the less confidence we had for any SNPs in that window (e.g. 1 cDNA type => 0.25, 2 => 0.06, 3 => 0.03 using the equation above).

**Repetitive gene motifs**

If a window had more cDNA types than that estimated for the whole sequence (and the paralog count was not zero), the score was multiplied by $(c/b)^2$, or $(b/c)^2$ if the window had fewer cDNA types (Additional file 3). If a window contains more cDNA types than the whole database sequence, then it may mean that this window contains repetitive motifs or conserved sequence. Both interpretations decrease the confidence that a SNP found in a window with high numbers of cDNA types is real (e.g. 2 copies in window: 1 copy for whole gene => 0.7, 3:1 => 0.6 using the equation above). Low coverage and increased sequence similarity between paralogs may cause windows to have fewer cDNA types than the rest of the database sequence. This decreases the likelihood that a SNP is real in windows with fewer cDNA types.
**Sequencing error**

To distinguish SNPs from sequencing errors, the score was reduced by 75 percent if there was only a single representation (i.e. one cDNA read) of an allele (Additional file 3). If the scores were not reduced, the majority of the high scoring SNPs (> 0.25) were those with a single representative (data not shown). While these SNPs may be valid, they lack enough evidence to be considered high-confidence.

**Unequal representation and coverage**

The SNP score was also lowered if there was an overabundance of one allele compared to the other. It is assumed that alleles will have roughly similar expression profiles, and many sequencing errors will be unequally represented. The SNP score was multiplied by \((\frac{1+(x+y)}{2})/2 * \frac{4+x}{y}/5\), where 'x' is the number of the low count allele and 'y' is the number of the high count alleles (Additional file 3). This equation was used because it incorporates coverage and even representation, with deviations from equal representation lowering the score. At a lower coverage, unequal representation causes a greater deviation than it would at higher coverage (Additional file 3). We expect most sequencing errors to be at lower coverage.

**Sequencing quality score**

Every SNP score was modified by multiplying the score by the average of the sequencing quality scores (of all cDNA reads) assigned to the nucleotides at the SNP's location divided by the best score possible. If this score is near one, then it will not appreciably modify the SNP score, but if the sequencing quality scores are low, the SNP score will be proportionally lowered. If the sequencing score is low, then it is likely that the proposed SNP is actually a sequencing error and not real. Insertions and deletions were not considered candidate SNPs.

**Alignment Program Comparison**
Novoalign and Megablast use different alignment algorithms. To test how this affects how many alignments are produced for our dataset, we examined the combined tissues of the Clearwater clonal line. After the reference sequences were broken into windows as described above, we used custom Perl scripts to count the alignments in each window. Every alignment was binned into an appropriate sequence identity category based on the number of mismatches in that window. The same was done for the wheat dataset, except the data from the two wheat lines were combined after binning.

**SNP validation**

We downloaded cDNA sequences (NCBI sequence read archive files ERR045179 and ERR045180) and reference sequences (*Triticum aestivum* UniGene build #62) used by another research group to identify SNPs in polyploid wheat [36]. The strategy they used accounted for the polyploid nature of wheat, but they had an effective window size of one nucleotide. We then used custom Perl scripts to retrieve the sequences used in their SNP validation analysis (both the correctly identified SNPs and those that were monomorphic) from the downloaded UniGene database. These 125 sequences were used as a reference set. The cDNA sequence sets were transformed into FASTQ files using the NCBI’s sratoolkit version 2.1.9 [43] and paired-end reads were split into separate files at the same time. Fastq files were converted to fasta files with quality scores using custom Perl scripts for use with Megablast. Megablast and Novoalign were then used to align the cDNA sequences against the reference set.

Megablast uses a local alignment algorithm, Smith-Waterman, while Novoalign uses a global alignment algorithm, Needleman-Wunsch. Local alignment algorithms are generally used for dissimilar sequence alignment, while global alignment algorithms are more for similar sequence alignment. Default settings were used for Novoalign and Megablast. The different alignment program outputs were in SAM format or BLAST format. Custom scripts described above were used to identify potential SNPs. Perl scripts were used to compare the identified SNPs to the validated SNPs.
**Sequencing Validation**

Ten primers were designed [44] surrounding ten proposed rainbow trout SNPs with a range of scores (0.04 - 0.8 using Megablast alignments) to see if the genomic sequence corresponded to cDNA sequence. These proposed SNPs were selected to minimize insertions/deletions in between any known paralogs because they could have interfered with sequencing. The primers were found in ten reference sequences downloaded from the NCBI (BT045031, BT045191, FR677582, AF042218, AB208024, AJ627208, AY029216, AYS93999, AY370888, AY065837). The primers were checked against a preliminary draft of the rainbow trout genome to make sure they did not align to splice sites (obtained from Michael Miller, University of California, Davis).

**Availability of Supporting Data**

cDNA sequences were deposited in the sequence read archive of the NCBI (http://www.ncbi.nlm.nih.gov/sra/?term=SRP028233). Perl scripts and instructions on how they may be used are available in Additional file 2.

**Competing Interests**

The authors declare that they have no competing interests.

**Authors' Contributions**

KAC wrote the Perl scripts, conducted the bioinformatics analyses, and drafted the manuscript. JPB prepared cDNA from clonal lines and helped draft the manuscript. MJL and JD sequenced candidate SNPs. GHT and RBP conceived of and initiated the study and helped draft the manuscript.

**Acknowledgements**

We thank Paul Wheeler for producing the clonal rainbow trout. Thanks to Mark Wildung for 454 sequencing and providing the methods used for 454 sequencing. This research was funded by a USDA special research grant titled “Aquaculture, Idaho and Washington” awarded to GHT and RBP and in part
by Agriculture and Food Research Initiative Competitive Grants number 2009-35205-05067 and 2011-67015-30091 from the USDA National Institute of Food and Agriculture. Support for KAC came from the National Institute of General Medical Sciences, Award Number T32GM083864. Finally, we thank Bobbi Adams, Anna Heinke, Stacy Hathcox, Jon Mallatt, and the anonymous reviewers for their comments on the manuscript.

References


27. Studer B, Byrne S, Nielsen RO, Panitz F, Bendixen C, Islam MS, Pfeifer M, Lübberstedt T, Asp T: **A transcriptome map of perennial ryegrass (Lolium perenne L.).** *BMC Genomics* 2012, **13:**140.


34. Novocraft Technologies [www.novocraft.com]


41. 454 sequencing [www.454.com]


43. The Sequence Read Archive (SRA) [www.ncbi.nlm.nih.gov/sra]


Figures

**Figure 1** - A comparison between traditional approaches and our strategy of SNP identification

Panel A shows two homeologs (duplicated chromosomes) and the sequence of both alleles from each chromosome. The fish are homozygous and only have one allele from each homeolog. This means that SNPs can only be found by comparing sequences between the two homozygous fish. Any variation found within a fish is a paralogous sequence variant (i.e. the first nucleotide in the gene sequence). Panel B illustrates the differences between traditional approaches and ours. The traditional approach would scan every nucleotide position and identify any variant as a possible SNP. This leads to a 50% error rate for the illustrated case (SNP1 is actually a paralogous sequence variant). Our approach uses windows of sequence to identify potential SNPs. Variant windows, in this case 9 nucleotides, are compared between individuals, with windows found in both individuals not considered for SNP identification.
Figure 2 - UniGene copy number distribution in rainbow trout embryo and head kidney tissues.

The graph shows the percent of UniGenes that have evidence for specific copy numbers, from a single copy gene to ten or more copies. The alignment program used was Megablast with default settings.

Figure 3 - Alignment comparison

Panel A is a histogram of sequence similarity of alignments for the Clearwater clonal line. Alignments were produced by either Megablast or Novoalign and for every 80 nucleotide window the alignments were binned into their respective similarity category based on the number of mismatches in the window. The second peak of alignments around 92.5% with Megablast likely represents a set of genes duplicated during a whole genome duplication. Novoalign does not show this pattern because Novoalign uses a more stringent alignment algorithm. Panel B was produced in the same way, but it represents the combination of two polyploid wheat lines for 125 UniGenes used in SNP validation. The second peak for this panel is around 98.75% indicating a much more recent whole genome duplication. Both alignment programs have similar results with wheat because the paralogs have high sequence similarity.

Figure 4 - Effect of window size on distinguishing paralogs and the amount of useable data

Panel A shows two homeologs and the sequence of paralogs from a homozygous fish. Paralogous sequence variants are in bold text. Two windows are represented (1 and 16), only the 16 nucleotide window can differentiate between the paralogs. Panel B shows a multiple sequence alignment of the expressed cDNA reads aligned to the gene sequence from Chrom 1a. Depending on the alignment criteria and the sequence similarity, the two different paralogs will both align to the sequence from Chrom 1a. The number of useable sequences depends heavily on the window size. The larger the window size, the less likely that a read will span the window length, but the more likely the paralogs will be distinguishable.
Figure 3.1
Figure 3.2

- Embryo (11386 Unigenes)
- Head Kidney (7021 Unigenes)
- Both (16225 Unigenes)
Figure 3.3

A

B

Percent Identity

Count

Megablast

Novoalign
Figure 3.4

A

Chrom 1b

Chrom 1a

TGGCAATCGACCGACTAG
TGGCATTCGACCGGCTAG

Window Size  Paralog?

1  No

16  Yes

Gene Sequence

B

TGGCAATCGACCGACTAG
TGGCAATCGACCGACTAG
TGGCAATCGACCGAC
TGGCATTT
CATTCGACCGGCTAG
CGACCGGCTAG
AATCGACCGACTAG

Reads

Evidence

4 reads

2 reads
Tables

Table 3.1 - Counts of cDNA reads from embryo and head kidney tissues from five rainbow trout lines.

<table>
<thead>
<tr>
<th></th>
<th>Embryo</th>
<th>Head Kidney</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlee (YY)</td>
<td>96077</td>
<td>95196</td>
<td>191273</td>
</tr>
<tr>
<td>Clearwater (YY)</td>
<td>231981</td>
<td>108170</td>
<td>340151</td>
</tr>
<tr>
<td>Skamania (XX-male)</td>
<td>64002</td>
<td>126150</td>
<td>190152</td>
</tr>
<tr>
<td>Swanson (YY)</td>
<td>90235</td>
<td>12</td>
<td>90247</td>
</tr>
<tr>
<td>Whale Rock (XX)</td>
<td>88186</td>
<td>197793</td>
<td>285979</td>
</tr>
<tr>
<td>All</td>
<td>570481</td>
<td>527321</td>
<td>1097802</td>
</tr>
</tbody>
</table>

Table 3.2 - Gene duplication counts and SNPs for rainbow trout embryo and head kidney tissues

<table>
<thead>
<tr>
<th></th>
<th>Single Copy UniGenes</th>
<th>Multiple Copy UniGenes</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>5170 (7923)</td>
<td>754 (3458)</td>
<td>3337 (5034)</td>
</tr>
<tr>
<td>Head Kidney</td>
<td>2566 (5101)</td>
<td>399 (1906)</td>
<td>1376 (2284)</td>
</tr>
<tr>
<td>Both</td>
<td>7289 (11100)</td>
<td>1234 (5113)</td>
<td>5052 (7412)</td>
</tr>
</tbody>
</table>

The number of single and multiple copy UniGenes was estimated using the reads from all clones from the different tissues (The highest copy number from all individual clonal lines was used as the combined estimate for a UniGene). UniGenes without sufficient evidence of copy number are not represented. The SNP counts were found by comparing all clones against each other at the same time. These SNPs have a SNP score of 0.25 or greater, meaning that all of them have more than a single read representing each allele. The first set of numbers are from Novoalign alignments and the second set are from Megablast (in parentheses).

Table 3.3 - A comparison of SNPs between clonal rainbow trout lines for the combined tissues

<table>
<thead>
<tr>
<th></th>
<th>Arlee</th>
<th>Clearwater</th>
<th>Skamania</th>
<th>Swanson</th>
<th>Whale Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlee</td>
<td>-</td>
<td>2292</td>
<td>1460</td>
<td>1180</td>
<td>1628</td>
</tr>
<tr>
<td>Clearwater</td>
<td>(3300)</td>
<td>-</td>
<td>1425</td>
<td>1283</td>
<td>2412</td>
</tr>
</tbody>
</table>
The number of SNPs with SNP scores of 0.25 or greater using sequences from both embryo and head kidney tissues between the clonal lines of rainbow trout. The alignments, used to identify these SNPs, were produced using Novoalign with default settings (top) and Megablast (bottom in parentheses).

**Table 3.4** - SNP validation comparison using a transcriptome dataset from polyploid wheat.

<table>
<thead>
<tr>
<th></th>
<th>Original Study Window 80, SNP Score &gt; 0</th>
<th>Window 80, SNP Score &gt; 0.25</th>
<th>Window 40, SNP Score &gt; 0</th>
<th>Window 10, SNP Score &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validated SNPs</td>
<td>64</td>
<td>40 (37)</td>
<td>28 (22)</td>
<td>35 (34)</td>
</tr>
<tr>
<td>False-positives</td>
<td>61</td>
<td>3 (6)</td>
<td>1 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>False-positive rate</td>
<td>49%</td>
<td>7% (14%)</td>
<td>3% (0%)</td>
<td>5% (3%)</td>
</tr>
</tbody>
</table>

The number of SNPs from the original study and from different settings using our strategy. There are multiple numbers per parameter explored. These reflect different alignment programs used under these settings. The first set of numbers were from Novoalign alignments, and the second were from Megablast (in parentheses).
Additional files

**Additional file 1 — SNPs Found Between Five Rainbow Trout Lines in Rainbow Trout**

SNPs found between five rainbow trout clonal lines in UniGene build #34 reference sequences. This file contains the sequence and genotype information for the SNPs found using the Megablast alignment program. This file can be opened with a text editor or with a spreadsheet program, information is separated by tabs.

**Additional file 2 — Perl Scripts and Instructions**

This file must be decompressed before it is opened. It contains a folder with the various Perl scripts used for SNP discovery and instructions on how they may be used. The instructions may be opened with a text editor (e.g. notepad ++).

**Additional file 3 — SNP Score Generation**

Panel A illustrates the initial SNP score based on sequence similarity for three different nucleotide window sizes. In all cases the initial SNP score decreases quickly as the number of mismatches increases. Panel B shows how three different initial scores are modified if evidence for a nearly identical paralog is found. In all cases the score is lowered below a SNP score of 0.25. Panel C represents SNP scores after paralog information is analyzed. The X axis represents the number of estimated paralogs for a given window, while the different categories represent different numbers of estimated paralogs for the entire reference sequence. Panel D characterizes how the SNP score is modified by coverage and differential representation of the alleles. The X axis represents the coverage of allele 1 and the different categories represent the coverage of allele 2.
CHAPTER FOUR

SIGNATURES OF ANCESTRAL VERTEBRATE GENOME DUPLICATIONS DETECTED FROM TRANSCRIPTOME DATASETS

Abstract

Genome duplications are thought to be important for adding novelty to a genome and to have played a large role early in vertebrate evolution. We developed a new method for detecting genome duplications from transcriptomes sequenced with high-throughput technologies. With our strategy, we were effectively able to evaluate and confirm, using open-access transcriptome data, the results and conclusions from a large body of previous research. With transcriptome data accumulating at a great pace, this approach offers an inexpensive and effective strategy for analyzing genomic duplications in species that have yet to be characterized.

Introduction

Early in the history of vertebrates, two whole genome duplications are thought to have created diversity in the genes seen in contemporary descendants (Ohno 1970, reviewed in Otto and Whitton 2000, Otto and Yong 2002). Some ancestral gene copies, like the hox genes, can still be observed in the genomes of many vertebrates (reviewed in Van de Peer 2004). Additional genome duplications have played an important evolutionary role in several families of fishes (reviewed in Otto and Yong 2002, Koop et al. 2008, Kassahn et al. 2009, Wang et al. 2012) and a variety of other organisms (Ohno 1970, reviewed in Otto and Whitton 2000, reviewed in Otto and Yong 2002, Aury et al. 2006).

Extensive genome duplications identified in plants are thought to have been important for their vast species diversity (Otto and Whitton 2000, reviewed in Hegarty and Hiscock 2008). Many crop species are thought to have had relatively recent genome duplications (reviewed in Otto and Whitton 2000, Schlueter et al. 2004). The wide occurrence of genome duplication across taxa has created a broad interest in the study of evolution by genome duplication.
Many researchers have identified duplicated gene pairs that have been retained in organisms with ancestral genome duplications (Van de Peer 2004, reviewed in Séimon and Wolfe 2007, Kassahn et al. 2009). One goal of such research is to identify the duplicated pairs commonly retained in diverse organisms or from multiple events (Scannell et al. 2007, Kassahn et al. 2009, Buggs et al. 2012). A fundamental understanding of how and why certain duplicated genes pairs are retained may lead researchers to a better understanding of speciation and the origin of evolutionary novelty (reviewed in Otto and Whitton 2000, Scannell et al. 2007, reviewed in Séimon and Wolfe 2007, Ilut et al. 2012).

Interpretations from genomic research can be complicated by the existence of both segmental and whole genome duplications. Duplicated regions of the genome may often be mistaken for each other--creating problems during single nucleotide polymorphism (SNP) identification (Trick et al. 2008) and gene expression studies (Ilut et al. 2012). During SNP discovery, similar sequences of DNA or cDNA are aligned and variations in these aligned sequences are called as SNPs. If there are multiple regions in the genome with similar sequences, differences between these regions may be incorrectly called as SNPs (Trick et al. 2008, Christensen et al. in press). Likewise, duplicate copies of a gene may be expressed at the same time, and the expression of these copies may be measured incorrectly as being coded by a single gene. For these reasons, it is critically important to identify which genes have been duplicated.

If genomic resources are available for an organism, several strategies may be used for distinguishing between gene duplicates (reviewed in Van de Peer 2004). For example, if ancestral diploid genomes are available for allopolyploids, short cDNA sequences can be aligned to both of them and copies can be distinguished based on sequence identity to the different genomes (Ilut et al. 2012). Genomic sequence can by itself be used to identify and differentiate duplicated genomic regions (reviewed in Van de Peer 2004).
Previous attempts to identify gene copy pairs from cDNA transcripts often involved all versus all comparisons (where a set of sequences is aligned to itself) of a single dataset with long sequences (Blanc and Wolfe 2004, Wang et al. 2012). With these types of comparisons, the research goal is normally to find out how much of the genome is duplicated and when the duplication(s) occurred. The count of non-identical alignments can be used to estimate how much of the genome is duplicated (reviewed in Otto and Yong 2002, Blanc and Wolfe 2004, Aury et al. 2006). Synonymous mutations between the alignments produced during the all versus all comparisons are used to estimate when the majority of the copies were produced (Blanc and Wolfe 2004, Shoemaker et al. 2006, Wang et al. 2012). In order to increase the accuracy of these studies, the alignments often need to pass length and sequence similarity thresholds (Blanc and Wolfe 2004, Schlueter et al. 2004, Wang et al. 2012).

Many of these thresholds are longer than sequences produced by next generation sequencing; this can make modern transcriptome datasets unsuitable for estimating how much of the genome is duplicated unless quality contigs can be produced. Currently-produced transcriptome sequences, used without thresholds, would result in many alignments of repetitive sequences (either motifs or repetitive elements). Alignments of such repetitive sequences including repetitive elements may artificially inflate the estimation of how many genes have multiple copies. If these alignments including repetitive elements were used to estimate the amount of time since the last genome duplication, the analysis would likely over-estimate the time since the genome duplication or may create enough noise that a genome duplication would be impossible to identify.

One solution to the problem of short sequences would be to assemble quality contigs. For a polyploid, short sequences are often impossible to assemble into a complete or accurate contig because of high sequence identity between different genomic locations (Wang et al. 2012, Gruenheit et al. 2012). Paralog chimeras, created by assembling partial cDNA sequences from duplicated genes together into a
single contig, are expected when paralogs have high sequence identity, and would be very difficult to avoid using traditional assemblers (Gruenheit et al. 2012). Chimeras would likely reduce the estimated number of duplicated gene pairs found in a study utilizing contigs. The effect of chimeras on the estimation of when a genome duplication occurred probably is not straight-forward because the noise created would be random.

An approach is needed to prevent assembly of chimeras from paralogous sequences if the high-throughput transcriptome datasets are to be effectively utilized for quantifying genome duplication. Wang et al. (2012) may have had some success assembling transcriptomic sequences (454 platform, mRNA sequences, and ESTs from the common carp). First they identified 20 pairs of copied genes and found the sequence identity between each pair. They used the highest sequence identity found (97%) as a parameter in assembly, where overlapping regions would not be assembled if they were below that sequence identity. This strategy would eliminate chimeras if the paralogs uniformly accumulated enough mutations to have lower sequence identities than the threshold. If the mutations that have accumulated after the genes were duplicated were unevenly distributed, resulting in fewer mutations among some duplicate pairs, then chimeras are expected.

We present a bioinformatic method for analyzing high-throughput transcriptomic data to estimate the number of genes that have been duplicated in multiple organisms without using traditional assembly. Our results were consistent with most of the previously proposed genome duplication events.

**Materials and Methods**

the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA). UniGene databases were downloaded for the species available (Gallus gallus - Build 44, Xenopus laevis - Build 94, Xenopus tropicalis - Build 51, Carassius auratus - zebrafish dataset Build 125, Gasterosteus aculeatus - Build 7, Danio rerio - Build 125, Plecoglossus altivelis - rainbow trout Build 34, and Oncorhynchus mykiss - Build 34). The Japanese eel transcriptome was aligned to a predicted gene database of the European eel (version 1-1) from zfgenomics (zfgenomics) and the sea lamprey transcriptome was aligned to an Ensembl lamprey cDNA database (version 7.0.73). Alignments were produced between the transcriptome databases and the respective UniGene databases using discontinuous MegaBLAST version 2.2.28+ (Altschul et al. 1997).

Gene copy number was estimated using a strategy presented in Christensen et al. (in press). Briefly, perl scripts (slightly modified for BLAST+ 2.2.28 and available upon request) were used to break up alignments into overlapping windows for each UniGene sequence (Figure 1). For each window, unique sequences (i.e. sequences with the same variant patterns) were counted if there was more than one representative. The highest count was considered the gene copy number if it was seen in 50% of a gene. Otherwise, the highest count number was reduced by one and again checked for 50% coverage. This was done until a count number was identified or the count reached zero. Genes with a count of zero were not reported.

The number of reads and alignments was collected using custom perl scripts (available upon request) from alignment files produced by discontinuous MegaBLAST. The number of reads may not match the number reported for a transcriptome in SRA because alignments would run out of memory (~16 GB) depending on their size (a weakness of using MegaBLAST). The number of alignments were reported based on window size. If an alignment covered multiple windows, it would be reported multiple times. For example, if an alignment was 600 nucleotides long (possible if 454 technology was used), ten
nucleotide windows of 60 nucleotides could fit in this alignment (and would be reported that many times, though each window may have a different value).

Results

The average number of sequences per transcriptome that were used in this analysis was around 10 million reads, and the average number of alignments to all windows in a database was 12 million (Figure 2). In some cases, the number of reads greatly exceeds the number of alignments (e.g. Lamprey ~4:1). Illumina sequencing technology was used in both cases where the reads exceed the number of alignments. The average ratio of reads to alignments was 0.40 for Illumina reads if the chicken and lamprey datasets were excluded, and 0.43 for 454 reads.

The number of genes represented in this analysis was strongly correlated (r = 0.42) with the number of alignments (Figure 2). The average number of genes used per database was 5363 with a range from 288 to 14737. The mass of the different species' genome generated by various labs (Robinson et al. 1975, Thiébaud and Fischberg 1977, Rasch 1985, Ojima and Yamamoto 1990, Ronchetti et al. 1995, Vinogradov 1998, Collares-Pereira and Moreira da Costa 1999, Animal Genome Size Database) is shown in Figure 3.

The distribution of gene copy number is reported in Figure 3 for the different transcriptomes along with information about when the proposed genome duplication occurred. The first genome duplication is thought to have occurred before the split between jawed and jawless vertebrates and is referred to as the 1R duplication for one-round (reviewed in Sidow 1996) and so the sea lamprey is labeled as 1R in Figure 3. The second genome duplication in vertebrates occurred before the split between the lobe-finned (e.g. coelacanth) and ray-finned fish (e.g. goldfish) and is referred to as 2R for two-rounds (Sidow 1996), the chicken and Xenopus tropicalis belong to this category.
A ray-finned-specific genome duplication is thought to have occurred around 350 million years ago (reviewed in Meyer and Van de Peer 2005), referred to as 3R. The three-spined stickleback, Japanese eel, zebrafish, *Xenopus laevis*, and Ayu belong to this category. Although *X. laevis* experienced such an additional round of genome duplication, it is thought that it occurred separately from the ray-finned fishes around 50 million years ago (Evans et al. 2004). A fourth genome duplication in the rainbow trout (Allendorf and Thorgaard 1984) and goldfish (Ohno et al. 1967) was more recent and is referred to as 4R.

In Figure 3 the different categories are displayed along with information about copy number. The copy number was only reported for the first five copies. The various organisms fit well into their expected categories except for the three-spined stickleback and the Japanese eel. Based on our analysis of estimated gene copy number, these two appear as if they would fit better into the 2R category, though evidence for a whole genome duplication in stickleback exists (Cresco et al. 2003, Parmar and Wright 2013).

Looking at the distribution of alignments for the stickleback and the Japanese eel, it can be seen why they tend to fit into the 2R category when looking at gene copy number (Figures 3 and 4). The stickleback has a peak around 98.3% sequence identity, and the Japanese eel has a tail ranging from 73.3% to 55% sequence identity (Figure 4). With such high similarity between paralogs in the stickleback, our method fails to distinguish between paralog and allele, resulting in an overrepresentation of single copy genes. With the Japanese eel, many of the alignments are dissimilar enough that we suspect that alignments are failing to be produced for the majority of a gene. Without at least half of a gene being represented, it will not be considered in the analysis. This would result in an overrepresentation of single copy genes with high sequence identity.

In Figure 4, 70% of the sea lamprey’s alignments were with 100% similarity, with very few alignments noted in the remaining categories. If there was a whole genome duplication in the ancestor of the sea
lamprey, the remaining duplicates may be in the 85% sequence identity range or our analysis was unable to detect the duplication. There was a very small peak near 85% for all of the organisms in Figure 4. These sequences may have come from segmental duplication events, though.

A possible signal from the 2R genome duplication was seen in both the 2R organisms and 3R organisms around the 98.3 sequence identity region (Figure 4). This signal was not seen in the sea lamprey (1R), though it may be difficult to actually distinguish between the ancestral duplication and modern segmental duplications. Our prediction would be that ancient, ancestral duplications would have much less sequence similarity.

**Discussion**

With our strategy, we were effectively able to use open-access transcriptome data to reproduce the results and conclusions seen in a large body of research. With such data accumulating rapidly, this strategy offers a cheap and effective approach to broadly analyze genomic duplications of species that have been otherwise uncharacterized. This strategy may be used in time-series and tissue-specific analyses as well to characterize duplicated genes.

The majority of the predicted genome duplications were reaffirmed with our strategy for the various organisms we included in this study. These included the low amounts of duplicated genes that would be predicted in the 1R sea lamprey, though curiously there was not a large signal from an ancestral duplication detected in this analysis. This suggests that either the 1R duplication did not occur or that our strategy is not sensitive enough to detect extremely ancient duplications.

The chicken and the western clawed frog (*Xenopus tropicalis*) showed similar amounts of gene duplication, both greater than that seen in the sea lamprey. These organisms had a slight peak at 85% sequence identity like the sea lamprey, but additionally showed a peak around 98.3% absent in the sea lamprey that may correspond to the 2R genome duplication. Further research needs to be performed in
order to distinguish if these alignments originated from an ancestral (2R) whole genome duplication or from rounds of recent segmental duplications. The similarity between the sequences supports the model that these matches derive from segmental duplications, but because this peak was not seen in the lamprey it may also be from the second round of genome duplication.

Evidence from the stickleback transcriptome also suggests that the majority of duplicated genes from the 3R genome duplication have similar sequences, with a larger peak of alignments at the 98.3% sequence identity mark. These sequences were similar enough that the strategy for identifying gene duplications here was unsuitable for distinguishing between alleles and paralogs. In this case it favored alleles and rejected paralogs.

These peaks at 98.3%, if real, may indicate that genes retained in the stickleback genome also tend to have conserved sequence. This was not seen in any of the other 3R species. The trend of accumulating mutations as predicted by Ohno (1970) and Force et al. (1999) was seen for the ayu, zebrafish, and African clawed frog (Xenopus laevis). The zebrafish had a wide peak of alignments around 83.3% (data not shown) and had more gene copies than 2R organisms. Surprisingly, the Japanese eel showed a similar number of duplicated genes as the stickleback, but a different alignment profile, with the majority of alignments after the 2R peak (i.e. the peak at 98.3% seen in all 2R and 3R organisms) being evenly distributed. With the Japanese eel transcriptome being aligned to the European eel database, it makes this result uninterpretable because the variance of these alignments might come from nucleotide variation between species and not from paralogous sequence variation.

The 4R genome duplications predicted in the rainbow trout and goldfish were both strongly supported by our data. Both organisms had more gene copies than 3R organisms and the alignment pattern of rainbow trout was similar to the zebrafish except the peak was around 85% (rather than 83.3%) and was
wider (data not shown). These results support the use of this strategy in identification of historic gene duplication events.

In the future, we would like to analyze a time course of transcriptomes for a variety of tissues using this strategy. If properly modified, a similar strategy might be used to identify duplicate genes with enough confidence to distinguish between large regions of a paralog. If this was accomplished, paralog-specific expression could also be measured from transcriptome data. Nomenclature projects would benefit by knowing the number of copies of a gene there were in a genome.

Acknowledgements

Support for KAC came from the National Institute of General Medical Sciences, Award Number T32GM083864 and Washington State University.
References


Figures

**Figure 4.1.** Illustration of Methodology. A polyploid organism will have homeologs and from each homeolog set two alleles for a gene are possible. Our strategy was to use nucleotide windows (in this case 60 nucleotides) to look for variants of a gene and thereby identify the possible alleles of each homeolog set. Windows were used because modern sequencing tends to sequence partial segments of a gene. Also, in order to know that a comparisions is valid, the two sequences must fully overlap. For the 780 nucleotide gene, there would be 13 windows. If an allele had a sequence similarity of 99.5%, we would expect to detect (in the form of a sequence variants) that allele in 4 (31%) windows. Our threshold for paralog counting is set to 50% to avoid counting alleles. In this example, the allele would have to have a sequence similarity of < 99.2% to be called as a paralog. It seems unlikely that the variation between alleles would be evenly distributed over the entire gene though, and effectively the similarity would have to be even lower than the theoretical threshold.

**Figure 4.2.** Relationship Between Alignment Number and Genes used in Analysis. Panel A is a graph of the number of reads and alignments used for the different organisms. The reads were downloaded from NCBI's SRA, and the alignments were created using discontinuous MegaBLAST (version 2.2.28+). Panel B represents the number of genes where there was enough alignments (i.e. alignments covered at least 50% of the gene) to use in further analysis. After the name of the different organisms (in parentheses) is the window size used to look for alignments. The number was chosen based on the initial length of SRA reads (if a read sequence was less than 60 nucleotides, a smaller window was used).

**Figure 4.3.** Gene Duplications in a Variety of Organisms. This is a histogram of the percent of genes represented in five gene copy categories. The first category represents no duplications, and the others represent varying levels of duplication. In the first category, the number of historic genome duplications (from reviews) are given. The genome size is also given for available organisms (values for chicken come from domesticated chicken and genome size of the Japanese eel came from values reported for the European eel).

**Figure 4.4.** Distribution of Sequence Identity for Select Organisms. This is a histogram of the sequence alignments with the most similar sequences on the left. The alignments are from windows smaller than the original alignments. The original alignment may be represented more that once if it overlapped with multiple windows. The number of genome duplications (e.g. 1R - one round) thought to be associated with the organisms is given on the left side. The "Stickleback peak" is a peak only seen in the stickleback and in none of the other 3R (three rounds of whole genome duplication) organisms. The "Japanese eel tail" was only seen in the Japanese eel and none of the other 3R organisms.
Figure 4.1

Heterozygous

Homozygous

Chromosome 1b

Chromosome 1a

60 nucleotides

780 nt

Allele Similarity = 99.5 %
Paralog Similarity = 98.7%

<table>
<thead>
<tr>
<th>Count</th>
<th>Percent of gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
</tr>
</tbody>
</table>
Figure 4.2

A

Million

B

Thousand

- Read Number
- Alignment Number
- Number of Genes Represented

Lamprey (60 nt) 0.61
Chicken (60 nt) 0.288
Xenopus tropicalis (60 nt) 0.499
Three-spined stickleback (60 nt) 1.593
Xenopus laevis (60 nt) 0
Av (60 nt) 0
Goldfish (60 nt) 0
Rainbow Trout (60 nt) 0

86
Figure 4.3

Percent of Genes

Number of Copies

Genome Size

2.12 pg  Sea Lamprey
1.25 pg  Chicken
1.74 pg  Xenopus Tropicalis*
0.58 pg  Three-spined stickleback
1.67 pg  Japanese eel
1.78 pg  Zebrafish*
3.11 pg  Xenopus laevis*

Ayu*

1.88 pg  Goldfish
2.60 pg  Rainbow Trout
Figure 4.4
CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Over 500 high-throughput sequencing efforts (94 RNA, 465 DNA) have used rainbow trout as their model (SRA) and over 9,000 abstracts result from a NCBI query search of rainbow trout (NCBI). The interest in rainbow trout stems from their popularity as a game and food fish and their endangered species act listing (along with Pacific salmon). As a research species, the rainbow trout offers many advantages (Thorgaard et al. 2002). The salmonids are one of the few ancestral tetraploid vertebrates in the animal kingdom with a high incidence of functional duplicate genes, creating both a challenge and opportunity for researchers. In this body of research, the challenges of tetraploidy were addressed in efforts to understand gene copy variation, nucleotide variation, and the effects of hatcheries on antipredator behavior.

Hatcheries play an important and valuable role in maintaining large Pacific salmon and steelhead migrations. Ninety percent of the salmon and 88% of the steelhead caught in the Columbia River originate from hatcheries (WDFW). With these large supplementation efforts, the use of hatcheries has come under scrutiny in the last 40 years (Mobrand et al. 2005). The controversy largely centers on consistently observed alterations of traits between hatchery and wild fishes and questions about whether hatchery fish are negatively impacting the conservation of wild fishes (Neely et al. 2012, Skaala et al. 2012, Pulcini et al. 2013).

Evidence presented in Chapter Two and elsewhere (Jackson and Brown 2011) suggests that antipredator behavior and size (reflecting growth rate) have heritable components. Two clonal rainbow trout lines derived from recently wild populations showed significant differences in both antipredator behavior and size compared to hatchery-derived clonal lines. Based on QTL analyses presented in Chapter Two, both traits are partially controlled by several genomic loci. QTLs in a homologous chromosomal region also
affect behavioral traits in the zebrafish. Selective pressure on these loci may be to some extent responsible for the differences between wild and hatchery fishes.

Selective pressure differs between a hatchery and the wild, and as a corollary, alleles that are maladaptive to those pressures would be reduced in a population (Jackson and Brown 2011, Vasemägi et al. 2012). Selection on one trait may alter several more. Hatchery managers are thought to have selected for larger and faster growing individuals (Brannon et al. 2004) and may have inadvertently altered several other traits in the process. In Chapter Two, it was observed that some behaviors and size were correlated (from $r = 0.19$ to $-0.44$).

In certain environments and conditions, smaller size may be advantageous in predator avoidance and escape once attacked by a predator (reviewed in Abrahams 2006). A smaller fish may hide better, be of less interest to a predator (less energy available in small prey), and be more difficult to catch (Abrahams 2006). The behavioral differences between hatchery and wild fish may be intertwined with the size difference between these fish.

Brannon et al. (2004) suggested that hatcheries are not responsible for domestication, rather that the management of these hatcheries was altering wild traits (e.g. selecting for faster growing fishes). Traditional large-scale hatcheries alter so much of the environment that adaptation would be expected after only a few generations. In line with this view, evidence-based hatchery design and management could substantially reduce or eliminate the effects of a hatchery environment (Flagg and Mobrand 2010).

In regards to this proposition, it was observed in Chapter Two that the different clonal lines maintained different, but consistent behaviors throughout a month of observation. This suggests that rainbow trout have personality attributes as seen in many other vertebrates and observed before in salmonids (e.g., Iguchi et al. 2001). Specifically, there is a genetic contribution to whether a rainbow trout line exhibits a
shy (reactive) or bold (proactive) personality. In a traditional high-density hatchery setting, reactive fish are at a disadvantage and may not grow as fast as a bold fish.

The work presented in Chapter Two did not support the presupposition that hatchery fish are generally more bold, but reaffirmed that boldness is partially under genetic control. Perhaps a better understanding of how different fish in wild populations respond to predators would give us a better framework for interpretation of these results. Looking at different populations of poeciliids from areas under high and low predation, Brown et al. (2007) found that the population with greater predation was on average more bold, suggesting that bold personalities may not always be selected against in the wild. One can predict a frequency dependent effect, with benefits to both bold and shy individuals and relative tradeoffs varying with risk of predation.

The work presented in Chapter Two suggests that there is a genetic component of personality traits in rainbow trout. The identified QTLs may be the same loci that underlie behavioral differences seen between wild and domesticated zebrafish based on a comparative map produced by Rexroad et al. (2008). Further research into locating the source of these QTL would advance our understanding of antipredator behavior and allow hatchery managers or researchers to monitor alleles that are thought to influence these behaviors. Fine-mapping these traits in a population would isolate relatively small regions of a chromosome and candidate genes could then be tested from these regions. Molecular biology could be utilized to tell if the candidate gene is involved in antipredator behavior. Alternatively, several populations could be tested to see if there is a correlation between behavior and alleles at a specific locus.

Another contemporary issue in salmonid genetics relates to the detection of sequence-based genetic variation. In Chapter Three, allelic variation was identified between clonal lines of rainbow trout using transcriptome data. This strategy is often known as genotyping-by-sequencing and is becoming
increasingly popular (Elshire et al. 2011, Davey et al. 2011) because of the low cost of detecting variation using high-throughput sequencing. Duplicated regions of a genome, however, pose a problem when using this strategy if the sequences from one region are mistaken for those of another.

The strategy presented in Chapter Three for dealing with gene duplicates involves a multiple-step process that distinguishes between allelic and paralogous sequence variants. The heart of this method is the use of windows of sequence that can be compared to sequences from another individual. The use of large window comparisons allows even very similar paralogs to be differentiated from alleles. This strategy has been shown in Chapter Three to be superior to existing methods of analyzing transcriptomes in polyploid organisms.

The allelic variation found from Chapter Three may be used to develop hybridization methods of genotyping (e.g. SNP-chip), or could be used as a database for other genotyping-by-sequencing projects in rainbow trout. As mentioned in the introduction, segmental duplications in model organisms may prevent allelic variation from being identified correctly.

During SNP identification the levels of paralogous variation were quantified using the same scripts. In Chapter Four, the copy number of specific genes was estimated from high-throughput transcriptome data. This type of data typically consists of millions of short sequences transcribed from RNA and is often filled with sequencing errors (personal observation). These sequences were aligned to a Unigene dataset comprising non-redundant cDNA sequences. With proper alignments, each gene (single or multiple-copy) is only represented once in each Unigene dataset, allowing us to estimate the gene copy number for a dataset (a proxy for a genome).

Overlapping windows of a Unigene sequence were examined for unique sequence that was represented two or more times. The copy number was estimated when the number of unique sequences was similar for half of the Unigene. Requiring multiple copies of a unique sequence removed sequencing errors,
and with half of the Unigene being represented the effects of repetitive motifs were reduced. These measures make the use of long cDNA sequences and thus the use of fully sequenced genomes unnecessary and therefore broaden the base of species which can be analyzed in this manner.

An advantage of the use of transcriptomes in gene copy estimation is that they are relatively inexpensive to produce when compared to full-length gene or genome sequences. As discussed in the introduction, genome assemblies fare poorly in duplicated regions and any analyses with assemblies will likewise fare poorly when using these regions. Use of full-length genes would be ideal, but in order to represent the entire genome, multiple tissues and time-points would need to be sampled to document the full expression pattern of every gene. High-throughput sequencing of transcriptomes offers a much less expensive and accessible alternative.

The information obtained from this analysis (Chapter Four) mirrored that seen by other researchers and reinforces the historic pattern of whole genome duplications (2R, 3R, and 4R) that have been proposed for vertebrate animals. With the available transcriptome datasets increasing, many more organisms are expected to be available for this type of analysis. Perhaps more organisms, like stickleback, will be found to have a unique aspect of their duplicated genes that will expand our understanding of what happens to the genome after a duplication event.

Further investigations into expression in different tissues at different times may be valuable in understanding how evolution of duplicated genes occurs. Also, modifications to this strategy may be used to categorize genes that may be valuable in gene nomenclature and gene expression studies. Such resources form the basis of what it means to be a model organism.

Advances in genomics are facilitating research in many non-model organisms such as salmonids. These studies have illustrated the application of several novel genomic approaches to contemporary
challenges in these important animals. The approaches described, including the ability to quantify large-scale gene duplication, are also broadly applicable to a wide array of other non-model species.
References


