IDENTIFICATION OF GENES INVOLVED IN HEAT STRESS IN ARCTIC CHARR

by

Nicole Lisa Quinn
M.Sc, McMaster University 2004
B.Sc, McMaster University 2002

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

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Fall 2011

Approval

Name: **Nicole Lisa Quinn**

Degree: **Doctor of Philosophy**

Title of Thesis: **Identification of genes involved in heat stress in Arctic charr**

Examining Committee:

Chair: **Dr. Michel Leroux**
Professor, Department of Molecular Biology and Biochemistry

Dr. William Davidson
Senior Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Bruce Brandhorst
Supervisor
Professor and Chair, Department of Molecular Biology and Biochemistry

---

Dr. John Reynolds
Supervisor
Professor, Department of Biology

Dr. Robert Devlin
Supervisor
Adjunct Professor, Department of Zoology
University of British Columbia

Dr. Fiona Brinkman
Internal Examiner
Professor, Department of Molecular Biology and Biochemistry

Dr. Patricia Schulte
External Examiner
Professor, Department of Zoology
University of British Columbia

Date Defended/Approved: **December 16, 2011**

Partial Copyright Licence

SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

STATEMENT OF ETHICS APPROVAL

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

(a) Human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

(b) Advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

(c) as a co-investigator, collaborator or research assistant in a research project approved in advance,

or

(d) as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Simon Fraser University
Burnaby, BC, Canada

Last update: Spring 2010

Abstract

I set out to identify candidate genes that can be used to develop genetic markers associated with Upper Temperature Tolerance (UTT) for use in a genomics-assisted broodstock program for Arctic charr (*Salvelinus alpinus*) and for examining wild populations at risk due to climate change. This was accomplished using genomic resources available for Atlantic salmon (*Salmo salar*), which allowed me to identify and examine genomic regions and specific genes of interest. In addition, I conducted expression profiling of Arctic charr exposed to acute and chronic thermal stress. Using comparative genomics, I identified several Atlantic salmon fingerprint scaffolds containing markers associated with UTT in Arctic charr and rainbow trout (*Oncorhynchus mykiss*). One of these was fully sequenced using 454 GS FLX next-generation sequencing and annotated, which identified nine genes in the putative Quantitative Trait Locus (QTL) region of the Atlantic salmon genome. This analysis also provided evidence that the 454 sequencing technology was suitable for partial assembly and gene annotation, but not for *de novo* whole genome sequencing of a complex salmonid genome. Next, I conducted expression profiling of phenotypically tolerant and intolerant Arctic charr. The differentially expressed genes were compared with those identified within the UTT QTL sequenced previously, which suggested *COUP-TFII* as a particularly interesting candidate gene. Heat shock proteins (Hsps) and hemoglobins were also significantly associated with acute thermal stress. Concurrently, I performed expression profiling of Arctic charr exposed to moderate, chronic temperature stress that mimicked a more realistic situation. Again, Hsps were identified in the thermal stress response, as well as ribosomal proteins, which were up-regulated throughout the exposure and the recovery period. Finally, I identified and fully annotated all of the hemoglobin genes in Atlantic salmon. This identified substantially more hemoglobin genes in this species than in any other fish analyzed to date, and included several non-Bohr beta hemoglobins, which may be used in emergency response situations. Combined, the findings of my research have substantial implications for developing a temperature tolerant Arctic charr broodstock and for examining wild populations of salmonids for responses to temperature stress brought by climate change.

Keywords: Arctic charr; aquaculture; expression profiling; heat shock protein; hemoglobin; microarray; next generation sequencing; ribosomal protein; salmonid; quantitative trait locus; upper temperature tolerance

Layman's Abstract

Arctic charr is an especially attractive aquaculture species because it features the desirable tissue traits of other salmonids such as Atlantic salmon, sockeye and rainbow trout, but is bred and grown at inland freshwater tank farms year round, a relatively sustainable alternative to the floating net pet method currently used for many salmonids. However, Arctic charr is a cold water species, which limits where it can be grown without the added expense and resources involved in cooling aquaculture tanks. This is an especially relevant concern given the recent 'locally grown food' movement as well as the added threat to both cultured and wild populations brought by climate change. Therefore, it is of interest to develop a genetics-based selective breeding program to generate strains of Arctic charr that are more tolerant to warm temperatures.

For my PhD research, I identified genes that are associated with temperature tolerance and recovery from heat stress in Arctic charr. Specifically, I conducted two heat trials with live Arctic charr. For the first trial, Arctic charr were exposed to acute, lethal temperature stress, which allowed me to identify temperature intolerant and tolerant individuals by selecting the first and last fish to show signs of stress. For the second trial, I subjected Arctic charr to prolonged, moderate temperature stress that mimicked a realistic heat wave scenario. The sampled fish from both trials were genetically screened to identify potential genes involved in tolerance to heat as well as the heat response in general. These genes were compared with those that I had identified within Atlantic salmon in a sequenced region of DNA that was previously suspected to contain important temperature tolerance genes in other salmonids. This identified *COUP-TFII*, a gene for which there is little information in salmonids, as a particularly interesting temperature stress gene, although its specific role remains unclear. Next, I identified three well-known gene families - hemoglobins, heat shock proteins and ribosomal proteins - as potentially playing a role the response to heat stress in Arctic charr. Hemoglobins, which govern the capacity of an organism to deliver oxygen to its tissues, were under-expressed in tolerant fish. The heat shock proteins are a group of proteins that maintain cellular health during stress conditions. The expression levels of these genes were elevated in temperature tolerant fish as well as throughout the exposure to moderate heat stress. Finally, ribosomal genes, which play a role in maintaining protein production, were implicated in both temperature tolerance as well as recovery from prolonged, moderate heat stress. Given their known functions in various aspects of stress response and the overall maintenance of organism health, all of these gene families are logical candidates for temperature tolerance genes, which is supported by the results of my research. However, their specific roles in the trait remains to be clarified by further analyses with Arctic charr of a variety of genetic backgrounds as well as additional experimental conditions.

Finally, given their suspected role in temperature tolerance, I identified and described all of the hemoglobin genes in Atlantic salmon. This revealed that the Atlantic salmon hemoglobin repertoire has several unique characteristics. Specifically, there are more hemoglobin genes in Atlantic salmon than have been identified in any other fish species to date. Additionally, there are several 'non-Bohr' hemoglobins, which are suspected to be involved in stress response, present within the Atlantic salmon genome but not in any other fish. Combined, these features may reflect the unique, dynamic life cycle of Atlantic salmon.

The findings of this work stand to benefit both the aquaculture industry as well as wild populations of Arctic charr by facilitating the development of a temperature tolerant Arctic charr strain and by providing tools for examining wild populations of salmonids for responses to temperature stress brought by climate change.

Acknowledgements

I want to start by expressing my heartfelt appreciation towards Dr. William Davidson, my Senior Supervisor. It is difficult for me to find adequate words to express my gratitude to Willie. Put simply, I could not have asked for a better scientist, teacher, colleague or friend to work with for my PhD. The knowledge and lessons that I have had the privilege of learning from Willie will no doubt transcend into all of my roles - scientist, teacher, mentor, colleague, friend, parent and partner - for the rest of my life. Willie's mantra: "the harder you work, the luckier you get", has become mine.

Second, I want to thank my Supervisory Committee, Drs. Robert Devlin, John Reynolds and Bruce Brandhorst. I feel honoured to have had such absolute experts in their fields advising me throughout my research. The array of knowledge and expertise that spans this group is astounding, and I am grateful for having had the opportunity to study under their mentorship. I would also like to thank Drs. Fiona Brinkman and Patricia Schulte for acting as internal and external examiners, respectively, during my Defence.

Next, I want to thank some of the people that helped me with specific aspects of the project. Dr. Krzysztof Lubieniecki was invaluable almost every step of the way, providing training, troubleshooting and expert knowledge. Keith Boroevich and Will Chow helped with the bioinformatics, and these papers would not have been to the standards that they are without their work. Dr. Colin McGowan is an expert on Arctic charr, and helped with the study design, the implementation of the temperature trials, and the data analysis. Glenn Cooper taught me how to make a beautiful microarray, and what to do when they're not so beautiful; without him, there would be no expression data. Dr. Ben Koop provided expert guidance, scientific advice, and behind-the-scenes support in the form of funding applications. Finally, I thank the members of the Davidson and Koop laboratories who provided technical, knowledge and moral support over the years.

I would not have been able to complete this degree without my friends and family. My girlfriends, most of whom are scientists too, provided hours of essential downtime, usually combined with food, exercise, conversation or all three, that were essential in keeping me balanced. No doubt, without those hundreds of miles run around the city or dozens of trips up and down the beautiful mountains surrounding Vancouver, I would have burned out years ago. My mother provided moral support and guidance, and unquestioning love that gave me boosts of confidence and determination when things were stressful or busy. To my dad, I owe my work ethic, without which I would not have survived the past 5 years. And, last, but absolutely not least, Michael, my husband, my best friend, my mentor, my partner in life. Michael keeps me happy, focused, motivated and comforted. Michael and I are a team, and even though this degree is in my name, and I am the one that understands the big words and the complicated concepts, this accomplishment is just as much his as it is mine.

Table of Abbreviations, Terms and Acronyms

454 GS FLX DNA Sequencing: The next generation sequencing platform owned by Roche (formerly 454 Life Sciences) in 2008. The GS FLX (pronounced "flex") system replaced the GS system (launched in 2005), and has since been improved by the GS FLX Titanium platform, which uses a new series of reagents on the GS FLX instrument.

Annotation: Identification of the genomic position of intron-exon boundaries, regulatory sequences, repeats, gene names and protein products within a DNA sequence.

Bacterial Artificial Chromosome (BAC): A DNA construct in which a segment of genomic DNA (usually 150–250 Kbp) is inserted into a plasmid, which is transformed into a bacteria (usually *E. coli*).

BAC-end Sequence: The first 500–1,000 nucleotides (based on Sanger sequencing) of either end of the genomic DNA insert within a BAC.

BLAST: Basic Local Alignment Search Tool. A BLAST search will compare query sequence against a database (e.g., that of nucleotide sequences, translated nucleotides or proteins).

BSFU Markers: Microsatellite markers located within BAC-end sequences. The Atlantic salmon BSFU markers were identified using an automated sequence screening protocol, which was used to screen the BAC-end sequences from the Atlantic salmon BAC library to identify microsatellite-like elements.

Complementary DNA (cDNA): DNA synthesized from a mature mRNA template by reverse transcription.

COUP TF II (Chicken Ovalbumin Upstream Promoter Transcription Factor II): A transcription factor that is also known as NR2F2 (nuclear receptor subfamily 2, group F, member 2).

Expressed Sequence Tag (EST): A short sequence obtained from one shot sequencing of cDNA, corresponding to a fragment (~500 bp based on Sanger sequencing) of an expressed gene.

Fingerprint Scaffold (FPS): The predicted ordered arrangement of a group of BACs or series of contigs based on overlapping DNA sequences as determined by DNA fingerprinting. For the Atlantic salmon genome, FPSs were formally referred to as

contigs (referring to contiguous regions), which were predicted automatically, and the term fingerprint scaffold was later adopted.

Gene Expression Omibus (GEO): A database repository of high-throughput gene expression data and hybridization arrays, chips and microarrays.

Genomics: Discipline in genetics concerning the study of the genomes of organisms.

Hemoglobin (Hb): The iron-containing protein found within the blood of most vertebrates that binds and carries oxygen from the respiratory organs to the tissues, then carries CO_2 from the tissues to the respiratory organs for release into the environment.

High-throughput: Processes usually performed via increased levels of automation and robotics. In sequencing: involves the application of rapid sequencing technology at the scale of whole genomes.

Marker-assisted selection (MAS): The use of DNA markers linked to traits of interest to assist in the selection of individuals for breeding purposes.

Minimum Tiling Path (MTP): The minimum number of overlapping BACs required to span a given region of the genome that constitute a contiguous DNA sequence (i.e., no gaps).

Molecular marker: Specific fragments of DNA that can be identified within the whole genome. These can be associated with the position of a particular gene or the inheritance of a particular characteristic.

mRNA: Molecule of RNA encoding a specific protein product. mRNA is transcribed from a DNA template in the cell nucleus then moves to the cytoplasm where it is translated by the ribosomes.

Polymerase Chain Reaction (PCR): Technique designed to amplify a DNA fragment across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Quantitative PCR (qPCR) or Quantitative Real Time PCR: A technique the follows the principles of PCR, but that simultaneously amplifies and quantifies a targeted DNA molecule. Usually, qPCR is combined with reverse transcription of mRNA into cDNA to quantify mRNA levels in cells or tissues.

Quantitative trait locus (QTL): Stretch of DNA containing or linked to genes that underlie a quantitative trait of interest.

RAD (Restriction site Associated DNA): RAD tags are the sequences that immediately flank a restriction enzyme site. RAD markers refer to sequence variations identified within RAD tags that can be used as molecular markers (e.g., SNPs).

Reference genome: Nucleic acid sequence database, assembled into a whole genome and representative of a species' genetic code. Typically used as a guide on which new genomes are built.

RNA: Ribonucleic acid. Single stranded molecule transcribed from one strand of DNA. Unlike DNA, the sugar in RNA is ribose and one of the four bases, T (thymine) is replaced by U (uracil). There are several types of RNA mainly involved in protein synthesis (mRNA, tRNA, rRNA).

Sequence assembly: Refers to aligning and merging sequence fragments into a longer sequence.

Single Nucleotide Polymorphism (SNP): A DNA sequence variation occurring within a single nucleotide that differs between members of a species.

Synteny: Physical co-localization of genetic loci on the same chromosome within an individual or species.

Transcription: Process involved in the synthesis of mRNA from a DNA template, catalyzed by RNA polymerase.

Translation: Process in which the messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein.

Upper Temperature Tolerance (UTT): The highest temperature that an organism can tolerate before showing signs of stress, illness or death.

Whole genome sequence: The complete DNA sequence of the genome of an organism.

Table of Contents

Table of Figures

Tables

1: Background and objectives of thesis

1.1 Arctic charr

Arctic charr (*Salvelinus alpinus*) is a member of the family Salmonidae, which includes salmon, trout, charr, freshwater whitefish and grayling. Arctic charr are circumpolar in Arctic, sub-Arctic and alpine regions, and exhibit three different life histories: lake-resident fish, which have accessible migration routes, but choose to remain in a lake throughout their life cycle, land-locked fish, which lack access to migratory routes, and migratory populations, which move from lakes to marine environments and back annually (1).

Arctic charr are a particularly suitable species for aquaculture for several reasons. First, they exhibit desirable tissue traits, including a high omega-3 fatty acid content, flesh colour ranging from white to orangy-pink to deep red, and high palatability similar to other salmonids such as Atlantic salmon and rainbow trout, two of the most popular aquaculture species. Additionally, Arctic charr can thrive in freshwater year round. This allows for inland, closed-containment aquaculture facilities, which circumvents some of the adverse environmental impacts that are associated with the marine, open net-pen farming that is traditionally used for Atlantic salmon as well as some other salmonid species. This feature has been recognized by Arctic charr's position as a 'best choice' seafood on the Monterey Bay Aquarium's Seafood Watch list, as well as the recent recognition of Arctic charr as a 'sustainable seafood member' of the Vancouver Aquarium Ocean Wise program. Finally, Arctic charr naturally school at relatively high

densities, a behavioural characteristic that allows them to thrive in an aquaculture setting. These characteristics, combined with the increasing consumer demand for fish flesh, particularly that of salmonids, make Arctic charr an excellent species for the aquaculture industry.

Icy Waters Ltd., located in Whitehorse, Canada is a fully integrated operation that includes a Fisheries and Oceans certified broodstock facility, hatchery, tank farm and a Canadian Food Inspection Agency approved processing plant. Icy Waters Ltd. is a land based aquaculture facility that uses a gravity fed flow-through design with streams and springs as its water source. This is an energy efficient system that does not rely on any fuel-based technologies to transport, circulate, heat, cool or clean the water, which are the main concerns in terms of resource intensiveness and cost for other land based salmonid aquaculture facilities. The facility produces 120–200 MT of Arctic charr food products per annum, which are sold in restaurants and retail establishments throughout North America. However, the majority of Icy Waters' business is from the sale of ova in the form of eyed eggs to grow-out sites around the world. Approximately 80% of the Arctic charr produced in North America originates from the Whitehorse hatchery.

Given the quality of the product, as well as the sustainability of its production, Icy Waters, and Arctic charr aquaculture in general, are at an advantage in the market. It is unlikely that a substantial increase in the supply of farmed Arctic charr will exceed consumer demand, and therefore market prices should remain at a premium. However, Arctic charr is a cold water species that inhabits water temperatures from -1 to 14ºC. This presents substantial geographical limitations in terms of where Arctic charr can be grown at present. Although this means that, based on its geography and climate, Canada should

be capable of becoming the major producer of Arctic charr in the world, it also limits the range of distribution of eyed eggs and potential customers for Icy Waters Ltd. Tank farms that are otherwise equipped to grow and distribute freshwater fish species, including salmonids such as rainbow trout, often cannot accommodate Arctic charr due to an unsuitable climate and the high energy cost of maintaining tanks within the optimal temperature range for Arctic charr survival. In addition, fish forced to live in temperatures higher than their natural range show signs of stress, including reduced immune function, reduced appetite and growth, increased susceptibility to disease and ultimately death (2).

The challenge of finding suitable climates for Arctic charr aquaculture is an increasing concern as temperatures rise as a result of climate change. Indeed, at Icy Waters, there is no efficient cost effective way to control the temperature of the water flowing into the system. Thus, the tank temperature reflects that of the neighboring streams, springs and lakes. The optimal summer water temperature for growing and spawning healthy Arctic charr is between 11ºC and 13ºC. The observed upper lethal temperature limit for Arctic charr is between 25ºC and 26ºC. The summer of 2004 brought unprecedented high temperatures to the Yukon, which caused the water flowing into the tanks to reach a maximum temperature of 21.5ºC. In total, 24.86 MT of biomass were lost to heat-induced mortality compared to 4.63 MT in the previous year starting with a similar number of fish.

In terms of climate change, it is predicted that the most vulnerable habitats will be the northerly latitudes, putting Arctic charr habitats particularly at risk (3). As temperatures continue to climb and become less predictable, Canadian Arctic charr

hatcheries and tank farms throughout the world will be faced with an on-going struggle to keep fish alive, healthy and comfortable, and to maintain growth and spawning at the optimal rate. This raises the need for an aquaculture strain (broodstock) of Arctic charr that is more robust to temperature fluctuations, and that is more tolerant to elevated temperatures.

1.2 Genomics and aquaculture

1.2.1 Marker assisted selection (MAS)

In traditional selective breeding practices, individuals showing desirable phenotypic characteristics are bred to produce new strains of plants or animals that exhibit features such as faster growth, greater size, increased overall robustness, or improved aesthetic appeal. Farmers, through thousands of years of following the mantra "breed the best to the best and hope for the best" (quotation attributed to American Thoroughbred and Standardbred breeder, John E. Madden), have drastically changed natural populations, as exemplified by the now hundreds of breeds of domestic dog, or differing cattle strains for dairy or beef production. However, these phenotype-based breeding tactics come with inherent drawbacks. Specifically, they drain resources and time, as they often rely on trial and error. They also require vast numbers of individuals exhibiting extensive phenotypic variation, and often numerous generations of repetitive selective breeding are required to see an effect, which is especially difficult for species with long generation times and for which extensive parental investment is needed. Furthermore, often animals need to be sacrificed to detect or measure morphological characteristics such as flesh colour and tissue quality, and these, as well as disease challenged (i.e., to determine immune

function or disease resistance) animals cannot be bred. Finally, most characteristics of interest to aquaculture are complex traits – i.e., those that are governed by numerous genes and complex interactions between multiple pathways, or those for which morphological variation is based on environmental cues, and it is very difficult to select for multiple complex traits together based on phenotypic information alone. Thus, the amount of time and the costs associated with traditional, phenotypic approaches to selective breeding are large.

In contrast to phenotype-based selection, which directly selects for a given heritable trait, in genotype-based selection, or marker-assisted selection (MAS), traits are indirectly selected for based on variability at the DNA level. More specifically, the genomes of breeding populations are screened for multiple markers, or DNA tags, that are associated with genes of interest that work together to produce a desirable phenotype for a complex trait. This genetic screening can be accomplished using high-throughput genomics systems that incorporate tools such as polymerase chain reaction (PCR) or DNA sequence-based screening.

MAS has been used extensively for the genetic improvement of cultivated plant cultivars, such as wheat (4), corn (5) and soy (6), as well as cultured stocks in agriculture species such as swine (7) and cattle (8). As the availability of genomics resources has increased for aquatic species, more research is being done to improve broodstocks for many key aquaculture species worldwide. Specifically, genomics, and the use of genomic tools, enables one to examine the differences and similarities among organisms at the genotypic (DNA) level, as opposed to more traditional broad-scaled phenotype-based (appearance-based) approaches. This very fine-scaled perspective, looking at differences

in alleles as well as variability at non-coding loci, or positions within the genome, means that complex traits that are driven by more than one gene or pathway can be broken down into their components. Therefore, differences at the individual, family, population, species and even the organism level can be assessed in very fine detail. Such insight into the genetic factors that drive complex traits can facilitate the development of effective and efficient breeding methods, which have far-reaching implications for the aquaculture industry.

1.2.2 MAS for temperature tolerance in Arctic charr

The solution to the challenges facing the Arctic charr farming industry, especially in terms of the geographical range of the eyed egg market and the threat of massive losses of fish due to increasing temperature, lies largely in the development of a robust broodstock that tolerates higher than normal temperatures. The integration of a MAS approach into the traditional broodstock development program at Icy Waters Ltd. has the potential to dramatically increase the viability and productivity of Arctic charr farming. That is, the identification of molecular markers that are genetically associated with tolerance to temperature in Arctic charr will facilitate the development of an Arctic charr broodstock that exhibits a higher optimal temperature range and that is more robust to fluctuations in water temperature that are not normally observed in their natural habitat. Ultimately, such markers could improve animal health and welfare, while simultaneously reducing costs and resource use for the Arctic charr aquaculture industry by allowing its expansion into regions that are geographically closer to the target markets for the fish.

1.2.3 Implications for wild populations of Arctic charr

In addition to threatening Arctic charr aquaculture, climate change is also a threat to wild populations of Arctic charr, with potential impacts such as habitat loss, population instability and changes in species interactions, with results such as forced migration, population loss, and reduced genetic diversity (3, 9). Molecular markers for temperature tolerance in Arctic charr, such as those that could be integrated into a broodstock program as described above, could also be used to screen wild populations of the species that might be at risk, or suffering from the effects of climate change. Thereby, it may be possible to develop population-based conservation initiatives, as well as screen for robustness to climate variations, or to track migration, colonization or population declines based on genetic markers for temperature tolerance.

1.3 Thesis objectives and introduction to chapters

The goal of my PhD research was to identify candidate genes that can be used to develop genetic markers associated with elevated thermal tolerance or intolerance for use in a genomics-assisted aquaculture breeding program or for the genetic screening of wild populations at risk due to climate change. To accomplish this goal, I used the genomic resources available for the Atlantic salmon genome to identify and examine genomic regions and specific genes of interest, as well as conducted expression profiling on Arctic charr exposed to acute and chronic thermal stress. Each chapter describes a component of this research, as follows:

Chapter 2 describes the generation of several Atlantic salmon fingerprint scaffolds (FPS) putatively associated with upper temperature tolerance (UTT), as well as their

partial annotation by comparative synteny. Chapter 3 describes the full sequencing of one of these FPSs and the identification of seven putative UTT-associated genes in Atlantic salmon. For Chapter 4, I conducted expression profiling of phenotypically tolerant and intolerant Arctic charr, which identified several genes associated with tolerance and intolerance to acute, lethal temperatures, and compared these genes with those identified with the UTT quantitative trait locus (QTL) sequenced in Chapter 3. Chapter 5 describes the results of expression profiling of Arctic charr exposed to moderate, chronic thermal stress that mimics a realistic situation such as a heat wave at an aquaculture facility or in the wild. Hemoglobins were identified as playing a role in UTT in Chapter 4; however, at the time, there was little known about the hemoglobin genes in salmonids. Given their putative involvement in UTT in Arctic charr, as well as the insight that this well-studied group of proteins could provide to understanding salmonid evolution, for Chapter 6 I identified and fully annotated all of the hemoglobin genes in Atlantic salmon. Finally, Chapter 7 discusses the conclusions, implications and future directions of this work. Chapters 3 through 6 constitute published, peer-reviewed manuscripts, and are presented herein unaltered from their published forms with the exception that the tables and figures have been re-named according to the chapter numbers for clarity (e.g., Figure 3-2 refers to the second figure in Chapter 3).

1.4 References

1. Loewen TN, Gillis D, Tallman RF. Ecological niche specialization inferred from morphological variation and otolith strontium of arctic charr *salvelinus alpinus L.* found within open lake systems of southern baffin island, nunavut, canada. J Fish Biol. 2009;75(6):1473-95.

2. Pankhurst NW, King HR. Temperature and salmonid reproduction: Implications for aquaculture. J Fish Biol. 2010 Jan;76(1):69-85.

3. Prowse TD, Furgal C, Bonsal BR, Edwards TW. Climatic conditions in northern canada: Past and future. Ambio. 2009 Jul;38(5):257-65.

4. Gupta PK, Mir RR, Mohan A, Kumar J. Wheat genomics: Present status and future prospects. Int J Plant Genomics. 2008;2008:896451.

5. Tuberosa R, Salvi S, Sanguineti MC, Landi P, Maccaferri M, Conti S. Mapping QTLs regulating morpho-physiological traits and yield: Case studies, shortcomings and perspectives in drought-stressed maize. Ann Bot. 2002 Jun;89 Spec No:941-63.

6. Kim DH, Kim KH, Van K, Kim MY, Lee SH. Fine mapping of a resistance gene to bacterial leaf pustule in soybean. Theor Appl Genet. 2010 May;120(7):1443-50.

7. Liu G, Jennen DG, Tholen E, Juengst H, Kleinwachter T, Holker M, Tesfaye, D, Un, G, Schreinemachers, HJ, Murani, E, Ponsuksili S, Kim JJ, Schellander K, Wimmers K. A genome scan reveals QTL for growth, fatness, leanness and meat quality in a duroc-pietrain resource population. Anim Genet. 2007 Jun;38(3):241-52.

8. Veerkamp RF, Beerda B. Genetics and genomics to improve fertility in high producing dairy cows. Theriogenology. 2007 Sep 1;68 Suppl 1:S266-73.

9. Urban MC, Holt RD, Gilman SE, Tewksbury J. Heating up relations between cold fish: Competition modifies responses to climate change. J Anim Ecol. 2011;80(3):505-7.

2: *Salmo salar* as a reference genome for genomics and the development and partial annotation of minimum tiling paths of *S. salar* BACs associated with a UTT QTL

2.1 Introduction

The holy grail of any genomics program for a species is a whole genome sequence that is well assembled and annotated. The advantages of this are many, but are mainly centred around two things: 1) the wealth of data that is produced, including the full gene repertoire with additional information such as gene location and copy number, and 2) the ability of the sequenced genome to act as a reference genome, both for the sequenced species itself (i.e., such that the genomes of additional individuals can be easily re-sequenced using the original as a reference for assembly), as well as to provide information for other, closely related species. Currently, however, even with the great advances in sequencing technology that have come to light in recent years, obtaining a whole genome sequence remains an extremely difficult, costly and time-consuming undertaking. This is particularly true for fish species simply due to the evolutionary age of fish and the more than 20,000 extant species (1), factors that make fish genomes diverse and complex and complicate sequencing. Indeed, only six fish genome sequences have been reported to date (medaka, *Oryzias latipes*; tiger pufferfish, *Takifugu rubripes;* green spotted pufferfish, *Tetraodon nigriviridis*; zebrafish, *Danio rerio;* stickleback, *Gasterosteus aculeatus*, and most recently, Atlantic cod, *Gadus morhua*), although more are underway. However, five of these fish species were chosen for their abilities to act as

models for studying genetics, rather than for their utility for aquaculture. Specifically, the medaka (2) and zebrafish (NCBI BioProject PRJNA9557) (3) genomes were sequenced to provide model organisms for studying developmental biology, while the stickleback genome serves as a model for studying adaptive evolution (NCBI Bioproject PRJNA11772) and the two pufferfish represent the smallest known vertebrate genomes (4, 5). Figure 2-1 illustrates the phylogenetic relationships among these species as well as some key aquaculture species, and shows that the full spectrum of teleosts is not represented by the genome sequences that are currently available.

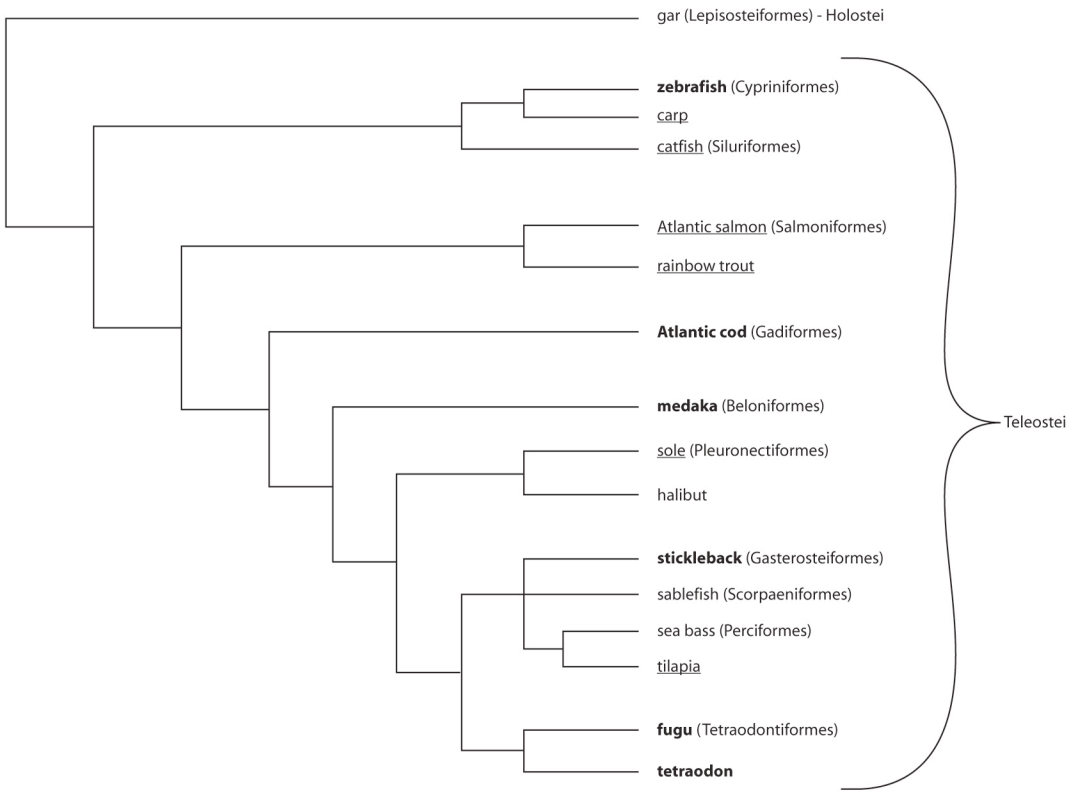

Figure 2-1 Schematic representation of the phylogenetic relationships among Teleost fish species.

Species in bold have publically available full genome sequences, while those that are underlined are currently being sequenced. Note that the full spectrum of teleosts is not represented by the genome sequences that are currently available. Species are listed by their common names with orders in parentheses. Gar is used as an outgroup.

When a whole genome sequence is not available, there are numerous genomics resources and tools that can be developed which, particularly when used in combination, can provide extensive insight into a genome and can be used for applications such as MAS, both for the species of interest, and for other closely-related species. In this research, to pursue the goal of identifying QTL for UTT in salmonids, particularly Arctic charr, I used Atlantic salmon (*Salmo salar*) and its genomics resources as a reference genome. Atlantic salmon is a particularly good model fish species for genomics because there are extensive genomics resources available, which were developed using standard methods and approaches that are applicable to other genomes. Furthermore, a strong argument has been made for obtaining the full genome sequence for Atlantic salmon, a project that is currently in progress (6). Thus, given that there are no salmonid species yet sequenced, Atlantic salmon can serve as a reference salmonid genome, providing extensive opportunities for cross-referencing, or comparative synteny analyses with other salmonids, particularly the Pacific salmon (*Oncorhynchus sp*.), rainbow trout (*Oncorhynchus mykiss*) and Arctic charr (*Salvelinus alpinus*).

Atlantic salmon also provides an example of some of the challenges that face fish genomics in general. Specifically, the common ancestor of salmonids underwent a whole genome duplication event between 20 and 120 million years ago (7, 8). Thus, whereas there are usually two copies of each gene within a genome, Atlantic salmon have four, and the duplicate copies are evolving into genes with new functions or non-coding DNA. The genome duplication also increased the size and the repetitiveness of the genome. These characteristics, combined with the lack of a closely related guide sequence, mean

that sequencing and assembling the Atlantic salmon genome are extremely challenging, and that genomics studies of salmonids are generally complicated.

2.1.0 Identification of UTT QTL in salmonids

Generally, characteristics that are desirable for selection, such as UTT, are complex, or 'quantitative', which means that they are not simply governed by a single gene, but rather are controlled by a suite of genes and regulatory factors that interact to produce a phenotypic characteristic. The resulting phenotype usually falls somewhere on a spectrum. This is not always the case; traits such as genetic disorders can be determined by a single gene (e.g., cystic fibrosis and Duchenne muscular dystrophy), but most phenotypic traits associated with health, growth, meat quality and the capacity for utilizing new and ecologically sustainable food items are complex traits that are strongly influenced by the environment, and thus show a range of phenotypes. QTL are regions of the genome that contain, or are linked to, genes that contribute to a particular trait (9). Identifying QTL for a trait is usually the first step when the goal is to develop molecular markers for a complex trait.

As the first step to identifying molecular markers associated with UTT in salmonids, I sought to explore and characterize the QTL associated with UTT that had previously been identified in Arctic charr and rainbow trout. Specifically, research carried out by R. Danzmann's group at the University of Guelph identified a polygenic basis for UTT in selected lines of rainbow trout (*Oncorhynchus mykiss*) with the detection of UTT QTL on at least nine linkage groups (10-12). These studies used a survival-based method to identify high and low temperature tolerant fish by subjecting them to upper lethal temperature tolerance experiments in temperature trial aquaria.

Subsequent work by the same group using a similar protocol identified two significant and seven suggestive marker UTT associations in Arctic charr (13). It was observed that six of the Arctic charr UTT QTL were shared with those previously identified in rainbow trout. Conserved synteny was found between Arctic charr linkage group 26 (AC-26) and rainbow trout linkage group 6 (RT-6) through comparative mapping of the microsatellite marker Cocl3LAV, which mapped to the same region as the UTT marker SsaF43NUIG on AC-26 and Ssa20.19NUIG on RT-6. This suggested that there are homologous regions of the genome that harbour genes associated with thermal tolerance in all salmonids.

2.1.1 Co-localization of SsaF43NUIG and SSa20.19 to LG 23 of Atlantic salmon

Given the suspected UTT QTL in the region of SsaF43NUIG and Ssa20.19NUIG in Arctic charr and rainbow trout, as well as the relative lack of genomics resources for these species, our lab group sought to identify the locations of these microsatellite markers in Atlantic salmon. Specifically, the markers were tested for variability within the two Atlantic salmon SALMAP mapping families, Br5 and Br6, each of which contains two parents and 46 offspring (14).

The microsatellite markers SsaF43NUIG and Ssa20.19NUIG were informative (i.e., variable) in both of the Atlantic salmon SALMAP mapping families, Br5 and Br6 (14) and mapped within 1.1 centimorgan of each other on linkage group 23 of Atlantic salmon, which corresponds to Atlantic salmon chromosome 16 (15) (see Figure 4-2). This provided strong evidence for a UTT QTL in this region of the Atlantic salmon genome. Thus, minimum tiling paths (MTPs) of the Atlantic salmon bacterial artificial

chromosomes (BACs) surrounding each of these markers were identified, and that spanning SsaF43NUIG was sequenced and annotated, as described in detail in Chapter 3.

2.1.2 BAC-end sequences and BSFU markers in the UTT QTL

Usually, once a BAC library is generated and the corresponding physical map is assembled, the next step is to conduct BAC-end sequencing. Because the sequence of the cloning vector within the BAC is known, the vector sequences at the junctions between the genomic DNA and the vector can act as sequence primers. Thereby, the first 500–1,000 nucleotides (based on Sanger sequencing) of the genomic DNA inserts—i.e., the BAC-ends—can be determined. End-sequencing a large subset of BACs can provide extensive insight into the full genome sequence. For example, the 207,869 BAC-end sequences for the Atlantic salmon BAC library cover approximately 3.5% of the whole genome sequence. This glance into a genome is very powerful, as it can provide information about the complexity of the genome (i.e., repeat content). Additionally, the BAC-end sequences can be a source of molecular markers. For the Atlantic salmon genome, approximately 20,000 microsatellite markers have been identified from BAC-end sequences, which are identified by the tag "BSFU" after the marker number. Finally, the BAC-end sequences can be used for comparative synteny analyses by aligning them against other fully sequenced genomes, which can provide insight into the gene content of the BACs, thus providing a partial, putative annotation of segments of the genome (16). This technique is particularly powerful if the two genomes being compared are closely related, as one can act as a reference genome (e.g., common carp and zebrafish, which are both Cypriniformes) (17). Even when there is no obvious reference genome, there is value in this exercise as it often suggests homologous regions among

phylogenetically distantly related species, and can lead to candidate genes being suggested for specific traits. The Atlantic salmon BAC end sequences have been made available in a publicly accessible searchable database (www.asalbase.org) along with comparative genomic information for four of the six published fish genomes.

2.1.3 Objectives

The goal of the following experiments was to use the available information, including the BAC-end sequences for the Atlantic salmon physical map, as well as the comparative synteny tools available within the Atlantic salmon database, ASalbase, to attempt to conduct gene annotation *in silico* for the UTT QTL associated with the markers SsaF43NUIG and Ssa20.19NUIG. Specifically, by aligning the BAC-end sequences within the associated contigs against the published medaka genome, I attempted to generate a list of putative UTT-related genes, which could subsequently be compared against the gene lists generated by microarray analysis in the subsequent chapters.

2.2 Methods

2.2.0 BAC identification

First, 40-mer probes corresponding to the microsatellite markers SsaF43NUIG and Ssa20.19NUIG markers were designed (SsaF34NUIG: GTGCTGTGTTTCAGGGCCACTGAGCAGCTTGTCC; Ssa20.19NUIG: CACACACTTGGTAGAGGAGAGGCTGTGCTGGGGAAACTAG) (R. Powell, personal communication). To identify the Atlantic salmon BACs containing the SsaF43NUIG and Ssa20.19NUIG microsatellite markers, the probes were labelled with γ^{32}P-ATP using T4 polynucleotide kinase (Invitrogen, Burlington, Ont. Canada) and

hybridized to filters containing the Atlantic salmon BAC library (18) (CHORI-214; CHORI, BAC-PAC Resources, Oakland, CA, USA.). Filters were exposed to phosphor screens that were scanned and visualized using ImageQuant™ software, giving an image of the ^{32}P-labeled hybridization-positive BACs containing the probe sequence. All hybridization-positive BACs were verified using PCR with the corresponding primers for the probe of interest. The locations of these BACs within the Atlantic salmon physical map (19), i.e., their corresponding fingerprint scaffolds (FPSs) were determined using the "BAC to FPS" tool within ASalbase.

I also identified four microsatellite markers within BAC-end sequences (i.e., BSFU markers) that mapped within zero centimorgans of the SsaF43NUIG and Ssa20.19NUIG markers on the Atlantic salmon linkage maps: Ssa0150BSFU, Ssa0245BSFU, Ssa0250BSFU and Ssa0266BSFU (see Figure 4-2, although note that several additional BSFU markers, which were subsequently identified, are indicated in this figure). Given that these microsatellites originated from BAC-end sequences, their source BACs, and the corresponding FPSs were known.

2.2.1 Chromosome walking to generate FPSs

The Atlantic salmon physical map was generated automatically using a highly stringent cut-off likelihood score of $1e^{-16}$, and end-to-end joints with a minimum score of $1e^{-10}$ were accepted to merge contigs (19). This inherently stringent method can sometimes exclude BACs from a contig (now referred to as an FPS), or separate contigs (i.e., omit legitimate contig joints) as a result of below-threshold band-sharing. In addition, the repetitive nature of the Atlantic salmon genome complicates the assembly of a physical map as numerous BACs from different regions of the genome can exhibit common

banding patterns. Thus, it was of interest to try to manually expand the contigs or, ideally, join them using 'chromosome walking'. For this, 40-mer oligonucleotide probes were designed from the BAC-end sequences of the outer-most BACs in the contigs of interest (i.e., those containing the SsaF43NUIG and Ssa20.19NUIG markers and the BSFU microsatellite markers surrounding them), as determined by the BAC order predicted by the physical map. These markers were then used to probe the Atlantic salmon BAC library to identify any adjacent BACs that were either not included in the physical map (i.e., singletons), or to join two adjacent contigs together into one FPS. All of the BACs identified as positives by hybridization were verified by PCR using the corresponding primers for the BAC-end sequence. Note that when two or more FPSs were joined, the FPS assumed the numerical name of the lowest contributing FPS. This 'chromosome walking' was done for each of the six FPSs of interest until a 'dead-end' was reached on both sides of each. A dead-end was declared when probing the BAC library with the probe designed from the outermost BAC either resulted in no adjacent BACs identified, or many (i.e., too many to accurately score the filters) BACs were hybridization-positive, which indicated a repetitive region within the BAC.

2.2.2 Generation of minimum tiling paths (MTPs)

Based on the predicted order and overlap of the BACs within the physical map, a series of putatively 'tiled' or overlapping BACs was selected that would represent the minimum number of BACs necessary to span each FPS without any gaps in DNA sequence (i.e., a minimum tiling path; MTP). These putative MTPs were verified by PCR by designing primer sets for sequence tag sites (STSs) in both the SP6 and T7 ends of selected BACs. Using these primers, BACs that were predicted to overlap with the STS source BAC as

well as other surrounding BACs were screened using PCR, thereby establishing relative BAC orientation and confirming any overlap. In situations that these predicted MTPs failed, i.e., the BACs were shown not to overlap by PCR, new BACs were selected and the MTP was re-designed accordingly.

2.2.3 Comparative synteny to generate candidate gene lists

Within the Atlantic salmon database, ASalbase (www.asalbase.org), developed by the Davidson lab, it is possible to conduct a BLAST (20) search of the Atlantic salmon BAC-end sequences within an FPS against four of the six published fish genomes (medaka, *Oryzias latipes;* green spotted pufferfish, *Tetraodon nigriviridis*; zebrafish, *Danio rerio* and stickleback, *Gasterosteus aculeatus*). The purpose of this feature is to provide an indication of the orthologous chromosomes(s) and region(s) within the chromosome(s) of the other fish species to the FPS of interest in Atlantic salmon, as well as a rough annotation of the genes within the FPS based on the genes annotated within the syntenic region(s) of the other fish.

I tested the utility of this experimental annotation tool for generating lists of candidate UTT genes from the six UTT QTL-associated FPSs identified above. The medaka genome was chosen for comparison against Atlantic salmon because, at the time, it was one of the most complete and fully annotated fish genomes available, and it was therefore expected to provide the most reliable results. I used FPS 483, which contains nine BACs that were fully sequenced and annotated (see Chapter 3), as a reference. Specifically, using the ASalbase BLAST tool, the FPS 483 BAC-end sequences were aligned by performing a nucleotide BLAST against the medaka genome to test the parameters necessary to identify as many of the nine genes that were annotated from the

fully sequenced FPS as possible (see Chapter 3). A single BAC-end sequence within FPS 483 (i.e., Soo10J14_T7) aligned with medaka chromosome 6 between bases 13116629—13117640. However, this initial alignment only spanned one BAC-end sequence and thus did not encompass all of the nine known genes within medaka chromosome. In addition, because there was only one BAC-end sequence from Atlantic salmon FPS 483 that aligned with the medaka genome, I was unable to determine the relative orientation of the two alignments based on the position of the BAC-end sequence within the FPS. Therefore, I manually expanded the region of the medaka genome by 518,000 bp, or approximately 25% of the estimated length of Atlantic salmon FPS of 2 Mbp (note that ASalbase predictes FPS 483 to be 2 Mbp long, although sequencing the MTP of the region in Chapter 3 revealed that it is actually 1 Mbp long), on either side of the initially aligned region. This enabled the incorporation of eight of the nine known genes (all except for "*novel protein similar to vertebrate perillipin*") that had been annotated within the sequenced orthologous Atlantic salmon FPS 483. Given that the region was expanded in both directions (because I was unable to determine the relative position and orientation of the single BAC-end sequence hit within the medaka genome), the alignment incorporated 24 additional gene hits upstream of the first gene (*gonadotropin releasing hormone receptor*) that was annotated within the sequenced FPS 483 MTP in Chapter 3. Therefore, for the rest of the FPSs (i.e., those not sequenced), in an attempt to ensure that all genes within the FPS were identified, the orthologous medaka regions were expanded by 25% of the estimated size of the Atlantic salmon sequence in both directions, despite the known risk of including genes that were actually outside of the corresponding FPS. Any annotated medaka genes within >60% identity and an e-value $<e^{-10}$ to the Atlantic

salmon BAC-end sequences within the corresponding FPSs were included in the gene lists. Using this method, I attempted to generate lists of putative UTT-associated genes, which could be subsequently compared to those that would be identified by microarray analysis in future experiments. Note that, at the time this research was conducted (2009), the ASalbase BLAST alignment tool utilized the Ensembl release 53, which accessed the medaka1.0 genome release (October 2005). Additionally, we chose to use the medaka genome for this initial test because, at the time, it provided the best annotation, and thus was expected to be the most informative for this purpose.

2.3 Results

2.3.0 MTPs of FPSs

The goal of 'chromosome walking' was to expand each microsatellite-containing (i.e., SsaF43NUIG, Ssa20.19NUIG or BSFU markers) FPS as much as possible, ideally until another microsatellite-containing FPS was reached, thus joining the two FPSs, or alternatively, until a dead-end was reached. Unfortunately, none of the microsatellite-containing FPS could be joined before dead-ends were reached; however, several of them were significantly expanded, and many contigs were joined to generate larger FPSs (note again that when two FPSs are joined, they assume the name of the lower contributing FPS). This information for each FPS is summarized in Supplemental File 2-1, and the MTPs for FPS 358 and FPS 617 are illustrated schematically in Figures 2-2 and 2-3, respectively, while that for FPS 483 is provided in Chapter 3 (Figure 3-4).

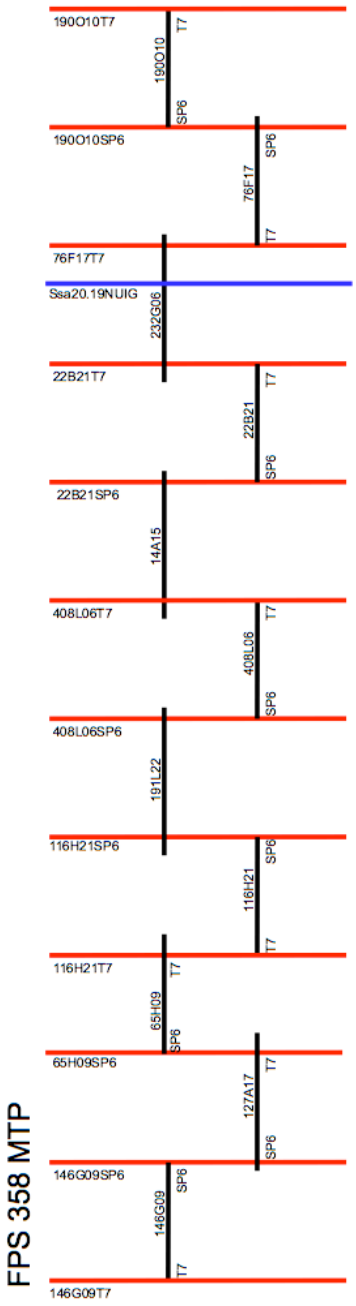

Figure 2-2 Minimum tiling path of FPS 358

Minimum tiling path BACs (black bars) of FPS 358, indicating sequence tag sites (STSs; red bars) and BAC orientation when known. The approximate location of marker Ssa20.19NUIG is indicated by the horizontal blue bar.

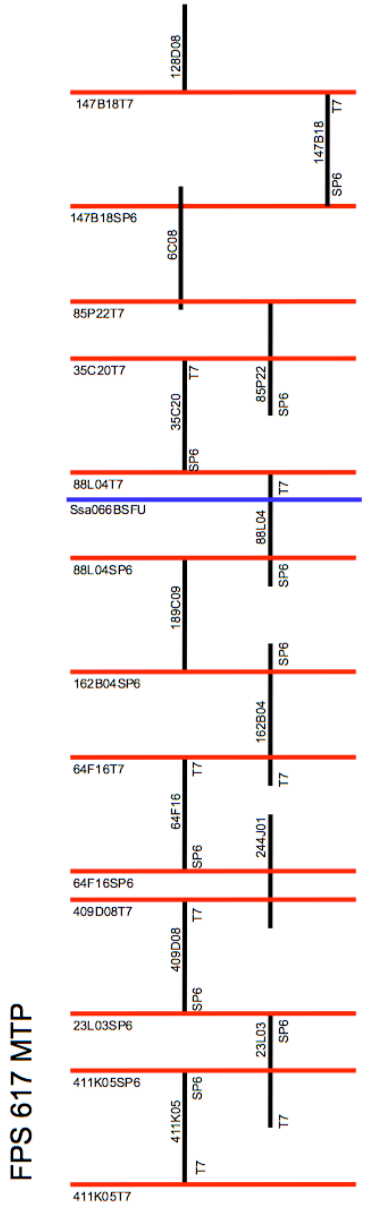

Figure 2-3 Minimum tiling path of FPS 617

Minimum tiling path BACs (black bars) of FPS 617, indicating sequence tag sites (STSs; red bars) and BAC orientation when known. The approximate location of marker Ssa0266BSFU is indicated by the horizontal blue bar.

2.3.1 Candidate genes

Supplemental File 2-1 also lists the orthologous medaka chromosome(s) to the Atlantic salmon FPS, the predicted base range (numbers based on the syntenic medaka chromosome) and range size, the expanded (25% in each direction) range boundaries and size, and the number of annotated genes within the expanded syntenic medaka region. The gene lists are provided in Supplemental File 2-2.

2.4 Discussion

It was reassuring that eight of the nine genes within the reference FPS were found within the expanded syntenic region of medaka chromosome 6, and that their order was well conserved. However, as evidenced in the summarized results in Supplementary File 2-1, there were several challenges with the results of the comparative synteny analysis between the Atlantic salmon BAC-end sequences within the FPSs of interest and the sequenced medaka genome. Firstly, despite the (rather liberal) minimum alignment criteria of 60% identity with an e-value e-value $<e^{-10}$, often very few (or no) Atlantic salmon BAC-end sequences aligned with the medaka genome, meaning that it was impossible to assess the size of the conserved QTL regions, and often impossible to get an idea of the relative positioning and orientation of the BAC-end sequence(s) within the syntenic regions. This may suggest that the BAC-end sequences provided insufficient coverage of the FPSs to align with the medaka genome, or that they represented unique sequences that are not found within that genome. Secondly, an Atlantic salmon FPS often aligned to more than one medaka chromosome, or more than one distinct region on a

chromosome, indicating that there has been extensive shuffling and rearrangement of chromosome segments between these two genomes, or that the BAC-end sequence contained repetitive sequences that are found multiple times within the genome. Thirdly, the differences in genome size (~3 Gbp for Atlantic salmon, vs. 700 Mbp for medaka), as well as the Atlantic salmon genome duplication and the extensive phylogenetic distance between these two species complicated the alignment, making it difficult set parameters in terms how far to expand the alignment in order to encompass all relevant genes. Finally, as demonstrated by the 24 additional genes that were included in the region of the medaka chromosome 6 that was syntenic to FPS 483 and that was expanded by ~500 Kbp in each direction to encompass all of the known genes within this FPS (see Chapter 3), this method incorporated a large number of genes that were outside of the corresponding FPS region, and therefore that could be outside of the QTL region.

The results of this exercise indicate that, at present, the method appears to be of limited use. This is particularly true given that, in practice, there likely would not be a fully sequenced reference FPS available on which to base the alignment parameters and assess the results. Nevertheless, given that the method is largely automated and involves minimal time and resources, it may still be worthwhile to conduct the exercise for the purpose of comparing the gene lists with any putative UTT genes identified by other more reliable means, such that further analyses could be performed should the gene lists overlap. It may also be useful for providing an indication of whether a region of interest contains a particular gene, and thus determining whether further analysis of that region is warranted. Additionally, it should be noted that, after this research was conducted, a similar approach was successfully used to identify candidate genes for resistance to

infectious salmon anemia (ISA) in Atlantic salmon (21). Therefore, the method does indeed appear to have some merit, although clearly its success is dependent on the nature of the genomic region of interest (noting that the region used here appeared to be relatively gene-poor, with only nine genes within the 1 Mbp reference MTP). Furthermore, given a more closely related reference genome, particularly that of another salmonid, this method may be more consistently valid, and could even eliminate the need for sequencing a genome or a region of interest. Indeed, it is possible that, once the full sequence of the Atlantic salmon genome is available, it could act as a reference for aligning BAC-end sequences of other species, and thereby enable partial annotation of other salmonids or teleosts for which full genome sequencing is not feasible.

2.5 Supplementary Material

Appendix File 1 *Supplemental File 2-1*

This table lists the markers, corresponding FPSs, former FPSs, the predicted size based on the Atlantic salmon physical map (from ASalbase), the number of BACs and BAC-end sequences within each FPS, the orthologous medaka chromosome, the original range start and end points within the medaka chromosome resulting from BLAST alignment of the Atlantic salmon BAC-end sequences, the expanded range and the number of genes within the final gene list.

Appendix File 2 *Supplemental File 2-2*

Gene lists generated by comparative homology analysis between Atlantic salmon UTT QTL MTPs and the medaka genome.

2.6 References

1. Nelson JS. Fishes of the world. New Jersey: John Wiley and Sons Inc.; 2006.

2. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007 06/07;447(7145):714-9.

3. Fishman MC. Zebrafish--the canonical vertebrate. Science. 2001;294(5545):1290-1.

4. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffeis A, Rash S, Hoon S, Smit A, Gelpke M, Roach J, Oh T, Ho I, Won M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith S, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowan L, Hood L, Tan YH, Elgar G, Hawkins T, Vankatesh B, Rokhsar M, Brenner S. Whole-genome shotgun assembly and analysis of the genome of *fugu rubripes*. Science. 2002;297(5585):1301-10.

5. Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, Mauceli E, Bounear L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Sugurens B, Dasilva C, Sakabiybat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie M, Biemont C, Duprat S, Brottier P, Coutancear J-P, Couzy, J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff J-N, Guigo R, Zody MC, Mesirov J, Linbald-Toh K, Birren B, Nusbaum Ch, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H. Genome duplication in the teleost fish *tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature. 2004 10/21;431(7011):946-57.

6. Davidson WS, Koop BF, Jones SJ, Iturra P, Vidal R, Maass A, Jonassen I, Lien S, Omholt SW. Sequencing the genome of the atlantic salmon (*salmo salar*). Genome Biol. 2010;11(9):403.

7. Allendorf FW, Thorgaard GH. Tetraploidy and the evolution of salmonid fishes. In: Evolutionary Genetics of Fishes. New York, NY: Plenum Press; 1984. p. 55.

8. Ohno S. Evolution by gene duplication. New York, NY: Springer-Verlag; 1970.
9. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer Associates Inc.; 1998.

10. Danzmann R, Jackson T, Ferguson M. Epistasis in allelic expression at upper temperature tolerance QTL in rainbow trout. Aquaculture. 1999;173:45-58.

11. Jackson, TR, Ferguson MM, Danzmann RG, Fishback AG, Ihssen PE, O'Connel M, Crease TJ . Identification of two QTL influencing upper temperature tolerance in three rainbow trout (*oncorhynchus mykiss*) half-sib families. Heredity. 1998;80:143,144-151.

12. Perry GM, Danzmann RG, Ferguson MM, Gibson JP. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Oncorhynchus mykiss*). Heredity. 2001 Mar;86(Pt 3):333-41.

13. Somorjai IM, Danzmann RG, Ferguson MM. Distribution of temperature tolerance quantitative trait loci in arctic charr (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*). Genetics. 2003 Nov;165(3):1443-56.

14. Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Koop BF, Hoyheim B, et al. Distribution of ancestral proto-actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (rainbow trout and atlantic salmon). BMC Genomics. 2008 Nov 25;9:557.

15. Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF, Danzmann RG, Davidson WS. Assignment of atlantic salmon (*salmo salar*) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*oncorhynchus mykiss*). BMC Genet. 2009 Aug 18;10:46.

16. Sarropoulou E, Franch R, Louro B, Power DM, Bargelloni L, Magoulas A, Senger F, Tsalavouta M, Patarmello T, Galibert F, Kotoulas G, Geisier R. A gene-based radiation hybrid map of the gilthead sea bream sparus aurata refines and exploits conserved synteny with *tetraodon nigroviridis*. BMC Genomics. 2007 Feb 7;8:44.

17. Xu P, Li J, Li Y, Cui R, Wang J, Wang J, Zhang Y, Zhao Z, Sun X. Genomic insight into the common carp (*cyprinus carpio*) genome by sequencing analysis of BAC-end sequences. BMC Genomics. 2011 Apr 14;12:188.

18. Thorsen J, Zhu B, Frengen E, Osoegawa K, de Jong PJ, Koop BF, Davidson WS, Hoyhelm B. A highly redundant BAC library of atlantic salmon (*salmo salar*): An important tool for salmon projects. BMC Genomics. 2005 Apr 4;6(1):50.

19. Ng SH, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJ, de Jong PJ, Koop BF, Krzywinski MI, Lubieniecki K, Marra MA, Mitchell LA, Mathewson C, Osoegawa K, Parisotto SE, Phillips RB, Rise ML, von Schalburg KR, Schein JE, Shin H, Siddiqui A, Thorsen J, Wye N, Yang G, Zhu B. A physical map of the genome of Atlantic salmon, *salmo salar*. Genomics. 2005 Oct;86(4):396-404.

20. Altschul S, Gish W, Miller W, Myers E, J L. Basic local alignment search tool. J Mol Biol. 1990;5:403-10.

21. Li J, Boroevich K, Koop B, Davidson W. Comparative genomics identifies candidate genes for infectious salmon anemia (ISA) resistance in atlantic salmon (*Salmo salar*). Marine Biotechnology. 2011;13(2):232-41.

3: Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome

Published in: *BMC Genomics* (2008) Vol. 11, No.1, pp. 539, ISSN 1471-2164.

Author list: Nicole L Quinn[1], Natasha Levenkova[2], William Chow[1], Pascal Bouffard[2], Keith A Boroevich[1], James R Knight[2], Thomas P Jarvie[2], Krzysztof P Lubieniecki[1], Brian A Desany[2], Ben F Koop[3], Timothy T Harkins[4], and William S Davidson[1]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada
[2]454 Life Sciences, Branford, USA
[3]Department of Biology, University of Victoria, Victoria, Canada
[4]Roche Applied Science, Indianapolis, USA

Author contributions: NLQ, PB, TPJ, BD, JK, TTH, BFK and WSD conceived the project. NLQ established the minimum tiling path and prepared the DNA. PB was responsible for GS FLX pyrosequencing. NL, WC, KAB, JK, KPL and BD performed bioinformatics. NLQ, NL, WC, PB, JK, KAB, KPL and WSD analyzed and interpreted the data. NLQ, TTH and WSD prepared the manuscript.

3.1 Abstract

3.1.1 Background

With a whole genome duplication event and wealth of biological data, salmonids are excellent model organisms for studying evolutionary processes, fates of duplicated genes and genetic and physiological processes associated with complex behavioral phenotypes. It is surprising therefore, that no salmonid genome has been sequenced. Atlantic salmon (*Salmo salar*) is a good representative salmonid for sequencing given its importance in aquaculture and the genomic resources available. However, the size and complexity of the genome combined with the lack of a sequenced reference genome from a closely related fish makes assembly challenging. Given the cost and time limitations of Sanger sequencing as well as recent improvements to next generation sequencing technologies, we examined the feasibility of using the Genome Sequencer (GS) FLX pyrosequencing system to obtain the sequence of a salmonid genome. Eight pooled BACs belonging to a minimum tiling path covering ~1 Mb of the Atlantic salmon genome were sequenced by GS FLX shotgun and Long Paired End sequencing and compared with a ninth BAC sequenced by Sanger sequencing of a shotgun library.

3.1.2 Results

An initial assembly using only GS FLX shotgun sequences (average read length 248.5 bp) with ~30× coverage allowed gene identification, but was incomplete even when 126 Sanger-generated BAC-end sequences (~0.09× coverage) were incorporated. The addition of paired end sequencing reads (additional ~26× coverage) produced a final

assembly comprising 175 contigs assembled into four scaffolds with 171 gaps. Sanger sequencing of the ninth BAC (~10.5× coverage) produced nine contigs and two scaffolds. The number of scaffolds produced by the GS FLX assembly was comparable to Sanger-generated sequencing; however, the number of gaps was much higher in the GS FLX assembly.

3.1.3 Conclusion

These results represent the first use of GS FLX paired end reads for *de novo* sequence assembly. Our data demonstrated that this improved the GS FLX assemblies; however, with respect to *de novo* sequencing of complex genomes, the GS FLX technology is limited to gene mining and establishing a set of ordered sequence contigs. Currently, for a salmonid reference sequence, it appears that a substantial portion of sequencing should be done using Sanger technology.

3.2 Background

The salmonids (salmon, trout and charr) are of considerable environmental, economic and social importance. They contribute to ecosystem health by providing food sources for predators such as bears, eagles, sea lions and whales. As an increasingly popular food choice for humans, salmonid species contribute to local and global economies through fisheries, aquaculture and sport fishing. In addition, they have distinct social importance as they are a traditional food source for indigenous peoples, and play a significant role in their culture and spirituality. Salmonids are also of great scientific interest. The common ancestor of salmonids underwent a whole genome duplication event between 20 and 120 million years ago [1,2]. Thus, the extant salmonid species are

considered pseudo-tetraploids whose genomes are in the process of reverting to a stable diploid state. More is known about the biology of salmonids than any other fish group, and in the past 20 years, more than 20,000 reports have been published on their ecology, physiology and genetics. Salmonids, with their genome duplication and wealth of biological data, are excellent model organisms for studying evolutionary processes, fates of duplicated genes and the genetic and physiological processes associated with complex behavioral phenotypes [3]. It is surprising therefore, that no salmonid genome has been sequenced to date.

The Atlantic salmon (*Salmo salar*) is an ideal representative salmonid for genome sequencing given the popularity of this species for aquaculture as well as the extensive genomic resources that are available. The current genomic resources include: a BAC library, restriction enzyme fingerprint physical map comprising 223,781 BACs in ~4,300 contigs [5], 207,869 BAC-end sequences that cover ~3.5% of the genome sequence, a linkage map with ~1,600 markers, ~600 of which are integrated with the physical map [6], and > 432,000 ESTs [7,8]. The haploid C-value for Atlantic salmon is estimated to be 3.27 pg [9], or a genome size of approximately 3×10^9 bp, which is very comparable to the sizes of mammalian genomes. The Atlantic salmon genome is highly repetitive, and at least 14 different DNA transposon families whose members are ~1.5 kb have been described [10]. Although five fish genomes have been sequenced (medaka, *Oryzias latipes*; tiger pufferfish, *Takifugu rubripes;* green spotted pufferfish, *Tetraodon nigriviridis*; zebrafish, *Danio rerio* and stickleback, *Gasterosteus aculeatus*), they represent euteleostei lineages, and often very derived species that have been separated from salmonids for at least 200 million years [11]. The complexity of the Atlantic salmon

genome combined with the lack of a closely related guide sequence means that sequencing and assembly will be extremely challenging.

Conventional Sanger sequencing of paired end templates (2–4 kb plasmids, 40 kb fosmids, or ~150 kb BACs) using fluorescent di-deoxy chain terminators and capillary electrophoresis revolutionized the field of genomics (reviewed in [12]). Although this approach remains the gold standard for sequence and assembly quality, limitations with respect to cost, labor-intensiveness and speed, which are largely due to the necessity of generating and arraying cloned shotgun libraries and isolating template DNA for sequencing, have fueled the demand for new approaches to DNA sequencing. In recent years, several novel high-throughput sequencing platforms have entered the market including the SOLiD system by Applied Biosystems [13], the Solexa technology [14], now owned by Illumina, the recently released true Single Molecule Sequencing (tSMS) platform by Helicos [15] and the 454 platform [16], now owned by Roche. Most of these are targeted to the goal of re-sequencing an entire human genome for < $1,000 [17]. This next generation of genome sequencing stands to have major scientific, economic and cultural implications with respect to applications such as personalized medicine, metagenomics and large-scale polymorphism studies on organisms of commercial value whose genomes have already been sequenced. However, the ability of these technologies to sequence the genomes of complex organisms *de novo* remains unknown.

A common feature among the new generation of sequencing procedures is the elimination of the need to clone DNA fragments and the subsequent amplification and purification of DNA templates prior to capillary sequencing. Rather, sequence templates are handled in

bulk, and massively parallel sequencing by synthesis or ligation allows the generation of hundreds of thousands to millions of sequences simultaneously.

With respect to *de novo* whole genome sequencing, perhaps the most promising new technology uses a pyrosequencing protocol [18] optimized for solid support and picolitre scale volumes (i.e., pyrosequencing using the 454 system [16]). The 454 pyrosequencing technology [both the Genome Sequencer (GS) 20 and FLX generation systems] has proven very successful for a number of applications such as complete microbial genome sequencing [19] metagenomic and microbial diversity analyses [20,21] ChIP sequencing and epigenetic studies [22,23], genome surveys [24], gene expression profiling [25] and even for sample sequencing fragments of Neanderthal DNA that were extracted from ancient remains [26,27]. Recent accomplishments include its contribution to a high quality draft sequence of the grape genome [28] as well as complete re-sequencing of an individual human genome, for which the assembly was accomplished by mapping 454 reads back to a reference genome [29].

Although several studies comparing 454 pyrosequencing with Sanger sequencing have shown that the per base error rates of the two technologies are similar [27,30], 454 pyrosequencing has limitations. The major concerns have been relatively short read lengths (i.e., as of 2007 an average of 100–200 nt compared to 800–1,000 nt for Sanger sequencing), a lack of a paired end protocol and the accuracy of individual reads for repetitive DNA, particularly in the case of monopolymer repeats [12]. Combined, these factors often make it impossible to span repetitive regions, which therefore collapse into single consensus contigs during sequence assemblies and leave unresolved sequence gaps. These issues have recently been addressed with the release of the GS FLX system

as well as the Long Paired End sequencing platform. The GS FLX system provides longer read lengths and lower per-base error rates than the previous systems. In addition, the 454 technology offers the longest read length of any of the next generation sequencing systems currently available. Thus, we chose to evaluate the ability of the 454 technology, as it stands, to sequence a complex genome without the aid of high-coverage Sanger-generated reads.

With respect to *de novo* assembly of a complex genome, the most relevant test to date of the capability of the 454 pyrosequencing technology (GS 20 system) involved sequencing four BACs containing inserts of the barley genome, two of which had previously been sequenced using the traditional Sanger approach [30]. The barley genome is relatively large (5.5×10^9 bp) and is comprised of more than 80% repetitive DNA, posing a significant challenge for sequencing. Whereas each BAC contained approximately 100 Kb of genomic DNA, the cumulative size of all consensus sequence contigs per BAC did not reach the actual size of the BAC clones for any of the 454-based assemblies. This was largely due to the pooling of repetitive sequences into single contigs. Thus, while the 454 technology proved useful for identifying genes, it was of limited value for producing long contiguous sequence assemblies [30].

Given the significant and ongoing improvements in the 454 technology since the barley BAC analysis, which include longer read lengths and higher sequence accuracy attributable to the release of the GS FLX system, as well as the availability of a paired end protocol, we set out to assess the feasibility of using this technology to sequence the Atlantic salmon genome. Here we report the results of using the GS FLX pyrosequencing system to sequence *de novo* a 1 Mb region of Atlantic salmon DNA covered by a

minimum tiling path comprising eight BACs. We discuss the integration of Atlantic salmon genomic resources such as BAC-end sequences as well as assembly techniques and annotation tools given the lack of a closely related guide sequence. We also address the ability of the GS FLX Long Paired End technology to establish the order of sequence contigs and assemble them into large scaffolds. Finally, we compare the GS FLX assemblies with and without the addition of paired end reads to a Sanger-generated assembly of a ninth BAC from the same region of the genome. This is the first application of the GS FLX Long Paired End system for *de novo* assembly of a large region from a complex genome. This study represents the most difficult challenge for 454 pyrosequencing thus far, and the results we present can be used to assess the feasibility of this technology for sequencing the Atlantic salmon genome *de novo*.

3.3 Methods

3.3.0 Establishment of minimum tiling path and DNA preparation

We initially chose contig 570 of the Atlantic salmon physical map for analysis due to the presence of the microsatellite marker SsaF43NUIG, which is linked to upper temperature tolerance in rainbow trout [31,32] and Arctic charr [33]. Contigs 2469 and 483 were joined to contig 570 using 'chromosome walking'. Specifically, 40-mer oligonucleotide probes were designed from the BAC-end sequences of the outer-most BACs in the contigs, as determined by the contig order predicted by the physical map, beginning with contig 570. The probes were labeled with γ^{32}P-ATP using T4 polynucleotide kinase (Invitrogen, Burlington, Ont. Canada) and hybridized to filters containing the Atlantic salmon BAC library [4] (CHORI-214; CHORI, BAC-PAC Resources, Oakland, CA, USA.). Filters were exposed to phosphor screens that were

scanned and visualized using ImageQuant™ software, giving an image of the ^{32}P-labeled hybridization-positive BACs containing the probe sequence. All hybridization-positive BACs were verified using PCR with the SsaF43NUIG primers [34]. The minimum tiling path across Atlantic salmon contig 483 was established by designing primer sets for sequence tag sites (STSs) in both the SP6 and T7 ends of selected BACs. Using these primers, we screened the BACs that were predicted to overlap with the STS source BAC given the predicted assembly from the Atlantic salmon physical map using PCR, thereby establishing relative BAC orientation and overlap. The minimum tiling path was then established by selecting the minimum number of overlapping BACs required to span the entire contig. We isolated approximately 5 μg of cloned Atlantic salmon BAC DNA from the minimum tiling path BACs using Qiagen's Large Construct kit as per the manufacturer's directions (Qiagen, Mississauga, Ont. Canada). The kit includes an exonuclease digestion step to eliminate *E. coli* genomic DNA.

3.3.1 454 Shotgun pyrosequencing

The shotgun sequencing protocol using the 454 sequencing system has been described previously [16]. The salmon BAC results presented here were generated on the GS FLX (454 Life Sciences, Branford, CT) whereas the results presented previously [16] were generated on the GS 20 sequencer, the previous generation instrument. The GS FLX instrument is capable of generating 100 million bp of sequence in approximately 250 bp reads in a 7.5 hour run. Additionally, the GS FLX system has a significantly lower error profile than the GS 20 system.

Briefly, to generate the GS FLX shotgun library, the isolated Atlantic salmon BAC DNA was mechanically sheared into fragments, to which process specific A and B adaptors were blunt end ligated. The adaptors contain the amplification and sequencing primers necessary to the GS FLX sequencing process. After adaptor ligation, the fragments were denatured and clonally amplified via emulsion PCR, thereby generating millions of copies of template per bead. The DNA beads were then distributed into picolitre-sized wells on a fibre-optic slide (PicoTiterPlate™), along with a mixture of smaller beads coated with the enzymes required for the pyrosequencing reaction, including the firefly enzyme luciferase. The four DNA nucleotides were then flushed sequentially over the plate. Light signals released upon base incorporation were captured by a CCD camera, and the sequence of bases incorporated per well was stored as a read. DNA extractions were performed at Simon Fraser University (Burnaby, BC, Canada), and library generation and sequencing were performed at 454 Life Sciences (Branford, CT, USA).

3.3.2 GS FLX Long Paired End DNA library generation and sequencing

GS FLX Long Paired End library generation for 454 sequencing has been described previously [23]. Briefly, DNA was sheared into ~3 kb fragments, EcoRI restriction sites were protected via methylation, and biotinlylated hairpin adaptors (containing an EcoRI site) were ligated to the fragment ends. The fragments were subjected to EcoRI digestion and circularized by ligation of the compatible ends, and subsequently randomly sheared. Biotinlyated linker containing fragments were isolated by streptavidin-affinity purification. These fragments were then subjected to the standard 454 sequencing on the GS FLX system. The paired end reads are recognizable as the known linker (originating

from the two hairpin adaptors) surrounded by BAC sequence. When sequenced on the GS FLX, this protocol generates two, ~100 bp tags known to be ~3 kb apart. These paired end reads were used to build the original contigs and to assemble the contigs into scaffolds.

3.3.3 GS FLX assemblies

A previous version of the Newbler assembler used in performing the assemblies has been described previously [16], and the overall structure and phases of the assembler used here follows the structure described in that paper; however, the algorithms used for the specific phases of assembly have been upgraded. The upgraded Newbler assembler identifies pairwise overlaps between reads, and then uses them to construct multiple alignments of contiguous regions of the dataset. Boundaries where the read-by-read alignments diverge or converge (such as at the boundaries of repeat regions) define breaks in the contig multiple alignments (also called branch points). The resulting data structure consists of a graph, where each node is a contiguous multiple alignment, undirected edges exist between the 5' and 3' ends of the contig nodes, and reads form alignments along paths of the graph. The assembler builds this multiple alignment graph using an adjustable greedy algorithm of taking a 'query' read, finding the pairwise overlaps to it, constructing a multiple alignment of those overlaps, then choosing a subsequent 'query' read from the overlapped reads that are only partially aligned so far (thereby extending the multiple alignment). If any pairwise overlap alignments conflict with the current multiple alignment graph, corrective algorithms use the conflicting alignments to either ignore the new pairwise overlap (if the graph is more consistent) or to correct the constructed multiple alignment (if the new pairwise overlap identifies a

misalignment in the graph). These overlaps and multiple alignment algorithms use a combination of nucleotide-space (i.e., the bases of the reads) and flow-space (i.e., the 454 flowgram signal intensities of the reads), where available, to perform the multiple alignment construction.

Following the construction of the multiple alignment graph, a series of 'detangling' algorithms are used to simplify the complex regions of the graph, such as overly collapsed regions shorter than the length of the reads (i.e., parts of reads that happened to be near-identical to each other by chance, and so produced overlaps that collapsed into a single multiple alignment region). The nodes in the resulting graph after detangling are considered to be the 'contigs' by the assembler, and those longer than 500 bp are output as the 'large contigs' of the assembly (those longer than 100 bp are output in the set of 'all contigs').

If paired end reads are included in the data set (either 454 or Sanger paired ends), then an additional scaffolding step is performed after detangling, to create chains of contig nodes using the paired end information. The pairs from each library where both halves of the pair occur in the same contig are used to calculate expected pair distances for the library. The scaffolding algorithm then performs a greedy algorithm of identifying pairs of nodes where at least two paired end reads have their halves aligned at the ends of the pair of nodes, with the correct alignment direction and expected distance from each other. In addition, the set of paired end reads aligned at those two contig ends must support the unambiguous chaining of the two nodes as immediate neighbors in a scaffold, with fewer than 10% of the paired end reads aligning to other contig nodes in the

assembly. The chains of contig nodes found by this greedy algorithm are output as the scaffolds of the assembly.

3.3.4 Gene mining of 454 GS FLX assemblies using syntenic regions

Sequence contigs > 1,000 bp were analyzed using a variety of sequence similarity searches and gene prediction algorithms that have been incorporated into an in-house computational pipeline and database [35]. Sequences entering this pipeline were screened (masked) for repetitive elements using RepeatMasker 3.1.8 [36] and were searched against the NCBI nr (non-redundant) and Atlantic salmon EST [8] databases using BLAST [37]. A GENSCAN gene model prediction algorithm [38] was used to predict introns and exons, and the resulting predictions were searched against the Uniref50 (clustered sets of sequences from UniProt Knowledgebase) database [39]. Finally, a rps-BLAST against the NCBI CDD (Conserved Domain Database; [40]) was conducted to provide additional information with respect to the predicted genes [see Additional File 3-1].

3.3.5 Use of BAC-end sequences to confirm GS FLX scaffold builds and order

The final scaffold assembly incorporating all data (GS FLX shotgun, paired end and BAC-end reads) was verified by conducting BLAST searches of the 126 BAC-end sequences against the four scaffolds > 10,000 bp and comparing the alignment positions with those predicted by the Atlantic salmon physical map. This method was also used to establish relative scaffold order and to confirm the gene order predicted by the BLAST searches of the 454 shotgun and BAC-end sequence contigs against four published fish genomes.

3.3.6 Sanger shotgun sequencing, assembly and annotation

The ninth BAC (S0022P24) of the minimum tiling path was sequenced using standard Sanger sequencing of a shotgun library. Briefly, the purified BAC DNA was sheared by sonication and blunt-end repaired. The sonicated DNA was size fractioned by agarose gel electrophoresis and 2–5 kb fragments were purified using the QIAquick Gel Extraction Kit (Qiagen, Mississauga, Ont. Canada). DNA fragments were ligated into pUC19 plasmid that had been digested with SmaI and treated with shrimp-alkaline phosphatase to produce de-phosphorylated blunt ends. The ligation mixture was used to transform supercompetent *E. coli* cells (XL1-Blue; Stratagene, La Jolla, CA. USA). Transformed cells were cultured overnight at 37°C on LB/agar plates supplemented with ampicillin (200 mg/L) and 1,920 (5 × 384 well plates) clones were sent to the Michael Smith Genome Sciences Centre for sequencing. The sequences were analyzed for quality using PHRED [41], assembled using PHRAP [42], and viewed using Consed version 15.0 [43]. The S0022P24 assembly was annotated using the same protocol as the GS FLX assemblies (see above).

3.4 Results and Discussion

3.4.0 Selection of BACs for GS FLX pyrosequencing

Using chromosome walking, we joined contigs 2469 and 483 to contig 570, and by convention, the new contig was named after the lowest numbered contig within it (i.e., contig 483). Contig 483 contains 195 BACs and includes 126 BAC-end sequences with an average read length of 660 bp. A contig summary can be found in the Atlantic salmon database [6]. Nine BACs were required to span the contig in a minimum tiling path (Fig. 3-1); eight tiled BACs were selected for GS FLX pyrosequencing and the final (ninth)

BAC was sequenced using standard Sanger sequencing of a shotgun library. The estimated length of the minimum tiling path, based on HindIII banding patterns and accounting for overlap between BACs was 1,119,000 bp, with the eight BACs sequenced by GS FLX pyrosequencing accounting for ~950,000 bp. This is probably an underestimate of the true length as doublet and triplet bands may be counted only once.

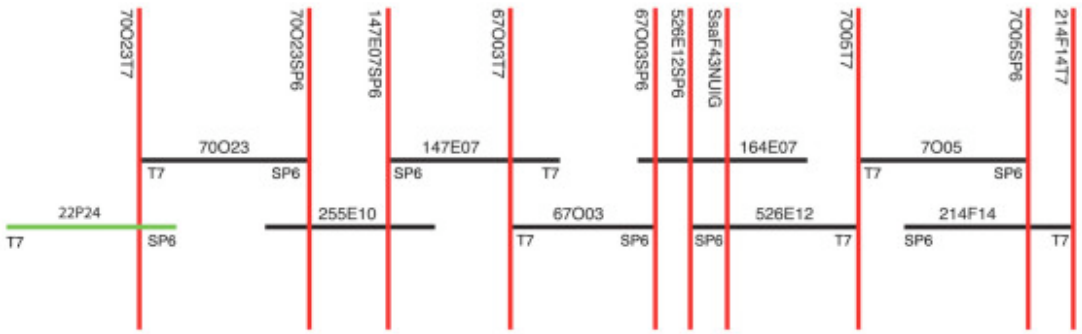

Figure 3-1 Nine BACs within the minimum tiling path (MTP) of Atlantic salmon contig

483

Using the BAC-end sequences, primers were developed to amplify sequence tag sites (STSs – vertical lines), which were used to design and verify a minimum tiling path across the contig. BAC S0022P24 (green line) was sequenced using traditional Sanger sequencing of a shotgun library and the remaining eight BACs (black lines) were sequenced using the GS FLX platform.

3.4.1 GS FLX shotgun assemblies with and without BAC-end sequences

We created a GS FLX shotgun library using eight pooled BACs belonging to a minimum tiling path that spanned approximately 1 Mb of the Atlantic salmon genome. The shotgun run produced 141,746 high quality reads with an average read length of 248.5 bp (Fig. 3-2a). After filtering for vector and *E. coli* sequences, 101,705 reads with a total of 30,549,147 bases were assembled into 803 contigs, 149 of which were > 500 bp and therefore defined as large contigs. Note that this definition of a large contig would include all Sanger-generated reads, which typically range from 500–800 bp. The average contig size was 6,381 bp and the largest contig comprised 34,471 bp. The N50 contig size, defined as the largest contig size at which half of the total size of the contigs is represented by contigs larger than the N50 value, was 11,497 bp (Table 3-1). The second assembly incorporated an additional 89,095 bp in the form of 126 Sanger-generated BAC-end sequences with an average read length of ~660 bp. This effectively added 126 large contigs to the 149 generated by GS FLX shotgun sequencing. Assembling the GS FLX shotgun data with the BAC-end sequences enabled contig joins, thereby decreasing the number of large contigs to 138 and increasing the N50 contig size to 13,455 bp. The average contig size for the second assembly was 6,827 bp and the largest contig size was 38,211 bp. Both assemblies produced an estimated total length of ~1,080,000 bp not including sequence gaps, which is in agreement with the estimate derived from HindIII fragments (Fig.3-3). The GS FLX shotgun sequencing produced ~30× coverage of the region and the BAC-end sequences provided an additional ~0.09× coverage.

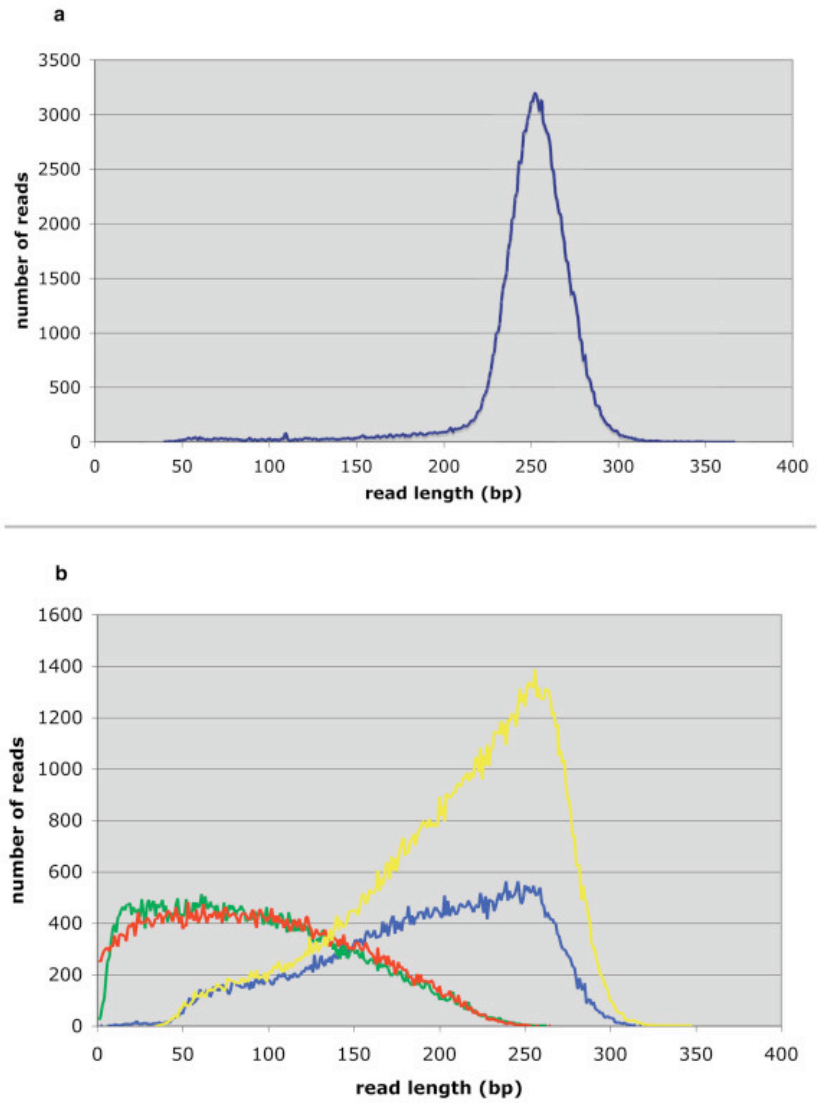

Figure 3-2 Read lengths for GS FLX shotgun and Long Paired End sequencing

a. Distribution of the read lengths for the GS FLX shotgun sequencing (average 248.5 bp)

b. Distribution of read lengths of the GS FLX Long Paired End sequencing. The yellow curve represents the raw reads (average read length 210 bp). These were separated into those containing the linker sequence and those without. The reads containing the linker sequence were separated into two paired end reads, one to the left of the linker (green curve; average read length 93 bp) and those to the right of the linker (red curve; average read length 96 bp). Reads without the linker sequence (blue curve, average read length 191 bp) were added to the assembly as additional shotgun reads.

Summary of GS FLX shotgun assemblies

	SG	SG+BE
Reads assembled	101705	102953
Singleton reads	2795	2870
Large contigs[a] (> 500 bp)	149	138
Total number of contigs	803	811
Bases in large contigs	950826	942244
Total bases covering region	1088103	1081281
Average contig size (bp)	6381	6827
N50 contig size[b] (bp)	11497	13455
Largest contig (bp)	34471	38211
> Q40 bases (bp)	947699	939244

Table 3-1 Summary of GS FLX shotgun assemblies

GS FLX shotgun assembly alone (SG) and when combined with 126 BAC-end sequences (SG+BE). [a]Contigs are defined as more than one read joined by overlapping sequence. Large contigs defined as greater than 500 bp. [b]The N50 contig size is defined as the largest contig size at which half of the total size of the contigs is represented by contigs larger than the N50 value.

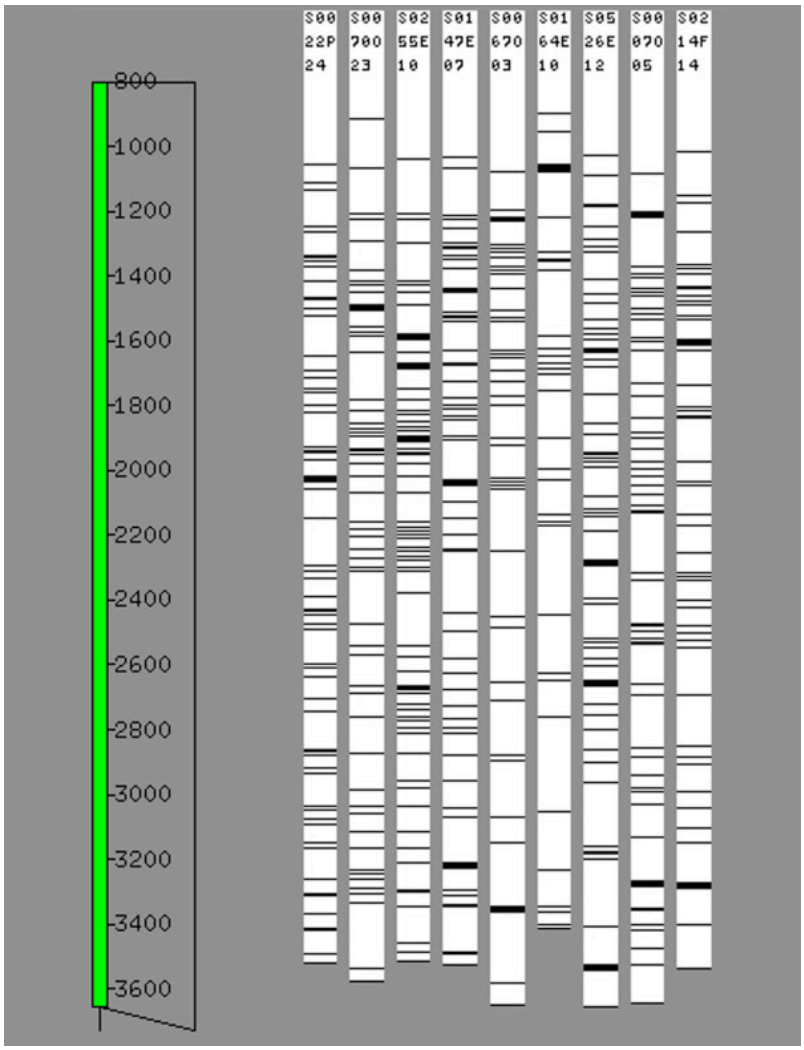

Figure 3-3 HindIII banding patterns of the nine BACs that comprise the minimum tiling path of contig 483 of the Atlantic salmon physical map.

Adjacent lanes share some common bands indicating overlap, whereas lanes separated by more than one lane do not share common bands except when HindIII fragments are of the same size by chance. Scale indicates migration distance. The nine tiled BACs were estimated to span 1,119,000 bp with the eight BACs sequenced by the GS FLX system accounting for approximately 950,000 bp as determined by summing the unique bands in each lane.

3.4.2 Annotation of GS FLX shotgun contigs > 1,000 bp

BLAST results for four fish genomes (medaka, *Oryzias latipes*; tiger pufferfish, *Takifugu rubripes;* zebrafish, *Danio rerio* and stickleback, *Gasterosteus aculeatus*) against the large contigs from the GS FLX shotgun and BAC-end sequence assembly revealed hits to seven well annotated genes and one hypothetical gene (Fig. 3-4a). BLAST results against the *Tetraodon nigriviridis* genome were inconclusive, as most sequence contigs matched to "un_random" sequences (sequence contigs and scaffolds that have not been mapped to any *Tetraodon* chromosome) that collectively spanned over 130 Mb. No genes were identified in any of the fish genomes that were not found in the Atlantic salmon sequence contigs and *vice versa*, indicating conservation of synteny for this genomic region for these four species. Gene order was conserved across three of the four fish species (medaka, zebrafish and the tiger pufferfish), whereas there were two apparent inversions in the stickleback genome relative to the other genomes (Fig. 3-4b), which may be an artifact of the preliminary, incomplete assembly of the stickleback genome. Using these results and assuming conservation of gene order among teleosts, we could predict the order of 12 gene-containing sequence contigs relative to one another; however, their order with respect to the remaining 126 large contigs could not be established. This confirmed the utility of GS FLX shotgun sequencing for gene discovery and highlighted the difficulty of using this approach alone to assemble the sequence of a complex genome *de novo*.

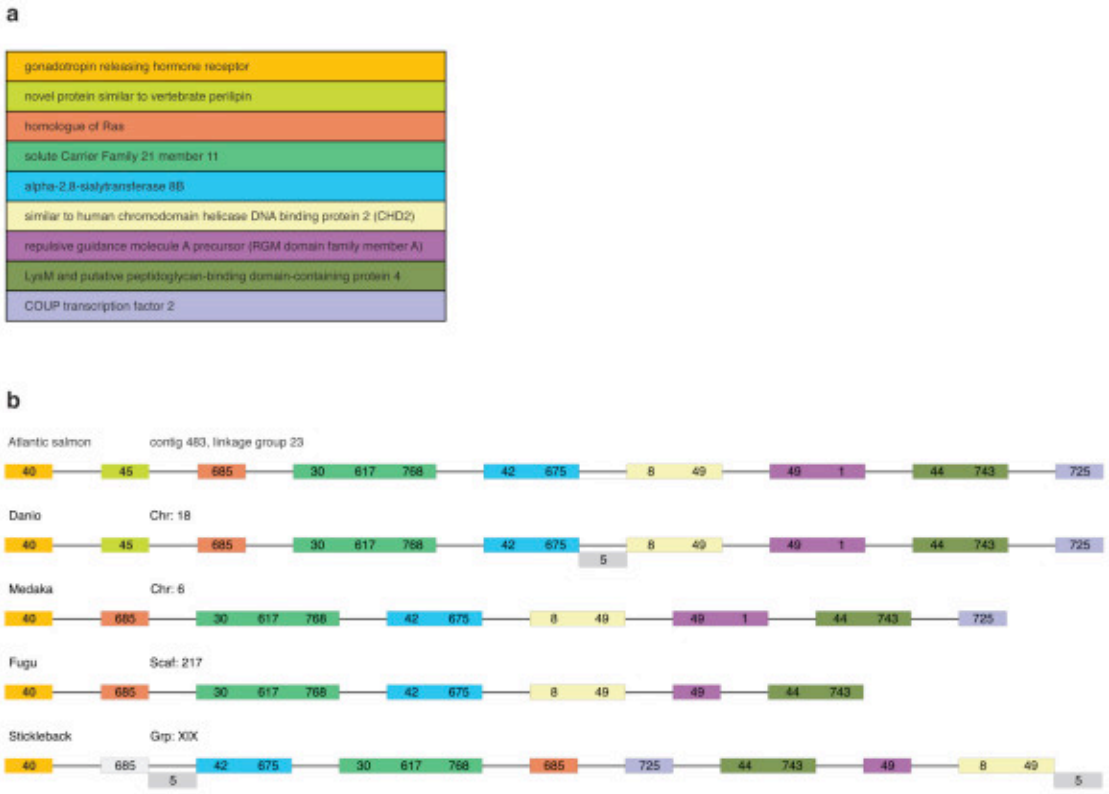

Figure 3-4 Gene annotation by comparative synteny

a. Genes identified in the nine BACs using our in-house annotation pipeline http://grasp.mbb.sfu.ca/. b. Order of the genes within the minimum tiling path. Comparative synteny analysis against the four published fish genomes (medaka, *Oryzias latipes*; tiger pufferfish, *Takifugu rubripes;* green spotted pufferfish, *Tetraodon nigriviridis*; zebrafish, *Danio rerio* and stickleback, *Gasterosteus aculeatus*) enabled the ordering of the gene-containing contigs in the GS FLX assembly of shotgun reads only. This order was confirmed when contigs were assembled into scaffolds with the addition of GS FLX Long Paired End reads. Numbers correspond to contig identity in the Atlantic salmon assemblies; colors coordinate with genes listed in Figure 4a. The grey boxes that correspond to sequence contigs 5 and 685 indicate matches to hypothetical genes. The genes for gonadotropin releasing hormone receptor and the novel protein similar to vertebrate perilipin were found within the Sanger-sequenced BAC and the remaining genes were within the eight BACs sequenced by GS FLX pyrosequencing.

3.4.3 Assemblies incorporating GS FLX Long Paired End data

We constructed a GS FLX Paired End library using DNA from the eight tiled BACs to test its ability to improve the shotgun assembly. After trimming for *E. coli* and vector sequences, the GS FLX Long Paired End sequencing produced 149,035 high-quality reads with an average read length of 210 bp (Fig. 3-2b). Of these, 66,739 contained the linker sequence used to construct the paired end library; therefore, they represented the two paired ends of DNA separated by linker. The average read lengths of the paired ends were 93 and 96 bp for left and right sides of the linker, respectively (Fig. 3-2b). The remaining reads (i.e., those not containing linker) had an average read length of 191 bp (Fig. 3-2b) and were used in the assembly as additional shotgun reads. After splitting each linker-containing read into two paired ends and adding the remaining reads, 213,118 usable reads were obtained. When assembled, these produced 310 contigs, 203 of which were assembled into six large scaffolds (i.e., > 10,000 bp) with an N50 scaffold size of 197,327 bp and the largest scaffold was 227,111 bp (Table 3-2). When combined with the GS FLX shotgun reads, the assembly yielded 289 large contigs, 106 of which were assembled into three large scaffolds with an N50 scaffold size of 361,606 bp and the largest scaffold size was 501,016 bp. Finally, when the 126 BAC-end sequences were incorporated, 286 contigs were produced, 175 of which were assembled into four large scaffolds [GenBank: EU481821] with an N50 and largest scaffold value of 538,994 bp. The GS FLX Long Paired End sequencing provided an additional ~26× coverage of the eight tiled BACs, which, when combined with the GS FLX shotgun data resulted in ~56× coverage of the region. So far, the only published use of the GS FLX Long Paired End

technology has been for revealing structural variations in the human genome [23]. The results presented here represent the first use of this technology for *de novo* genome sequence assembly.

The combination of GS FLX shotgun and Long Paired End reads provided approximately 56× coverage of the 1 Mb region of the salmon genome. We speculate that this represents extensive over-coverage and that similar results could be obtained using fewer reads and less coverage of the region. However, further studies that examine various combinations of coverage from shotgun and paired end libraries are necessary to test this hypothesis and to determine the optimal combination of the two GS FLX read types for genome assembly.

Summary of GS FLX Long Paired End assemblies

	PE only	PE+SG	PE+SG+BE	S0022P24
Large contigs[a] (> 500 bp)	310	289	286	14
Average contig size (bp)	2686	3058	3149	8885
N50 contig size[b] (bp)	4160	4728	5635	32866
Contigs assembed into scaffolds[c]	203	186	175	9[h]
Total scaffolds	9	3	4	2
Large scaffolds[d] (> 10 Kb)	6	3	4	2
Average large scaffold size (bp)	96257	299378	226679	112155
Largest scaffold size (bp)	227111	501016	538994	137857
N50 scaffold size[e] (bp)	197327	361606	538994	137857
Total gaps[f]	194	183	171	8
Maximum gap size (bp)	1,881	2,100	2,131	unknown
Minimum gap size (bp)	4	4	8	unknown
Pair distance average[g] (bp)	2680	2776	2782	N/A
Pair distance deviation (bp)	670	694	696	N/A
Total bases covering region	958507	1002840	1000926	231017
Depth of coverage	~26×	~56×	~56×	~10.5×

Table 3-2 Summary of GS FLX Long Paired End assemblies

Results for GS FLX Long Range Paired End (PE) assembly alone and when combined with the GS FLX shotgun (SG) data and BAC-end (BE) sequences. [a]Contigs are defined as more than one read joined by overlapping sequence. Large contigs are greater than 500 bp. [b]The N50 contig size is defined as the largest contig size at which half of the total size of the contigs is represented by contigs larger than the N50 value. [c]A scaffold is defined as two or more contigs associated by paired ends. [d]Large scaffolds are those consisting of more than 10,000 bp among all contigs therein. [e]The N50 scaffold size is defined as the largest scaffold size at which half of the total size of the scaffolds is represented by scaffolds larger than the N50 value. [f]Gaps represent unsequenced regions between two contigs known to be adjacent due to associated paired ends. [g]Average pair distance is the average distance between two sections of BAC DNA separated by linker sequence. [h]Assembly based on large contigs (> 500 bp) consisting of ≥3 reads each.

3.4.4 Use of BAC-end sequences and minimum tiling path to confirm assembly and order of scaffolds

The accuracy of the final scaffold assembly was verified by conducting a BLAST search of the 126 BAC-end sequences against the scaffold builds. This also established the order of the four scaffolds relative to one another and confirmed that the aligned sequences followed the order predicted by the minimum tiling path of the eight BACs. These results provided further support for conservation of synteny and gene order of the seven genes in the genomes of Atlantic salmon, medaka, zebrafish and tiger pufferfish. Fig. 3-5 provides a visual summary of the data, including the minimum tiling path, sequence contigs, scaffolds, predicted genes and BAC-end sequences in the 1 Mb region.

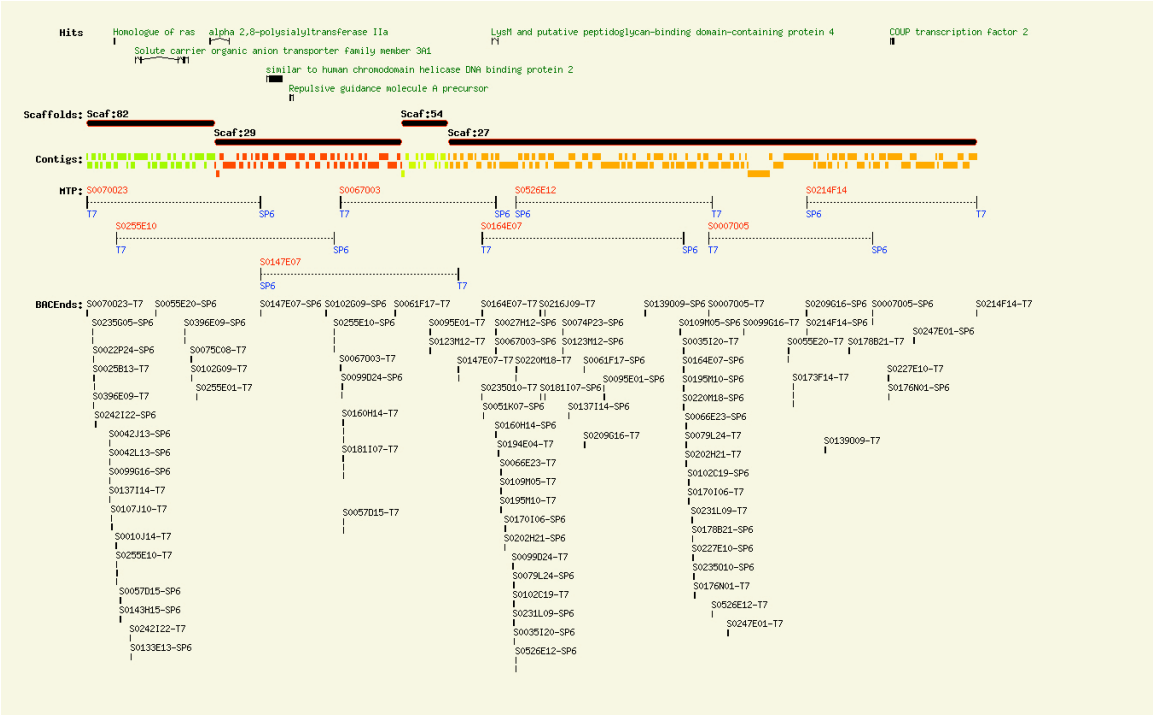

Figure 3-5 Summary of the 1 Mb sequenced region for the final assembly incorporating the GS FLX shotgun and paired end data with the 126 BAC-end sequences.

This figure summarizes all genes identified within the 1 Mb region and their position, the arrangement of the large scaffolds (order and orientation) as confirmed by the BAC-end sequences, the sequence contigs aligned against the scaffolds, the eight BACs of the minimum tiling path (MTP) including established overlap, and the BAC-end sequences within the region in the order predicted by the Atlantic salmon physical map.

3.4.5 Assembly and Annotation of the ninth BAC

Sanger sequencing of the shotgun library of the ninth BAC (S0022P24) in the minimum tiling path produced 3,524 confirmed reads and an average confirmed read length of 693.3 bp. PHRAP defines a confirmed read as verification of a read by another read with different chemistry or by an opposite-strand read [44]. This produced a ~10.5× depth of coverage given the estimated BAC size of 231,979 bp. The confirmed reads were assembled into 20 contigs with an average contig size of 8,885 bp and an N50 contig size of 32,866 bp; 14 contigs were defined as large contigs (i.e., > 500 bp). Nine large contigs consisting of three or more reads were assembled into two large scaffolds based on corresponding paired end reads from cloned inserts [GenBank: EU873552]. The average and N50 scaffold sizes were 112,155 and 137,857 bp, respectively. The two scaffolds were oriented relative to one another based on the locations of the T7 and SP6 BAC-end sequences.

The Sanger assembly produced a much larger average contig size and N50 contig size than any of the GS FLX assemblies (i.e., with and without paired end and BAC-end sequence reads), which corresponds to fewer contigs produced. This is likely because of the larger average read length of the Sanger sequences. The Sanger assembly produced two scaffolds with eight gaps for a ~230,000 bp region, whereas the final GS FLX assembly produced four scaffolds with 171 gaps for a ~1 MB region. Thus, with respect to the ability to establish the order and orientation of sequence contigs relative to one another, the GS FLX assembly was comparable to a Sanger-based assembly. This, however, was offset by the numerous gaps between contigs within the GS FLX assembly.

Sequence annotation using our in-house pipeline (described above) revealed hits to two genes: gonadotropin-releasing hormone receptor type I and a novel protein similar to vertebrate perilipin (Fig. 4-3a), with the latter located next to the final gene in the BACs sequenced by GS FLX. When the region was compared with regions that were previously identified as being syntenic with other sequenced fish genomes, only that of the zebrafish (*Danio rerio*) contained both genes. The remaining genomes (medaka, *Oryzias latipes*; tiger pufferfish, *Takifugu rubripes*; and stickleback, *Gasterosteus aculeatus*) only contained the gonadotropin-releasing hormone receptor type I gene with no evidence of the novel protein similar to perilipin or any other genes (Fig. 3-4b).

3.4.6 Nature of gaps in GS FLX assembly

A major concern is that 171 gaps remain between the GS FLX-sequenced contigs within the four final scaffolds. Given that GS 20, and by extension GS FLX, pyrosequencing is known to provide good coverage of genic regions [24], these gaps likely represent repeat regions rather than missed genes. This was supported by synteny analysis, which indicated that the initial assembly covered all genes present within this region in sequenced fish genomes, and by conducting a BLAST search of gap ends, which revealed that many of the gaps bordered known salmonid repetitive elements [10]. A comparison of the overlapping region between the BAC sequenced by the Sanger method and the corresponding region sequenced by GS FLX pyrosequencing (i.e., the region between the BAC-ends S0070O23-T7 and S0022P24-SP6 in Fig. 3-6), identified two gaps of 893 and 151 bp in the GS FLX assembly. These regions of the Sanger assembly were completely masked by the salmonid-specific repeat masker [45], thus verifying that the GS FLX technology has difficulty with repetitive regions.

Figure 3-6 Summary of the Sanger-sequenced BAC (S0022P24).

The two genes within the ~200,000 bp region are indicated as well as the nine sequence contigs and two scaffolds (indicated by red and green contigs). The relative orientation of these scaffolds was determined knowing the SP6 and T7 BAC-end sequences. The BAC-end sequences within the region are indicated in the order predicted by the Atlantic salmon physical map. Note that this BAC overlaps with the remainder of the MTP (i.e., that sequenced by GS FLX) at the 70O23-T7 BAC-end.

3.5 Conclusion

With 30–40% repetitive content and its pseudo-tetraploid nature due to a whole genome duplication event [2], the Atlantic salmon genome poses a significant challenge for sequencing. To date, the strategies to sequence complex vertebrate genomes have been Sanger sequencing of whole genome shotgun libraries (e.g., dog genome [46]), the generation of a library of cloned inserts such as BACs, followed by a 'map-first, sequence second' approach (e.g., pig genome [47]), or a combination of whole genome shotgun sequencing and pooled BAC sequencing [48]. These strategies are dependent on the minimal ability to sequence and assemble a full BAC insert. However, to date, this has proved unsuccessful with respect to complex genomes with any technique other than Sanger sequencing of a subcloned shotgun library [30].

The purpose of this study was to assess the feasibility of GS FLX pyrosequencing for *de novo* assembly of the Atlantic salmon genome given recent advances in read length and the availability of GS FLX Long Paired End technology. We demonstrated that without the inclusion of GS FLX Paired End reads, the GS FLX shotgun technology alone was substantially inferior to Sanger sequencing given the size and number of contigs produced and the inability to establish the relative order and orientation of the contigs. However, the addition of GS FLX Paired End reads vastly improved the capability of 454 pyrosequencing by enabling the assembly of contigs into large scaffolds. Indeed, in terms of the number of scaffolds produced, the GS FLX assembly that included the combined shotgun and paired end reads was comparable to the Sanger assembly. Moreover, the order of the GS FLX scaffolds could be established from information from BAC-end

sequences and the Atlantic salmon physical map. However, numerous gaps remained within the scaffolds, which is undesirable when a complete or reference genome sequence is one of the goals. Currently, if the Atlantic salmon genome is to provide a reference sequence for all salmonids, then a substantial proportion of the sequencing will have to be carried out using Sanger technology.

3.6 Supplementary Information

Appendix File 3 *Additional File 3-1*
Summary of information used for sequence annotation. Species, Ensembl names, assembly release date, Genebuild and database versions for all genome sequences used for comparative synteny analyses of the GS FLX shotgun + BAC-end sequence-generated contigs.

3.7 Acknowledgements

We gratefully acknowledge Kathy Bantle for her assistance with coordination of the project as well as Ken Dewar for comments on the manuscript. Roche/454 provided the GS FLX shotgun and Paired End sequencing and the authors affiliated with Roche/454 assisted with the study design, data collection, data analysis (bioinformatics) and the preparation of pertinent parts of the Methods section of the manuscript. All interpretation of the data and the decision to submit the manuscript for publication were done by researchers at Simon Fraser University independent of Roche/454. Funding for cGRASP (Consortium for Genomic Research on All Salmonids Project) was provided by Genome Canada and Genome BC.

3.8 References

1. Ohno S. Evolution by Gene Duplication. New York: Springer-Verlag; 1970.

2. Allendorf FW, Thorgaard GH. Tetraploidy and the evolution of salmonid fishes. In: Turner BJ, editor. Evolutionary Genetics of Fishes. New York: Plenum Press; 1984. pp. 55–93.
3. Thorgaard GH, Bailey GS, Williams D, Buhler DR, Kaattari SL, Ristow SS, Hansen JD, Winton JR, Bartholomew JL, Nagler JJ, Walsh PJ, Vijayan MM, Devlin RH, Hardy RW, Overturf KE, Young WP, Robison BD, Rexroad C, Palti Y. Status and opportunities for genomics research with rainbow trout. Comp Biochem Physiol B Biochem Mol Biol. 2002;133:609–646.
4. Thorsen J, Zhu B, Frengen E, Osoegawa K, de Jong PJ, Koop BF, Davidson WS, Høyheim B. A highly redundant BAC library of Atlantic salmon (*Salmo salar*): an important tool for salmon projects. BMC Genomics. 2005;6:50.
5. Ng SH, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJ, de Jong PJ, Koop BF, Krzywinski MI, Lubieniecki K, Marra MA, Mitchell LA, Mathewson C, Osoegawa K, Parisotto SE, Phillips RB, Rise ML, von Schalburg KR, Schein JE, Shin H, Siddiqui A, Thorsen J, Wye N, Yang G, Zhu B. A physical map of the genome of Atlantic salmon, *Salmo salar*. Genomics. 2005;86:396–404.
6. Atlantic salmon genome database http://www.ASalBase.org
7. Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SR, Smailus DE, Jones SJ, Schein JE, Marra MA, Butterfield YS, Stott JM, Ng SH, Davidson WS, Koop BF. Development and application of a salmonid EST database and cDNA microarray: Data mining and interspecific hybridization characteristic. Genome Res. 2004;14:478–490.
8. Atlantic Salmon EST Database http://web.uvic.ca/grasp/
9. Hardie DC, Hebert PD. The nucleotype effects of cellular DNA content in cartilaginous and ray finned fishes. Genome. 2003;46:683–706.
10. de Boer JG, Yazawa R, Davidson WS, Koop BF. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics. 2007;8:422.
11. Steinke D, Salzburger W, Meyer A. Novel relationships among ten fish model species revealed based on phylogenomic analysis using ESTs. J Mol Evol. 2006;62:772–784.
12. Hutchison CA., III DNA sequencing: bench to bedside and beyon. Nucleic Acids Res. 2007;35:6227–6237.
13. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008;18:1051–63.
14. Bennet S. Solexa Ltd. Pharmacogenomics. 2004;5:433–8.
15. Blow N. DNA sequencing: generation next-next. Nat Methods. 2008;5:267–274.
16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP,

Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in open microfabricated high density picoliter reactors. Nature. 2005;437:376–380.
17. Service RF. Gene sequencing: The race for the $1000 genome. Science. 2006;311:1544–1546.
18. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281:363–365.
19. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, Barbadora K, Klimke W, Dernovoy D, Tatusova T, Parkhill J, Bentley SD, Post JC, Ehrlich GD, Hu FZ. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol. 2007;189:8186–95.
20. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui L, Hutchison SK, Simons JF, Egholm M, Pettis JS, Lipkin WI. A metagenomic survey of microbes in honey bee colony collapse disorder. Science. 2007;318:283–287.
21. Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. Microbial population structures in the deep marine biosphere. Science. 2007;318:97–100.
22. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature. 2007;446:572–576.
23. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Pair-end mapping reveals extensive structural variation in the human genome. Science. 2007;318:420–426. Swaminathan K, Varala K, Hudson ME. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. BMC Genomics. 2007;8:132–145.
24. Torres TT, Metta M, Ottenwalder B, Schlotterer C. Gene expression profiling by massively parallel sequencing. Genome Res. 2008;18:172–177.
25. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S. Analysis of one million base pairs of Neanderthal DNA. Nature. 2006;444:330–336.
26. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM. Sequencing and analysis of neanderthal genomic DNA. Science. 2006;314:1113.
27. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-

Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Peer Y Van de, Salamini F, Viola R. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PloS One. 2007;12:e1326.
28. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452:872.
29. Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N. 454 sequencing put to the test using the complex genome of barley. BMC Genomics. 2006;7:275. doi: 10.1186/1471-2164-7-275.
30. Jackson TR, Ferguson MM, Danzmann RG, Fishback AG, Ihssen PE, O'Connell M, Crease TJ. Identification of two QTL influencing upper temperature tolerance in three rainbow trout (*Oncorhynchus mykiss*) half-sib families. Heredity. 1998;80:143–151.
31. Perry GML, Danzmann RG, Ferguson MM, Gibson JP. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Onchorhynchus mykiss*) Heredity. 2001;86:333–341.
32. Somorjai ML, Danzmann RG, Ferguson MM. Distribution of temperature tolerance quantitative trait loci in Arctic charr (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*) Genetics. 2003;165:1433–1456.
33. Sanchez JA, Clabby C, Ramos D, Blanco G, Flavin F, Vazquez E, Powell R. Protein and microsatellite single locus variability in *Salmo salar L*. (Atlantic salmon) Heredity. 1996;77:423–432.
34. Genomic Research on Atlantic Salmon Project (GRASP) website http://grasp.mbb.sfu.ca/
35. Repeatmasker http://www.repeatmasker.org
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
37. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78–94.
38. Uniprot http://www.pir.uniprot.org/database/nref
39. NCBI Conserved Domains Database http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd
40. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8:175–85.
41. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research. 1998;8:186–194.
42. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998;8:195–202.
43. PHRED/PHRAP instruction manual http://www.phrap.org/phredphrap/phrap.html
44. Salmonid-specific repeat masker http://grasp.mbb.sfu.ca/GRASPRepetitive.html

45. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, 3rd, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SM, Sutter NB, Thomas R, Webber C, Baldwin J, Abebe A, Abouelleil A, Aftuck L, Ait-Zahra M, Aldredge T, Allen N, An P, Anderson S, Antoine C, Arachchi H, Aslam A, Ayotte L, Bachantsang P, Barry A, Bayul T, Benamara M, Berlin A, Bessette D, Blitshteyn B, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Brown A, Cahill P, Calixte N, Camarata J, Cheshatsang Y, Chu J, Citroen M, Collymore A, Cooke P, Dawoe T, Daza R, Decktor K, DeGray S, Dhargay N, Dooley K, Dooley K, Dorje P, Dorjee K, Dorris L, Duffey N, Dupes A, Egbiremolen O, Elong R, Falk J, Farina A, Faro S, Ferguson D, Ferreira P, Fisher S, FitzGerald M, Foley K, Foley C, Franke A, Friedrich D, Gage D, Garber M, Gearin G, Giannoukos G, Goode T, Goyette A, Graham J, Grandbois E, Gyaltsen K, Hafez N, Hagopian D, Hagos B, Hall J, Healy C, Hegarty R, Honan T, Horn A, Houde N, Hughes L, Hunnicutt L, Husby M, Jester B, Jones C, Kamat A, Kanga B, Kells C, Khazanovich D, Kieu AC, Kisner P, Kumar M, Lance K, Landers T, Lara M, Lee W, Leger JP, Lennon N, Leuper L, LeVine S, Liu J, Liu X, Lokyitsang Y, Lokyitsang T, Lui A, Macdonald J, Major J, Marabella R, Maru K, Matthews C, McDonough S, Mehta T, Meldrim J, Melnikov A, Meneus L, Mihalev A, Mihova T, Miller K, Mittelman R, Mlenga V, Mulrain L, Munson G, Navidi A, Naylor J, Nguyen T, Nguyen N, Nguyen C, Nguyen T, Nicol R, Norbu N, Norbu C, Novod N, Nyima T, Olandt P, O'Neill B, O'Neill K, Osman S, Oyono L, Patti C, Perrin D, Phunkhang P, Pierre F, Priest M, Rachupka A, Raghuraman S, Rameau R, Ray V, Raymond C, Rege F, Rise C, Rogers J, Rogov P, Sahalie J, Settipalli S, Sharpe T, Shea T, Sheehan M, Sherpa N, Shi J, Shih D, Sloan J, Smith C, Sparrow T, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Stone S, Sykes S, Tchuinga P, Tenzing P, Tesfaye S, Thoulutsang D, Thoulutsang Y, Topham K, Topping I, Tsamla T, Vassiliev H, Venkataraman V, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Yang S, Yang X, Young G, Yu Q, Zainoun J, Zembek L, Zimmer A, Lander ES. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;8:803–819.
46. Porcine Genome Sequencing Project http://www.sanger.ac.uk/Projects/S_scrofa/
47. Rat Genome Sequencing Project Consortium Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004;428:493–521.

4: Identification of genes associated with heat tolerance in Arctic charr exposed to acute thermal stress

Published in: *Physiological Genomics* (2011) Vol. 43, No.11, pp. 685–96 ISSN 1471-2164.

Author list: Nicole L. Quinn[1], Colin R. McGowan[2], Glenn A. Cooper[3], Ben F. Koop[3] and William S. Davidson[1]
[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada
[2]Icy Waters Inc. Km. 4.2 Fish Lake Road, P.O. Box 21351, Station Main, Whitehorse, Yukon, Canada
[3]Department of Biology, University of Victoria, Victoria, British Columbia, Canada

Author contributions: NLQ, CRM, BFK and WSD conceived the project. NLQ and CRM performed the temperature trials and dissections. NLQ and GAC conducted the microarray analysis, NLQ conducted qPCR analysis. NLQ and GAC performed statistical analyses. NLQ, CRM, WSD and BFK interpreted the data. NLQ and WSD wrote the manuscript.

4.1 Abstract

Arctic charr is an especially attractive aquaculture species given that it features the desirable tissue traits of other salmonids, and is bred and grown at inland freshwater tank farms year round. It is of interest to develop upper temperature tolerant (UTT) strains of Arctic charr to increase the robustness of the species in the face of climate change, and to enable production in more southern regions. We used a genomics approach that takes advantage of the well-studied Atlantic salmon genome to identify genes that are associated with UTT in Arctic charr. Specifically, we conducted an acute temperature trial to identify temperature tolerant and intolerant Arctic charr individuals, which were subject to microarray and qPCR analysis to identify candidate UTT genes. These were compared with genes annotated in a QTL region that was previously identified as associated with UTT in rainbow trout and Arctic charr, and that we sequenced in Atlantic salmon. Our results suggest that small heat shock proteins as well as HSP-90 genes are associated with UTT. Furthermore, hemoglobin expression was significantly down-regulated in tolerant compared to intolerant fish. Finally, QTL analysis and expression profiling identified *COUP-TFII* as a candidate UTT gene, although its specific role is unclear given the identification of two transcripts, which appear to have different expression patterns. Our results highlight the importance of using more than one approach to identify candidate genes, particularly when examining a complicated trait such as UTT in a highly complex genome for which there is no reference genome.

4.2 Introduction

The salmonids (salmon, trout and charr) are of substantial environmental, economic and social value. They contribute to ecosystem health as well as to local and global economies through fisheries, aquaculture and sport fishing. Their increasing popularity as a food choice for humans has created a demand for salmonid flesh such as that of Atlantic salmon, rainbow trout, Pacific salmon species and, more recently, Arctic charr, which cannot be sustained by wild populations alone. This, combined with increasing environmental threats to wild populations has fueled the demand for sustainable and effective aquaculture methods. Genomics tools have long been successful in facilitating selective breeding for genetic improvement of cultured stocks (broodstocks) in agriculture species [e.g., swine (23) and cattle (47)]. As the availability of genomics resources has increased for aquatic species, more research is being done to improve aquaculture broodstocks for species such as Atlantic salmon (3), Atlantic cod (15), rainbow trout (46) and catfish (36, 41).

Arctic charr (*Salvelinus alpinus*) is an especially attractive aquaculture species given that it features the desirable tissue traits of other salmonids, commands a high market value, and is bred and grown at inland freshwater tank farms year round. This circumvents some of the adverse affects of marine net pen aquaculture, the current method used for most species of salmon (2). However, Arctic charr is a cold-water species that thrives in water temperatures from 0.1–14ºC, which presents substantial geographical limitations in terms of where this species can be grown at present. Tank farms that are otherwise equipped to grow and distribute freshwater fish species,

including salmonids such as rainbow trout, often cannot accommodate Arctic charr due to an unsuitable climate during at least part of the year and the high energy cost of maintaining tanks within the optimal temperature range for their survival. In addition, fish forced to live at temperatures higher than their natural range show signs of stress, including reductions in immune function, appetite, growth and reproduction, as well as susceptibility to disease and ultimately death (30). This is a problem of increasing concern, even in temperate regions where the species is currently farmed, as temperatures are rising as a result of climate change. As temperatures continue to climb and become less predictable, Arctic charr hatcheries and tank farms throughout the world will be faced with an on-going battle to keep fish alive, healthy and comfortable, and to keep them growing and spawning at the optimal rate. Arctic charr with different genetic backgrounds show markedly different abilities to withstand thermal stress (unpublished observations, CR McGowan). Understanding the genes involved in upper temperature tolerance (UTT) in Arctic charr as well as other salmonids stands to benefit both the aquaculture industry by facilitating the development of more robust broodstock, as well as natural populations, as such knowledge can feed into population-based conservation initiatives against the impacts of climate change.

The common ancestor of salmonids underwent a whole genome duplication event between 20 and 120 million years ago (1). Thus, the extant salmonid species are considered pseudo-tetraploids whose genomes are in the process of reverting to a stable diploid state. This, combined with the repetitive nature of the salmonid genomes in general (10) and the lack of a fully sequenced reference genome (9) makes identifying the genes responsible for complex traits such as UTT difficult. Common approaches to such

a task include the identification and analysis of quantitative trait loci (QTL), as well as expression analyses, which include microarray analysis and qPCR (22). However, individually, each of these methods has advantages and shortcomings, and may not provide the most accurate or comprehensive results on their own (see Discussion). In an attempt to circumvent these drawbacks while providing added confidence to QTL and expression data, it has become increasingly popular to use a combination of these methodologies, with the goal of identifying overlap between differentially expressed genes and QTL regions (17). The effectiveness of combining QTL and expression approaches for positively identifying candidate genes depends on many factors, including the resolution of the QTL analysis and the genome coverage provided by the expression analysis. In addition, the ability to detect a correlation between gene expression and QTL depends on the nature of the factor driving the QTL, an issue that is addressed within the Discussion of this paper.

Here, we adopted an approach that combines previously published QTL identification with expression profiling using the 32K GRASP microarray (20) and qPCR analysis to identify genes associated with UTT in Arctic charr. Specifically, QTL for UTT were previously identified in rainbow trout and Arctic charr (31, 42, 45). We mapped one of these QTL (that associated with markers SsaF43NUIG and Ssa20.19NUIG) to the same location in a homologous linkage group in Atlantic salmon (linkage group 23). We then used the Atlantic salmon genomic resources (7, 27, 39, 44) to identify nine BACs spanning a portion (that surrounding the SsaF43NUIG marker) of this QTL within the Atlantic salmon genome. The BACs were sequenced and annotated (38) thereby generating a list of putative UTT genes. In this study, we conducted an acute thermal trial

to identify tolerant and intolerant Arctic charr. RNA extracted from the gills of these fish was reverse transcribed into cDNA and used for microarray analysis, thus expanding the list of putative UTT genes identified previously. Gill tissue was chosen based on the results of a preliminary test, which indicated that of gill, liver and muscle tissues, gill would likely identify the most differentially expressed genes. Moreover, there is evidence in the literature to indicate that gill plays an important role in stress tolerance (4, 5). We reasoned that any genes identified by both the previously conducted sequencing of the QTL region and the current microarray study would be particularly strong candidates for UTT involvement. Finally, qPCR analysis was used to examine the behaviors of specific genes and to test the results of the microarray study. This combination of approaches enabled us to conduct an examination of the Arctic charr genome, and thus identify genes putatively involved in UTT with higher confidence than would be provided by a single approach. In addition, our results highlighted the strengths and potential pitfalls of each of these methods, particularly when studying a complex, duplicated genome for which there is, as yet, no reference sequence (9).

4.3 Materials and Methods

4.3.1 Mapping of SsaF43NUIG and Ssa20.19NUIG in Atlantic salmon

Previous reports identified UTT QTL in rainbow trout (31, 45) and Arctic charr (42), which were associated with markers Ssa20.19NUIG and SsaF43NUIG, respectively. These two markers, through their common association with marker CoCl3LAV, were found to map to the same location on homologous linkage groups (i.e., RT-10F and AC-

26F in rainbow trout and Arctic charr, respectively). We tested these markers for variability within the two Atlantic salmon SALMAP mapping families, Br5 and Br6, each of which contains two parents and 46 offspring (7). The forward primer for each pair contained an M13 sequence tag that was used for genotyping analysis. Genotyping results were analyzed with LINKMFEX ver. 2.3 (8).

4.3.2 Experimental design and tissue collection

All experiments were conducted according to the Canadian Council for Animal Care Guidelines, and were approved by the Animal Care Committee at Simon Fraser University, Canada. The temperature trial experiments were conducted at Icy Waters Inc., Whitehorse, Yukon, Canada in September, 2008 using 2006 young of the year Nauyuk Lake Arctic charr. Tanks were set up with a constant flow-through system (0.33 L/s) with fresh spring water at ambient temperature (approximately 6°C) and ambient oxygen levels (10.0–11.0 ppm). Approximately 200 fish were transferred to an experimental tank (diameter: 1.86 m, depth 50 cm) and left to acclimate for 48 h at ambient temperature. After acclimation, 10 fish were removed to act as a control group (hereafter referred to as Control fish), then water that had been diverted through a heat exchanger was added to the flow-through system to increase the water temperature in the tank by 6°C/h until it reached 22°C, then 0.5°C every 30 min until the water reached 25°C, the observed lethal temperature for these fish. Dissolved oxygen was allowed to fluctuate naturally and decreased from approximately 10.3 ppm to a minimum of 8.1 ppm, during the trial. Fish were not fed after being transferred to the experimental tank to avoid confounding gene expression results due to food metabolism.

When the water temperature reached 25°C, the temperature was held constant and the fish were closely monitored for signs of stress. The first and last 10 individuals to show loss of equilibrium (LOE) were quickly removed from the tank for sampling, thus representing the 5% least and most temperature tolerant fish, respectively (hereafter referred to as the Intolerant and Tolerant treatment groups, respectively). This temperature regime mimicked that conducted by the previous experiments that identified the UTT QTL (31, 42, 45) with some minor changes due to differences in the available equipment. The first LOE was observed after approximately 30 minutes at 25ºC, and the last fish showed LOE approximately 2 hr thereafter. Thus, it should be noted that Tolerant fish were exposed to lethal temperatures for up to 2 hr longer than Intolerant fish, and we therefore recognize that any genes identified as differentially expressed between Tolerant and Intolerant groups of fish may reflect this unavoidable difference in exposure time, rather than UTT itself. Fig. 4-1 is a schematic diagram of the experimental design. Fish were euthanized by a swift blow to the head, then weighed and their fork lengths were measured. Blood was withdrawn from the caudal vein of the fish (maximum possible volume ~200 μL), the entire lower half of outer-most gill arch was removed, the entire liver was sampled, and an approximately 1 cm^2 section of muscle from above the lateral line and behind the dorsal fin of the fish was removed, in that order. Tissues were placed into RNAlater® (Ambion Inc.) and were stored at room temperature for 24 h to allow RNAlater® to penetrate the tissues, and then moved to -80°C for storage until use as per the manufacturer's instructions.

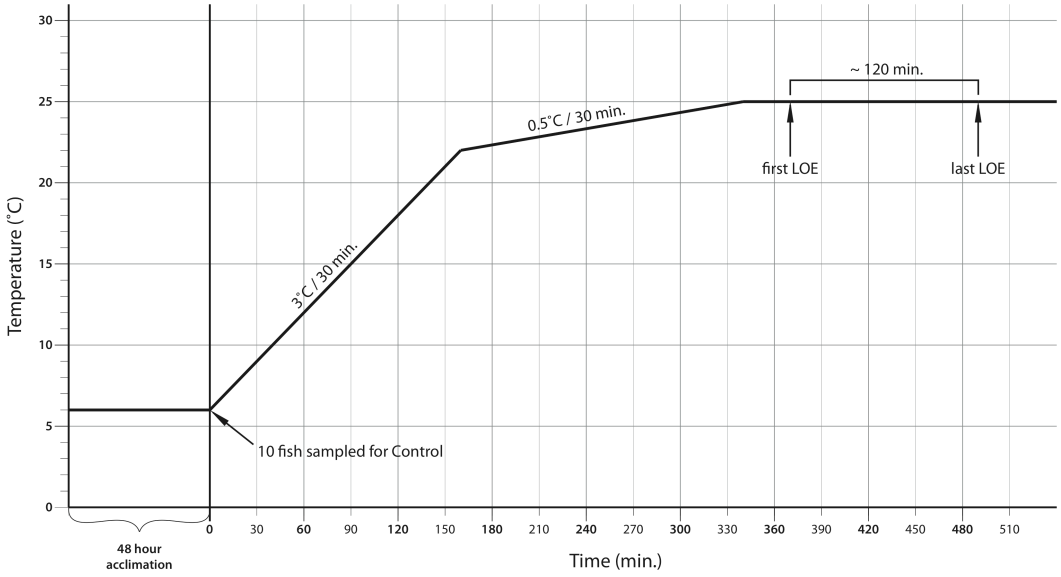

Figure 4-1 Schematic representation of the temperature profile for the UTT trials

Approximately 200 fish were transferred to an experimental tank (diameter: 1.86 m, depth 50 cm) and left to acclimate for 48 h at ambient temperature (6ºC). After acclimation, 10 fish were removed to act as a control group, then water that had been diverted through a heat exchanger was added to the flow-through system to increase the water temperature in the tank by 6°C/h until it reached 22°C, then 0.5°C every 30 min until the water reached 25°C, the observed lethal temperature for these fish. Dissolved oxygen was allowed to fluctuate naturally and decreased from approximately 10.3 ppm to a minimum of 8.1 ppm, during the trial. Fish were not fed after being transferred to the experimental tank to avoid confounding gene expression results due to food metabolism. When the water temperature reached 25°C, the temperature was held constant and the fish were closely monitored for signs of stress. The first and last 10 individuals to show loss of equilibrium (LOE) were quickly removed from the tank for sampling, thus representing the 5% least and most temperature tolerant fish, respectively (Intolerant and Tolerant treatment groups, respectively). The first LOE was observed after approximately 30 minutes at 25ºC, and the last fish showed LOE approximately 2 hr thereafter.

4.3.3 RNA Isolation

RNA isolations and microarray analysis were conducted at the University of Victoria, Canada. Total RNA was isolated from gill, muscle and liver tissue samples. Briefly, tissue samples were removed from RNAlater®, blotted on a clean Kimwipe™ to remove excess solution, and disrupted and homogenized in 1 mL TRIzol reagent using a Mixer-mill (Retch® MM 301) with tungsten carbide beads. Phase separation was conducted using 200 µL chloroform, and RNA was purified using the RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. Purified RNA was treated with 1 µL RNase inhibitor (Invitrogen). RNA integrity was verified by agarose gel with ethidium bromide staining to visualize ribosomal bands and by measuring the 260/280 absorbance ratio (>1.9) using a Nano Drop (ND-1000 Spectrophotometer, Thermo Scientifc) then stored at -80ºC until use.

4.3.4 Microarray analysis

The microarray study followed a reference design format. cDNA prepared from 1 µg gill RNA from six samples from each treatment group (Tolerant, Intolerant and Control; 18 slides in total) using Invitrogen's SuperScript Indirect cDNA labeling system. Treatment groups were compared indirectly against one another using a common reference sample that was hybridized to each microarray alongside the sample cDNA. The reference sample, designed to hybridize to as many spots on the array as possible, was comprised of high-quality RNA isolated from Atlantic salmon gonad, brain and spleen tissues that had been amplified using Ambion's Amino Allyl Message Amp™ aRNA kit, then

quantified, combined in equal amounts and divided into per-use aliquots to avoid degradation due to repeated freeze-thawing.

The GRASP 32K cDNA microarray was used (20). Details of the microarray hybridization process can be found at the University of Victoria cGRASP website (http://web.uvic.ca/grasp/microarray/array.html) within the .pdf document entitled *Invitrogen Indirect cDNA Labeling System version 3*. Briefly, slides were post-print processed by rinsing in 0.2% SDS and water and dried by centrifugation, then pre-hybridized in 5 x SSC, 0.1% SDS, 3% BSA, washed with water, dried again and stored in a dry oven at 49ºC until cDNA hybridization. cDNA (300 ng) and aRNA (500 ng) were labeled with Cy5 and Cy3 (Amersham Biosciences), respectively, using Invitrogen's SuperScript Indirect cDNA Labeling System, then combined with 2 x Formamide buffer and LNA dT blocker (Genisphere) to a total volume of 60 µL, which was heated to 80ºC then loaded on to the slide in the dark. Microarrays were incubated for 16 h at 49ºC in a dark, humidified chamber, then underwent a series of washes and were dried by centrifugation. Slides were scanned at 74 and 72 PMT for Cy3 and Cy5, respectively, using a ScanArray™ Express Microarray Scanner (Packard BioScience BioChip Technologies, model #ASCEX00) and spot intensity was calculated with ImaGene ver. 6.5.1.

A preliminary test was done to determine which tissue type would provide the most information for the analysis. Specifically, Cy5-labeled cDNA transcribed from RNA from gill, liver and muscle from a single individual from the Tolerant group was hybridized to three microarray slides along with the Cy3-labeled reference aRNA. The gill cDNA showed the highest number of Cy5 labeled spots, indicating that gill would

likely be the most informative tissue of the three in terms of revealing genes involved in thermal tolerance. Thus, gill tissues were used for all expression profiling (microarrays and qPCR).

4.3.5 Statistical analysis for identifying differentially expressed genes

All statistical analyses of the microarray data were conducted using Genespring ver. 7.3.1. Spots were identified as per the fully annotated gene ID file from the GRASP website (http://web.uvic.ca/grasp/microarray/array.html) (IE007 onwards; last modified on Nov. 3 2008). Signals were normalized per spot and per chip using an intensity-dependent (LOWESS) normalization, then per gene to normalize to the median. Spots were filtered on flags present, and only spots with signals greater or equal to the average base/proportional value of the raw channel were retained.

Pairwise student's t-tests were performed between each of the three treatment groups (i.e., Control vs. Tolerant, Control vs. Intolerant and Intolerant vs. Tolerant) with a *p*-value <0.01 (note that this is more stringent than a Bonferroni-corrected p-value of 0.05 with three comparisons), and any genes not meeting a two-fold differential expression between pairs were filtered out in the same step. Next, a Venn diagram was constructed to compare the resulting gene lists against one another. This approach enabled us to decipher transcripts that were identified by two or more of the pairwise comparisons (i.e., the overlapping portions of the Venn diagram), and thus would be more likely to represent general heat response genes. At the same time, we could generate lists of genes that only showed differential expression in one of the pairwise comparisons (i.e., the non-overlapping portions of the Venn diagram).

4.3.6 PCR, cloning and sequencing of multiple COUP-TFII transcripts

As described in the Results section, *COUP Transcription Factor II* (*COUP-TFII*) was identified by both sequencing of an UTT QTL region in Atlantic salmon and microarray analysis of Arctic charr as putatively playing a role in UTT. A search for the sequence of the *COUP-TFII* 596 bp EST from the microarray (GenBank accession number DW547089; hereafter referred to as transcript A) within the Atlantic salmon EST database (web.uvic.ca/grasp/, Project: *Salmo salar* – All 100/99) revealed two similar transcripts, the first corresponding to the EST spotted on to the array, and a second, contig19531 (hereafter transcript B), which consisted of a single read of 710 bp and showed 90% sequence identity with transcript A. Given the duplicated nature of the Atlantic salmon genome, we suspected that these two ESTs may represent duplicated genes. Although transcript B was not present in the 1 Mb region of the UTT QTL previously sequenced (38), and therefore the two transcripts do not reflect a tandem duplication, we do not know whether they are located further apart on the same chromosome, or whether they are located on separate chromosomes. Using the full Atlantic salmon genomic sequence of the *COUP-TFII* gene from our previous report (38) GenBank accession number EU481821.1), we designed PCR primers such that they would amplify the entire 596 bp segment of the EST that is on the microarray (transcript A). These primers were used to amplify the EST region by PCR using Arctic charr genomic DNA as a template. The PCR product was cloned using the pETBlue-1 AccepTor Vector kit (Novagen), individual clones were cultured, and the PCR inserts sequenced. This revealed that there were indeed two slightly different products in Arctic charr: one that was highly similar (97%) to the EST on the microarray (transcript A) and

one that was more similar to EST contig19531 (98%), or transcript B. Thus, qPCR reverse primers were specifically designed across and around an 11 bp gap in transcript B to specifically amplify transcripts A and B, respectively (the same forward primer was used for both transcripts; see Supplemental Table S4-1 for primer sequences). This ensured that only one product that was specific to a particular transcript was amplified by qPCR, which was verified by the presence of a single dissociation curve for each product.

4.3.7 Expression analysis using qPCR

We prepared cDNA for qPCR from 1 μg of total RNA using Invitrogen's SuperScript III Reverse Transcriptase kit following the manufacturer's instructions. The six RNA samples per treatment group that were used for microarray analysis were used along with RNA from three additional fish per treatment group (i.e., nine individuals per treatment group were tested with qPCR). qPCR primer pairs were designed for 24 genes selected based on interest in function as well as the degree of fold change observed in the microarray analysis. qPCR primer pairs were designed from the Atlantic salmon EST sequences used on the GRASP 32K array using Primer 3 version 0.4.0 (http://frodo.wi.mit.edu/primer3/). Primers were tested for amplification efficiency using cDNA generated from a single gill RNA sample using the qPCR conditions described below followed by a dissociation curve analysis to test for a single product for each primer pair and that no primer dimers were generated during the 40 amplification cycles. The 13 primer pairs meeting these criteria and showing the highest efficiencies (range 80.7–109.7; Supplemental File S4-1) were used for expression analysis of cDNAs from nine individuals from each treatment group. qPCR was conducted using the ABI 7900HT

system with Sybr green (Quanta Biosciences) under the following conditions: 95ºC for 3 min followed by 40 cycles of 95ºC for 15 s 60ºC for 30 s and 72ºC for 15 s, with one 96-well plate run per individual cDNA sample, which included, in triplicate, all 13 primers with corresponding no-template-controls (NTC), and two primer pairs for the endogenous control gene, $EF1A_A$. Specifically, we used the $EF1A_A$ primers designed by Olsvik et al. (29), which cross exons 5 and 6, and also designed primers that span exons 3 and 4 ($EF1A_A$3to4) (Supplemental Table S4-1). Having two primer sets within one gene serves as a control for the quality of the cDNA reverse transcription reaction because the expression of $EF1A_A$ should be the same using both primer pairs. The amplification efficiencies of the endogenous control primer sets $EF1A_A$ and $EF1A_A$3to4 were 98.6% and 98.8%, respectively. Each plate also contained a no-reverse-transcriptase control (i.e., RNA from the individual being tested that had gone through the steps of the cDNA preparation but lacking reverse trancriptase) to test for genomic DNA contamination of the cDNA. Also included was a linker sample, i.e., cDNA from a single individual amplified with $EF1A_A$, which was compared across all plates to test for technical variations between plates (average CT = 21.54 SD = 0.44).

4.3.8 Statistical analysis of qPCR results

qPCR results were analyzed using the ΔΔCt method (32) and calibrated for individual primer amplification efficiencies, producing a relative quantification (RQ) compared to a calibrator individual (a selected untreated control individual). RQ values were tested for outliers using the Box-Whisker method such that any data points falling outside of the 90% range were eliminated from the analysis. Remaining RQ values were log

transformed to meet the assumptions of the statistical tests. Pair-wise student's t-tests were performed between pairs among the three groups, Tolerant, Intolerant and Control. This statistical approach was used to maintain consistency with the microarray analyses, and enabled us to determine the degree to which the gene of interest (GOI) behaved similarly or differently between the microarray and qPCR analyses. All statistical analyses were performed using Graphpad Prism Ver.5.

4.4 Results

4.4.1 Mapping of SsaF43NUIG and Ssa20.19NUIG in Atlantic salmon

The microsatellite markers SsaF43NUIG and Ssa20.19NUIG were informative (i.e., variable) in both of the Atlantic salmon SALMAP mapping families, Br5 and Br6 (7), and mapped within 1.1 centimorgan of each other on linkage group 23 of Atlantic salmon, which corresponds to Atlantic salmon chromosome 16 (33) (Fig. 4-2). This provided strong evidence for a UTT QTL in this region of the Atlantic salmon genome. Thus, tiled Atlantic salmon BACs surrounding these markers were identified, nine of which spanning SsaF43NUIG were sequenced and annotated, as described in detail in our previous report (38).

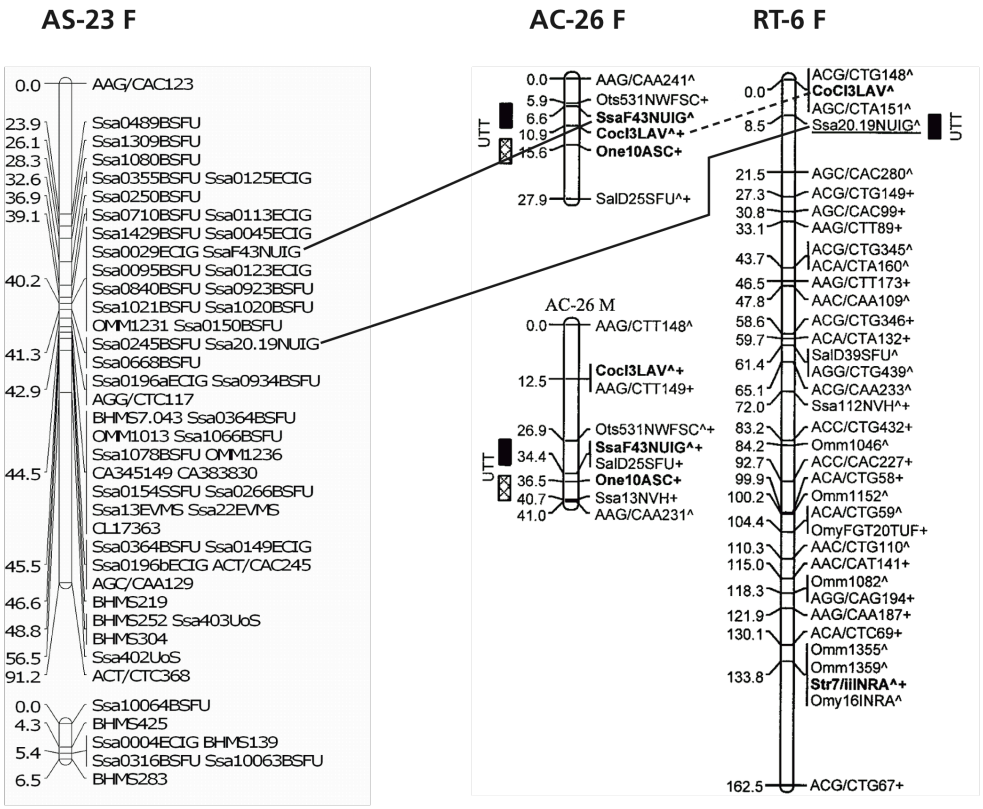

Figure 4-2 Comparative genetic analysis of UTT QTL markers in Arctic charr and rainbow trout and their location on the Atlantic salmon genetic map

Arctic charr linkage group AC-26F showing that microsatellite marker SsaF43NUIG is significantly associated with UTT and rainbow trout linkage group RT-6F showing that microsatellite marker Ssa20.19NUIG is significantly associated with UTT (42)These two linkage groups, through their common association with marker CoCl3LAV (dashed line), were found to be homologous. Both of the microsatellite markers SsaF43NUIG and Ssa20.19NUIG were informative (i.e., variable) in both of Atlantic salmon SALMAP mapping families, Br5 and Br6, and mapped within 1.1 centimorgan of each other on linkage group 23 of Atlantic salmon (solid lines), which corresponds to Atlantic salmon chromosome 16. This comparative genetic analysis provides strong evidence for a UTT QTL in this region of the Atlantic salmon genome.

4.4.2 Fish sizes

Fish weights per treatment group were as follows (average ± standard deviation): Tolerant 46.92 ± 10.62 g; Intolerant 42.89 ± 10.62 g; Control: 33.14 g ± 14.0 g. Fish lengths were: Tolerant 18.67 ± 0.97 cm; Intolerant 18.26 ± 14.72 cm; Control 16.32 ± 2.53 cm. There was no significant difference size between Tolerant and Intolerant fish (both weight and length), although Tolerant fish were larger than Control fish (1-way ANOVA followed by Tukey's post-hoc test p=0.0292 and p=0.0151 for weight and length, respectively; Supplemental Fig. S4-1).

4.4.3 Expression profiling of Tolerant, Intolerant and Control fish by microarray analysis

We were primarily interested in identifying genes that are associated with UTT, rather than genes that were differentially expressed regardless of the capacity of the individual fish to withstand thermal stress (i.e., general heat response genes). Thus, for the analysis of the microarray data, we focused on the gene lists contained in the three non-overlapping regions of the Venn diagram (Fig. 4-3), which represented the genes that were present in only one of the gene lists generated. However, all gene lists, including p-values and fold change values for the pair-wise comparisons are available in Supplemental Table S4-2. In addition, all microarray data (normalized as well as raw data) were deposited within Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE26306. From the

pair-wise comparisons among gene lists, genes were deemed noteworthy or interesting based on suspected function (i.e., genes that have been indicated in previous studies as playing a role in stress response, or those that play substantial roles in major biological pathways) as well as fold change and significance level (i.e., the most highly differentially expressed genes). These genes are indicated in the following paragraphs and some were further analyzed by qPCR as discussed later.

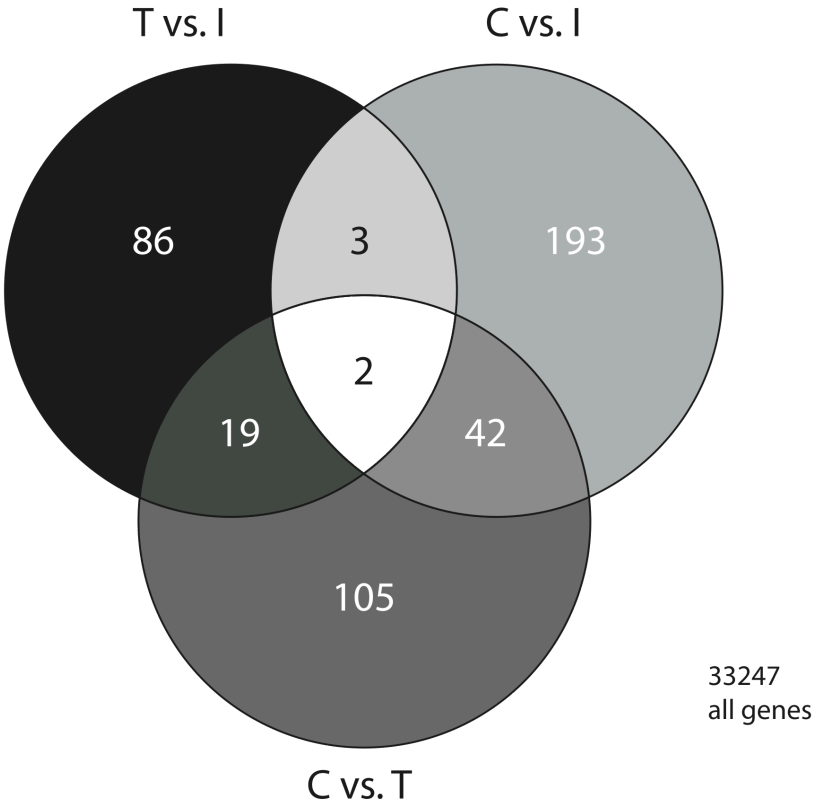

Figure 4-3 Venn diagram of three pair-wise comparisons of treatment groups

Venn diagram of three pair-wise comparisons of treatment groups (i.e., Tolerant vs. Intolerant, Tolerant vs. Control and Intolerant vs. Control; large circles). Numbers refer to the number of genes in that section. We were particularly interested in the gene lists in the non-overlapping regions, as these genes were more likely to be associated with differences in temperature tolerance, rather than a response to heat stress in general. The center of the diagram contained Hsp90-beta and Heat shock cognate 71 kDa protein (GenBank accession numbers CA062155 and EG813231, respectively). C, Control; I, Intolerant; T, Tolerant.

A total of 86 genes were differentially expressed only between Tolerant and Intolerant fish given the parameters assigned (i.e., 2 fold differential expression, $p<0.01$). Noteworthy genes that were up-regulated in Tolerant compared to Intolerant fish were *Pituitary homeobox 2a* (10.3 fold), *Actin - alpha cardiac muscle 1* (3.4 fold) and three heat shock proteins (Hsps), all of which belonged to the Hsp-beta family (2.03–2.8 fold). Genes that were down-regulated in Tolerant compared to Intolerant fish included seven beta hemoglobin subunits (2.7–6.6 fold) and seven alpha hemoglobin subunits (3.4–6.7 fold) (Fig. 4-4).

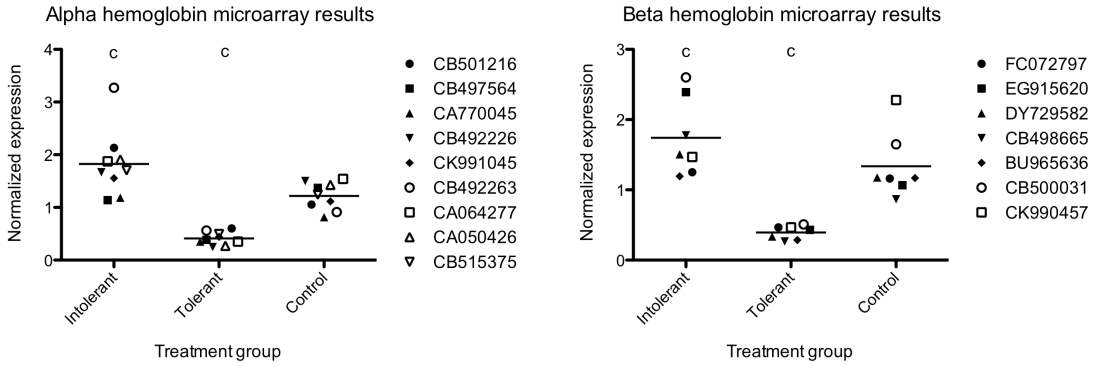

Figure 4-4 Hemoglobin microarray results

The mean (six individuals per treatment group) normalized (reference vs. raw fluorescence signal) expression values for alpha (A) and beta (B) hemoglobin genes significantly differentially expressed between Intolerant and Tolerant groups as determined by the microarray analysis.

A total of 105 genes were differentially expressed only between Tolerant and Control fish. Note that the Control group was comprised of randomly sampled fish that had not been subjected to heat stress, and thus contained an unknown mixture of genotypes. The genes on this list, therefore, may be associated with increased tolerance to heat stress, but do not necessarily distinguish temperature tolerant individuals from intolerant ones. Noteworthy genes up-regulated in Tolerant fish compared to untreated Controls included six Hsps [five Hsp90 genes (2.1–3.5 fold) and Hsp30 (3.41 fold)] as well as *ubiquitin* (2.5 fold) (Fig. 4-5), and *78 kDa glucose-regulated protein precursor* (3.1 fold). Genes that were down-regulated in Tolerant versus Control fish included two beta hemoglobin subunits (3.4 and 5.3 fold) and two alpha hemoglobin subunits (2.7 and 4.1 fold).

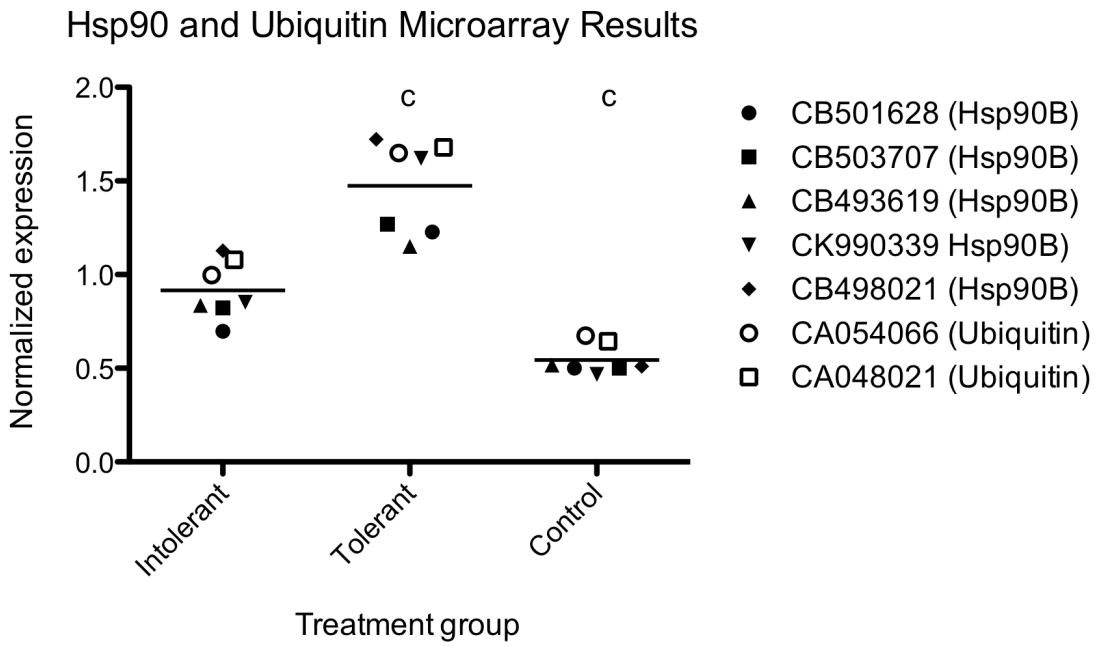

Figure 4-5 Hsp90-beta and ubiquitin microarray results

The mean (six individuals per treatment group) normalized (reference vs. raw) expression values plotted against treatment group for all Hsp90-beta genes as well as ubiquitin. C: Differential expression between groups meet significance parameters for microarray analysis (i.e., $p<0.01$, fold change >2.0)

The list of genes differentially expressed only between Intolerant fish and untreated Controls contained 196 genes. Of interest was that *COUP-TFII* was up-regulated in Intolerant fish compared to Controls (2.7 fold) as well as *Myosin heavy chain - fast skeletal muscle* and *Myosin heavy polypeptide 11 - smooth muscle* (3.6- and 2.2 fold, respectively) and *Heat shock 70 kDa protein* (5.7 fold). *Apoptosis-stimulating of p53 protein 1* was down-regulated in Intolerant compared to Control fish (4.2 fold).

4.4.4 Comparison of qPCR and microarray results among treatment groups

Supplemental Table S4-3 lists the genes tested by qPCR with their corresponding GenBank accession numbers, the microarray gene list that the genes were found on and the corresponding fold change. Also shown in Supplemental Table S4-3 are the results of the qPCR analysis with the average RQ value for each gene tested per treatment group and results (*p*-values) of pair-wise t-tests between groups. Results from the qPCR analysis that are in significant agreement with the microarray analysis are highlighted in yellow, while those showing the same trend as the microarrays are in orange, and those with no trend are left white. Of the 13 genes examined by qPCR, six showed differential expression in the same direction as indicated by the microarray analysis, six showed no trend and one showed significant differential expression in the opposite direction. Specifically, *Heat shock protein (Hsp)90-beta* showed highly significant ($p<0.0024$) up-regulation in Tolerant and Intolerant fish compared to Controls, with significantly higher expression in Tolerant compared to Intolerant fish (Fig. 4-6). Additionally, *78 kDa glucose-regulated protein precursor* showed down-regulation in both Tolerant and

Intolerant fish compared to controls (p<0.0001 and p=0.0378, respectively), and *Hsp11-beta* was significantly up-regulated in Tolerant vs. Intolerant fish (p=0.03). Also worth noting is that *Hsp90-beta*, which was one of two heat shock proteins in the centre of the Venn diagram and (i.e., significantly differentially expressed in all comparisons in the microarray analysis), showed the same patterns of significant differential expression by qPCR and microarray analyses.

Comparison of differential expressions of Hsp90-B (CA062155) as determined by microarray and qPCR analyses.

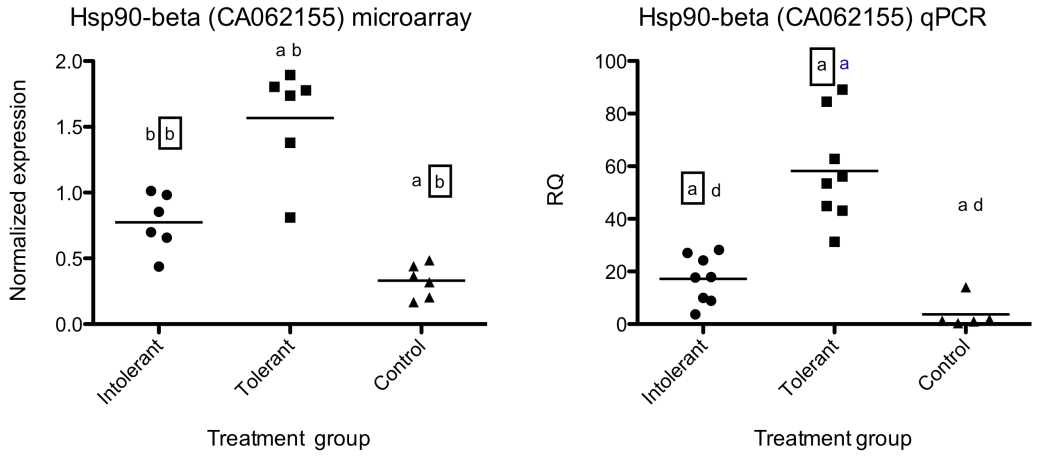

Figure 4-6 Microarray and qPCR results for Hsp90beta

Microarray and qPCR results for Hsp90beta (GenBank accession number CA062155) that was differentially expressed between all pair-wise comparisons (i.e., located in the center section of the Venn diagram). Corresponding boxed and unboxed lowercase letters indicate levels of significance for pair-wise comparisons at the following significance levels: a: $p<0.0001$, b: $p<0.001$, c: $p<0.01$, d: $p<0.05$.

The *COUP-TFII* transcript A (i.e., the transcript present on the microarray, see MATERIALS AND METHODS) showed significant up-regulation in Intolerant fish compared to Controls (p=0.0321), which is in accordance with the results from the microarray analysis. However, interestingly, the second transcript of *COUP-TFII*, transcript B, which was not on the microarray, exhibited significant differential regulation in the opposite direction, with decreased expression in Intolerant fish compared to Control fish as well as Tolerant fish (p=0.0018 and p=0.0003, respectively; Fig. 4-7). Finally, *Hsp7-beta* and *Actin, alpha cardiac muscle 1* showed trends of differential expression in the same direction as the microarrays, but did not reach statistical significance. Conversely, for *Apoptosis-stimulating of p53 protein 1* there was significant differential expression in the opposite direction as seen from the microarray analysis (i.e., up-regulated in Intolerant vs. Control fish; p=0.0022). The remaining genes showed very high p-values in all comparisons, (i.e., p>0.31), indicating that all treatment groups exhibited similar expression patterns.

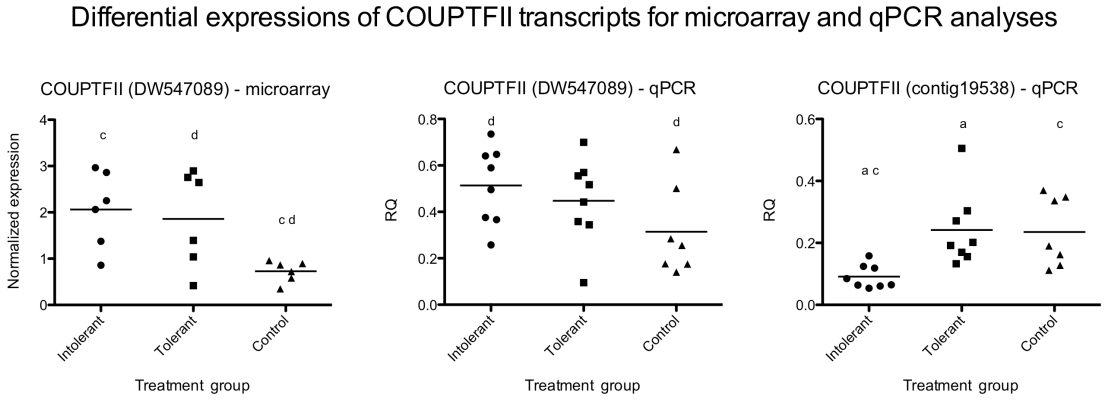

Figure 4-7 Microarray and qPCR results for *COUP-TFII*

Differential expression of two *COUP-TFII* transcripts between treatment groups for microarray and qPCR experiments. Horizontal lines indicate mean values. Corresponding lowercase letters indicate levels of significance for pair-wise comparisons at the following significance levels: a: $p<0.0001$, b: $p<0.001$, c: $p<0.01$, d: $p<0.05$.

4.5 Discussion

4.5.1 Up-regulation of heat shock proteins in thermo-tolerant fish

Hsps comprise a well-studied group of highly conserved, ubiquitously distributed proteins that are up-regulated when exposed to various stresses. In general, Hsps function as molecular chaperones, acting to maintain protein integrity during cellular stress conditions, including, but not limited to, elevated temperatures [reviewed in (40)]. In eukaryotes, Hsps are grouped into families according to their specific functions, sequence similarity and size (e.g., Hsp100, Hsp90, Hsp70 and Hsp60 with molecular weights of 100, 90, 70 and 60 kDa, respectively), along with the small Hsp (sHsp) group, which includes members that are 12–43 kDa (12, 40). Not surprisingly, the microarray analysis revealed several Hsps as differentially expressed in Arctic charr subjected to elevated temperatures. However, an interesting pattern emerged when the gene lists were analyzed using a Venn diagram.

First, sHsps were specifically associated with temperature tolerance as two *Hsp11-beta* transcripts and one *Hsp7-beta* were up-regulated in Tolerant vs. Intolerant fish by microarray analysis, and the qPCR results supported this relationship for one *Hsp11-beta* and the *Hsp7-beta* (Supplemental Table S4-2), whereas no other Hsps were present in this gene list. An additional *sHsp11* (also known as *Hsp30-beta*) gene was up-regulated in Tolerant vs. Control fish (microarray only; not tested by qPCR). sHsps exhibit a diverse range of structures, but share a conserved sequence of approximately 80 amino acid residues, the α-crystallin domain, located at the C-terminal of the protein

(14). Functionally, the sHsps display chaperone activities by interacting with unfolding proteins to maintain the folded state (16), while individual sHsps have been reported to play roles in cellular stress resistance and the inhibition of apoptosis (14). It has also been reported that *Hsp11-beta* was up-regulated in heat-shocked zebrafish (*Danio rerio*) embryos, whereas the expression *Hsp7-beta* did not change in response to heat shock (12, 24). Finally, *Hsp30* (a.k.a *Hsp11*) mRNA expression was elevated in the heart, brain, white muscle, red muscle and liver but not the blood of heat shocked adult rainbow trout (6). Thus, these sHsps appear to be transcribed in a tissue-specific manner in heat shocked fish. However, no reports of the activity of either of these genes in response to thermal stress in adult fish gills are currently available. Clearly, further analysis of these sHsps and their roles in heat tolerance in Arctic charr, including expression profiling at various life-stages in different tissues in response to a variety of heat stress regimes is warranted.

Second, the microarray analysis revealed that several members of the Hsp90 family followed the same pattern of expression as exhibited by the sHsps (i.e., Tolerant fish > Intolerant > Controls). This pattern was validated by qPCR for the *Hsp90-beta* transcript (GenBank accession number CA062155) that is located in the center of the Venn diagram (Fig. 4-3). For all other *Hsp90* genes, the increased expression between Tolerant and Control fish was statistically significant at $p<0.01$ (Fig. 4-5). In general, *Hsp90* genes function to maintain protein integrity, but they have also been reported to play roles in immune function, apoptosis and varying aspects of the inflammatory response in fish [reviewed in (40)]. Our results with the Tolerant fish showing elevated *Hsp90* expression compared to Intolerant and Control fish indicate that these particular

genes may be affiliated with temperature tolerance, and thus they merit further examination in terms of both determining the specific physiological roles of these Hsps as well as deciphering how, functionally and genetically, they differ from other *Hsp90*s.

Finally, we found that ubiquitin, a small regulatory protein found in all cells that tags proteins for recycling (18), showed the same pattern of expression as the *Hsp90* genes (i.e., Tolerant fish > Intolerant > Controls; Fig. 4-5). On one hand, this is not surprising given that ubiquitin is regularly used as an index of misfolded or damaged proteins, and thus is often found at levels similar to Hsps (35). On the other hand, it is interesting to note that Tolerant fish showed the highest levels of *ubiquitin* expression with significantly higher levels than Control fish.

It is worth noting again that, as a result of their prolonged survival, the Tolerant fish were exposed heat for longer, and therefore that the (or some of the) elevated expression of these stress-response genes may be a consequence of increased exposure to heat. Additionally, there is evidence that warm-adapted individuals of various taxa (including *Drosophila* and desert lizards) tend to show elevated constitutive levels of Hsps, rather than the extreme spikes that tend to be exhibited by cooler-adapted individuals [reviewed in (43)]. Thus, future research should focus on the particular role of ubiquitin as well as Hsps, and their relationships, in thermal tolerance (vs. thermal stress in general), as well as the genetic differences in upper temperature Tolerant compared to Intolerant Arctic charr.

4.5.2 Combining QTL and expression data

The effectiveness of combining QTL and expression approaches for positively identifying candidate genes depends on many factors, including the resolution of the QTL analysis and the genome coverage provided by the expression analysis. Additionally, the ability to detect a correlation between an expression QTL (eQTL) and a phenotypic QTL (pQTL) depends on the nature of the factor driving the QTL. Modifiers of gene expression can be cis- or trans-acting, which dictates whether the eQTL and pQTL coincide, and thus, whether a combined QTL/expression approach will result in Type II errors (i.e., false negative results) (48). Fig. 4-8 presents three scenarios to illustrate this point. Fig. 8A shows a situation in which the allele (represented by *) associated with the trait in question occurs in the coding region of the gene, which translates into a modified gene product. In this case, the mutation driving the pQTL is in cis (i.e., co-localized) with the measured gene. This situation results in a new gene product associated with the new allele, but no difference in the overall expression of the product. Thus, one would not expect to see any correlation between the pQTL and the eQTL, and a combination of QTL and expression analyses would not be useful in this situation. In the second scenario (Fig. 8B), in which there is a mutation in the promotor of the gene, again the mutation driving the pQTL and the gene in question are in cis; however, in this case, the binding affinity of the transcription regulator (e.g., transcription factor) would be altered, thus changing the expression of the gene, while the structure of the gene product itself remains unchanged. Thus, in this situation, the pQTL and eQTL co-localize, and the combination of QTL and expression analyses would be a powerful tool for identifying the gene of interest using these two independent methods. Finally, Fig. 8C illustrates a situation in

which the mutation driving the pQTL is in trans with the measured gene. Here, a change in a transcription factor results in differential expression of a gene at a separate locus. Although the eQTL and pQTL are not co-localized, an analysis of the genes that are differentially expressed may lead to the identification of a common pathway controlled at the level of transcription (13), and this form of pathway analysis could link the eQTL and pQTL. Therefore, combining or cross-referencing the results of genomics approaches, such as QTL and expression analyses, can be a powerful way to correlate results, particularly given the relatively high rates of Type I errors inherent in both of these approaches when conducted on their own, but one must be aware of the potential of Type II errors, and data interpretation as well as follow-up studies should be conducted accordingly.

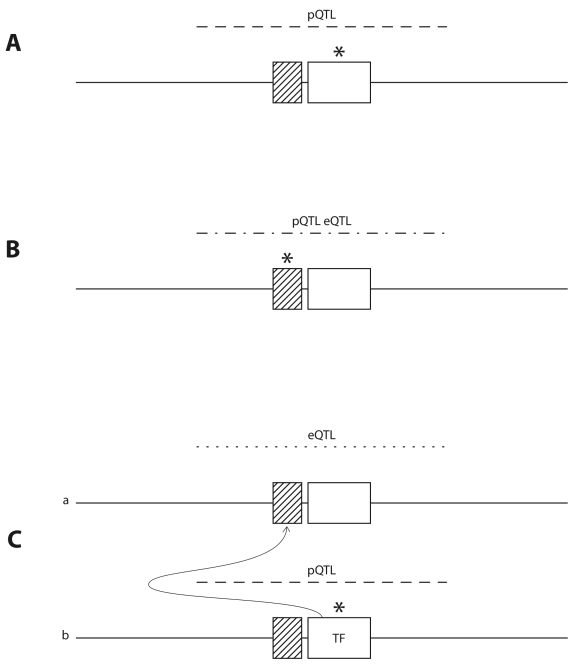

Figure 4-8 Possible associations between eQTL and pQTL

Fig. 4-8A shows a situation in which the variation associated with the trait in question (represented by *) occurs in the coding region of the gene, which translates into a modified gene product. In this case, the mutation driving the pQTL is in cis (i.e., co-localized) with the measured gene. This situation results in different gene products, one of which is associated with the trait of interest, but there is no change in the overall expression of the protein. Thus, the pQTL and qQTL will be independent of one another. Fig. 4-8B shows allelic variation in the promoter of a gene. The variation driving the pQTL and the gene in question are in cis; however, the binding affinity of a transcription regulator would be altered, resulting in differential expression of the gene, while the gene product itself remains unchanged. In this case, the pQTL and eQTL co-localize, and the combination QTL/expression approach would be a powerful tool for identifying the gene of interest using these two independent methods. Fig. 4-8C illustrates the situation in which the variation responsible for differential gene expression is in trans (i.e., a change in a transcription factor results in differential expression in a gene at a separate locus). Although the eQTL and pQTL are not co-localized, an analysis of the genes that are differentially expressed may lead to the identification of a common pathway controlled at the transcription level. This form of pathway analysis could link the qQTL and eQTL.

We were especially interested in the results for *COUP-TFII* (Chicken Ovalbumin Upstream Promoter Transcription Factor II), a member of the steroid/thyroid hormone receptor super-family, because 1) this gene was identified by both the sequencing of nine Atlantic salmon BACs spanning SsaF43NUIG (38), a marker which was significantly associated with a UTT QTL in rainbow trout (31, 45) and Arctic charr (42), and 2) our microarray analysis indicated that this gene was significantly up-regulated in Intolerant fish compared to untreated Control fish (2.7 fold; p=0.0025), and this was confirmed by qPCR (1.62 fold; p=0.0321) (Fig. 4-7). Thus, the pQTL affiliated with the marker SsaF43NUIG correlates with an eQTL associated with *COUP-TFII*. This situation corresponds to the scenario depicted in Fig. 4-8b, and the results thus implicate *COUP-TFII* gene as being associated with UTT in salmonids.

Although the physiological functions of *COUP-TF11* in mammals remain largely undefined due to the embryonic lethality of *COUP-TFII*-null mice, previous reports have implicated this gene in proper neuronal development as well as playing a role in apoptosis pathways (19), thus making it an ideal candidate for regulating the thermal stress response. The functions of *COUP-TFII* in fish, especially salmonids, remain largely unexplored, although there is some indication that it activates *hepatocyte nuclear factor* (*HPN*) *1*, which plays a pivotal role in liver development as well as in maintaining the differentiated hepatocyte phenotype (26). Further analyses of *COUP-TFII* and its role in UTT in salmonids are necessary, particularly with respect to the pathways in which it is involved.

We also found that the second transcript of *COUP-TFII*, corresponding to EST contig19538, which is not represented on the 32K GRASP micrarray and which we named Transcript B, showed significant differential expression in the opposite direction of the first transcript. That is, its expression was significantly down-regulated in Intolerant fish compared to Tolerant and Control fish. The apparently antagonistic expression patterns of the two *COUP-TFII* genes in salmonids subjected to elevated temperatures deserve more attention.

4.5.3 The role of hemoglobin genes in temperature tolerance in Arctic charr

It was previously predicted that decreased aerobic performance, a phenotype directly related to an organism's hemoglobin repertoire, would be a primary cause of extinction and relocation in warming habitats for marine fishes due to decreased oxygen supplies in warming waters (34). Our microarray analysis revealed that seven alpha and seven beta hemoglobin genes were significantly up-regulated in Intolerant vs. Tolerant fish, whereas two alpha and two beta hemoglobin genes were significantly down-regulated in Tolerant versus Control fish. Thus, Arctic charr that showed relative tolerance to acute thermal stress also showed significant down-regulation or reduced expression of hemoglobin genes (Fig. 4-4), with both alpha and beta hemoglobin genes behaving similarly. It is of particular interest that this differential expression of hemoglobin genes was observed in gill tissue, which is not considered a hematopoietic tissue and, to our knowledge, has not been identified as being transcriptionally active for hemoglobin. It is possible that these differences reflect hemoglobin transcription taking place in the blood trapped in the gills of the fish. That is, 1) the Intolerant fish could be expressing more hemoglobin genes in

their blood relative to Tolerant fish, or, 2) there could simply be more blood trapped in the gills of the Intolerant fish. With respect to 1) although we are not aware of any studies of hemoglobin expression in the blood of heat-shocked Arctic charr, Lewis et al. (21) conducted microarray analysis of blood isolated rainbow trout exposed to acute heat shock at various time points, and did not identify any hemoglobin genes as differentially expressed. This could suggest that hemoglobin transcription is not elevated in the blood of heat shocked salmonids, although clearly more research is needed to test this. With respect to 2), given that the Tolerant and Intolerant fish did not differ in size, we do not expect that more blood was trapped in the gills of the Intolerant fish due to the size of the gills alone, although it is also possible that Tolerant fish differed in their gill physiology, and were thus able to function with less blood in the gills, which could explain the lower levels of hemoglobin observed for these fish (28). These results therefore warrant further examination in the form of both expression profiling of individual hemoglobin transcripts in blood-only and gill-only RNA. This, however, would require transcript-specific hemoglobin primers for Arctic charr, which is not yet possible due to the lack of sequenced hemoglobin genes, as well as physiological studies examining the gill capacities of temperature Tolerant and Intolerant fish.

We have recently completed sequencing and annotating the full hemoglobin repertoire of the Atlantic salmon genome (37). There are more hemoglobin transcripts in Atlantic salmon than in any other fish genome studied thus far, and there are several non-Bohr beta hemoglobin genes in Atlantic salmon. We propose that these gene products act as emergency oxygen suppliers under conditions of high stress, such as that of increased temperature, decreased oxygen availability or in conditions where the fish is exerting

higher than normal levels of energy (37). Therefore, given that Arctic charr are a cold-adapted species and the results of this study suggest that reduced hemoglobin expression is associated with tolerance to acute heat stress, a full investigation of the Arctic charr hemoglobin repertoire, including sequencing of the genes as well as expression profiling, should be a priority for future research. Such an investigation would allow gene-specific hemoglobin primers to be designed, which could facilitate the identification of specific hemoglobin genes associated with UTT, a discovery with numerous implications for both cultured and wild Arctic charr and other salmonids.

4.6 Benefits and limitations of QTL analysis, microarrays and qPCR for identifying genes governing complex traits

QTL are useful as a starting point for identifying genes that govern a complex trait as they can provide information with regards to the chromosomal region of participating genes, as well as an estimate of the overall contribution of that region to the phenotype in question. However, gaining insight to the actual genes responsible for the QTL is resource intensive, particularly if there is no reference genome sequence available. Even if a genomic sequence is available and annotated, there are usually many genes within a QTL region and further experimentation is required to determine which, if any, of these genes is contributing to the trait in question. Expression profiling (microarray or qPCR analysis) provides a means of identifying eQTL. Cross referencing pQTL analysis with expression data may reveal whether *cis*- and *trans*-controlling elements determine the relative abundance of mRNA for a given gene (11). When pQTL and eQTL coincide, as is the case with the SsaF43NUIG UTT QTL (38, 42) and *COUP-TFII* in this study, this

provides cross-validation that the region, and the element identified therein, indeed contributes to the phenotype in question. However, as illustrated in Fig. 4-8, there are a number of situations in which a combination of approaches using QTL and expression analysis may not be informative, or may require further examinations, such as pathway analysis, to extract meaningful information.

On its own, microarray analysis acts as an excellent exploratory tool because thousands of genes can be tested at once. However, microarrays based on cDNA clones have inherent drawbacks, such that they are resource intensive and highly susceptible to technical variations as well as statistical pitfalls (25). The latter may result in Type I errors (false positive results), a particular concern when using a large microarray, as the false-positive rate increases as the number of spots increases. Indeed, this was a concern in this study, and might, at least in part, explain the lack of consistency between some of our microarray and qPCR data. Another challenge that we encountered was the potential for cross-hybridization between similar transcripts. It is possible, if not likely, that this occurred with the Hsp genes in the same families, the two *COUP-TFII* transcripts and the hemoglobin genes identified in our analysis. This may have muted evidence of differential expression between individual transcripts, and instead, produced the general trends observed for all highly similar genes on the microarray. In addition, this phenomenon may cause information to be lost if the elevated expression of a small number of family members is spread among all similar genes on the microarray such that no spots meet the statistical filters assigned (i.e., fold change and p-value cut-offs). The use of transcript-specific oligonucleotide arrays may overcome some of these challenges. An additional concern is introduced when the microarray being used was designed using

a different species than the one being tested, which increases the potential for cross-hybridization as well as lack of hybridization, another factor that may have played a role in our results. These are common concerns with microarray analysis, and as such, expression profiling using microarrays is sometimes described as an exploratory tool, for which the data must be interpreted with a caution, or which must be followed up or correlated with other types of analysis (21).

qPCR addresses some of the problems inherent in microarray analysis because it is highly specific and can tease apart the relationships between similar genes. However, to do so, the genes must first be identified through a process such as QTL or microarray analysis. Furthermore, as demonstrated by this study, when examining gene families or duplicated genes, it is often not feasible to design primers specific for each gene because the sequences of the genes must be available, a problem when there is no genome sequence or when the microarray only contains transcripts from a closely-related species.

4.7 Conclusion

We used a combination of approaches to identify candidate genes associated with UTT Arctic charr by cross-referencing genes identified within a sequenced UTT QTL region with expression data (microarray and qPCR) from temperature tolerant and intolerant individuals. Our results suggest that a number of sHsps as well as larger Hsp90 genes may be associated with tolerance to acute heat exposure. Furthermore, the microarray analysis provided clear evidence that hemoglobin genes (both alpha and beta) are significantly differentially expressed between Tolerant and Intolerant fish. *COUP-TFII* was identified by QTL sequencing as well as the microarray analysis as a candidate gene,

although its specific role is unclear given the subsequent identification of two transcripts, which show different expression patterns. Our results highlight the importance of using more than one approach to identify candidate genes, particularly when examining a complicated trait such as UTT in a species whose genome is highly complex and for which there is no genome sequence or even a suitable reference genome. Specifically, the lack of consistency between our microarray analysis and qPCR results strongly suggests that the results of microarray analysis should be further supported either by qPCR or QTL association, and that the use of microarray analysis should be limited to gene explorations by looking for groups of associated genes (e.g., gene families or those with shared pathways) that show similar trends in expression. Further examination of the physiological roles of these genes and their variants is necessary to be able to develop genomic markers associated with them. The results of this study can be incorporated into ongoing broodstock development programs to develop commercial strains of Arctic charr that can withstand warmer growing conditions, as well as used to screen wild populations for sensitivity to climate change, and potentially implement population-specific conservation initiatives.

4.8 Acknowledgements

The authors thank S Pavey, as well as the members of the Koop Laboratory, particularly K von Schalburg and B Sutherland, and members of the Davidson laboratory, particularly K Lubieniecki and K Johnstone, for technical assistance and help with data analysis. K Johnstone designed the EF1A_A3to4 primers.

4.9 Grants

This work was supported by funding for the Consortium for Genomic Research on All Salmonids Project (cGRASP) from Genome Canada and Genome BC and by a Strategic Grant from the Natural Sciences and Engineering Research Council of Canada.

4.10 Dislosures

CR McGowan is the Broodstock Development Manager for Icy Waters Inc. He accepts full responsibility for the conduct of the trial, has full access to all the data, and control over the decision to publish.

4.11 References

1. **Allendorf FW and Thorgaard GH.** Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes*Anonymous. New York, NY: Plenum Press, 1984, p. 55.

2. **Ayer NW and Tyedmers PH.** Assessing alternative aquaculture technologies: life cycle assessment of salmonid culture systems in Canada. 17: 362-363-373, 2009.

3. **Baranski M, Moen T and Vage DI.** Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (Salmo salar). *Genet.Sel.Evol.* 42: 17, 2010.

4. **Breves JP, Fox BK, Pierce AL, Hirano T and Grau EG.** Gene expression of growth hormone family and glucocorticoid receptors, osmosensors, and ion transporters in the gill during seawater acclimation of Mozambique tilapia, Oreochromis mossambicus. *J.Exp.Zool.A.Ecol.Genet.Physiol.* 313: 7: 432-441, 2010.

5. **Ching B, Chew SF, Wong WP and Ip YK.** Environmental ammonia exposure induces oxidative stress in gills and brain of Boleophthalmus boddarti (mudskipper). *Aquat.Toxicol.* 95: 3: 203-212, 2009.

6. **Currie S, Moyes CD and Tufts BL.** The effects of heat shock and acclimation temperature on hsp70 and hsp30 mRNA expression in rainbow trout: *in vivo* and *in vitro* comparisons. *J.Fish.Biol* 56: 398-408, 2000.

7. **Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Koop BF, Hoyheim B, Lien S, Lubieniecki KP, Moghadam HK, Park J, Phillips RB and Davidson WS.** Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). *BMC Genomics* 9: 557, 2008.

8. **Danzmann RG and Gharbi K.** Gene mapping in fishes: a means to an end. *Genetica* 111: 1-3: 3-23, 2001.

9. **Davidson WS, Koop BF, Jones SJ, Iturra P, Vidal R, Maass A, Jonassen I, Lien S and Omholt SW.** Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol.* 11: 9: 403, 2010.

10. **de Boer JG, Yazawa R, Davidson WS and Koop BF.** Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics* 8: 422, 2007.

11. **Druka A, Potokina E, Luo Z, Jiang N, Chen X, Kearsey M and Waugh R.** Expression quantitative trait loci analysis in plants. *Plant.Biotechnol.J.* 8: 1: 10-27, 2010.
12. **Elicker KS and Hutson LD.** Genome-wide analysis and expression profiling of the small heat shock proteins in zebrafish. *Gene* 403: 1-2: 60-69, 2007.

13. **Fisher P, Hedeler C, Wolstencroft K, Hulme H, Noyes H, Kemp S, Stevens R and Brass A.** A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Res.* 35: 16: 5625-5633, 2007.

14. **Franck E, Madsen O, van Rheede T, Ricard G, Huynen MA and de Jong WW.** Evolutionary diversity of vertebrate small heat shock proteins. *J.Mol.Evol.* 59: 6: 792-805, 2004.

15. **Hubert S, Higgins B, Borza T and Bowman S.** Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua). *BMC Genomics* 11: 191, 2010.

16. **Jakob U, Gaestel M, Engel K and Buchner J.** Small heat shock proteins are molecular chaperones. *J.Biol.Chem.* 268: 3: 1517-1520, 1993.

17. **Jouffe V, Rowe S, Liaubet L, Buitenhuis B, Hornshoj H, Sancristobal M, Mormede P and de Koning DJ.** Using microarrays to identify positional candidate genes for QTL: the case study of ACTH response in pigs. *BMC Proc.* 3 Suppl 4: S14, 2009.

18. **Kettern N, Dreiseidler M, Tawo R and Hohfeld J.** Chaperone-assisted degradation: multiple paths to destruction. *Biol.Chem.* 391: 5: 481-489, 2010.

19. **Kim BJ, Takamoto N, Yan J, Tsai SY and Tsai MJ.** Chicken Ovalbumin Upstream Promoter-Transcription Factor II (COUP-TFII) regulates growth and patterning of the postnatal mouse cerebellum. *Dev.Biol.* 326: 2: 378-391, 2009.

20. **Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmbhatt S, Rosner J, Rexroad CE,3rd, McGowan CR and Davidson WS.** A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* 9: 545, 2008.

21. **Lewis JM, Hori TS, Rise ML, Walsh PJ and Currie S.** Transcriptome responses to heat stress in the nucleated red blood cells of the rainbow trout (Oncorhynchus mykiss). *Physiol.Genomics* 42: 3: 361-373, 2010.

22. **Li J, Boroevich KA, Koop BF and Davidson WS.** Comparative Genomics Identifies Candidate Genes for Infectious Salmon Anemia (ISA) Resistance in Atlantic Salmon (Salmo salar). *Mar.Biotechnol.(NY)* 2010.

23. **Liu G, Jennen DG, Tholen E, Juengst H, Kleinwachter T, Holker M, Tesfaye D, Un G, Schreinemachers HJ, Murani E, Ponsuksili S, Kim JJ, Schellander K and Wimmers K.** A genome scan reveals QTL for growth, fatness, leanness and meat quality in a Duroc-Pietrain resource population. *Anim.Genet.* 38: 3: 241-252, 2007.

24. **Martin CC, Tsang CH, Beiko RG and Krone PH.** Expression and genomic organization of the zebrafish chaperonin gene complex. *Genome* 45: 5: 804-811, 2002.

25. **Martin LJ, Woo JG, Avery CL, Chen HS, North KE, Au K, Broet P, Dalmasso C, Guedj M, Holmans P, Huang B, Kuo PH, Lam AC, Li H, Manning A, Nikolov I, Sinha R, Shi J, Song K, Tabangin M, Tang R and Yamada R.** Multiple testing in the genomics era: findings from Genetic Analysis Workshop 15, Group 15. *Genet.Epidemiol.* 31 Suppl 1: S124-31, 2007.

26. **McNair A, Cereghini S, Brand H, Smith T, Breillat C and Gannon F.** Synergistic activation of the Atlantic salmon hepatocyte nuclear factor (HNF) 1 promoter by the orphan nuclear receptors HNF4 and chicken ovalbumin upstream promoter transcription factor I (COUP-TFI). *Biochem.J.* 352 Pt 2: 557-564, 2000.

27. **Ng SH, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJ, de Jong PJ, Koop BF, Krzywinski MI, Lubieniecki K, Marra MA, Mitchell LA, Mathewson C, Osoegawa K, Parisotto SE,**

Phillips RB, Rise ML, von Schalburg KR, Schein JE, Shin H, Siddiqui A, Thorsen J, Wye N, Yang G and Zhu B. A physical map of the genome of Atlantic salmon, Salmo salar. *Genomics* 86: 4: 396-404, 2005.

28. **Nilsson S and Sundin L.** Gill blood flow control. *Comp.Biochem.Physiol.A.Mol.Integr.Physiol.* 119: 1: 137-147, 1998.

29. **Olsvik PA, Lie KK, Jordal AE, Nilsen TO and Hordvik I.** Evaluation of potential reference genes in real-time RT-PCR studies of Atlantic salmon. *BMC Mol.Biol.* 6: 21, 2005.

30. **Pankhurst NW and King HR.** Temperature and salmonid reproduction: implications for aquaculture. *J.Fish Biol.* 76: 1: 69-85, 2010.

31. **Perry GM, Danzmann RG, Ferguson MM and Gibson JP.** Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (Oncorhynchus mykiss). *Heredity* 86: Pt 3: 333-341, 2001.

32. **Pfaffl MW.** A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29: 9: e45, 2001.

33. **Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF, Danzmann RG and Davidson WS.** Assignment of Atlantic salmon (Salmo salar) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (Oncorhynchus mykiss). *BMC Genet.* 10: 46, 2009.

34. **Portner HO and Knust R.** Climate change affects marine fishes through the oxygen limitation of thermal tolerance. *Science* 315: 5808: 95-97, 2007.

35. **Pratt WB, Morishima Y, Peng HM and Osawa Y.** Proposal for a role of the Hsp90/Hsp70-based chaperone machinery in making triage decisions when proteins undergo oxidative and toxic damage. *Exp.Biol.Med.(Maywood)* 235: 3: 278-289, 2010.

36. **Quiniou SM, Waldbieser GC and Duke MV.** A first generation BAC-based physical map of the channel catfish genome. *BMC Genomics* 8: 40, 2007.

37. **Quinn NL, Boroevich KA, Lubieniecki KP, Chow W, Davidson EA, Phillips RB, Koop BF and Davidson WS.** Genomic organization and evolution of the Atlantic salmon hemoglobin repertoire. *BMC Genomics* 11: 539, 2010.

38. **Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, Harkins TT and Davidson WS.** Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* 9: 404, 2008.

39. **Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SR, Smailus DE, Jones SJ, Schein JE, Marra MA, Butterfield YS, Stott JM, Ng SH, Davidson WS and Koop BF.** Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res.* 14: 3: 478-490, 2004.

40. **Roberts RJ, Agius C, Saliba C, Bossier P and Sung YY.** Heat shock proteins (chaperones) in fish and shellfish and their potential role in relation to fish health: a review. *J.Fish Dis.* 33: 10: 789-801, 2010.

41. **Serapion J, Kucuktas H, Feng J and Liu Z.** Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (Ictalurus punctatus). *Mar.Biotechnol.(NY)* 6: 4: 364-377, 2004.

42. **Somorjai IM, Danzmann RG and Ferguson MM.** Distribution of temperature tolerance quantitative trait loci in Arctic charr (Salvelinus alpinus) and inferred homologies in rainbow trout (Oncorhynchus mykiss). *Genetics* 165: 3: 1443-1456, 2003.

43. **Sørensen JG.** Application of heat shock protein expression for detecting natural adaptation and exposure to stress in natural populations. *Curr Zool* 56: 6: 703-713, 2010.

44. **Thorsen J, Zhu B, Frengen E, Osoegawa K, de Jong PJ, Koop BF, Davidson WS and Hoyheim B.** A highly redundant BAC library of Atlantic salmon (Salmo salar): an important tool for salmon projects. *BMC Genomics* 6: 1: 50, 2005.

45. **Timothy R. Jackson, Moira M. Ferguson, Roy G. Danzmann, Anthony G. Fishback, Peter E. Ihssen, Michael O'Connell and Teresa J. Crease.** Identification of two QTL influencing upper temperature tolerance in three rainbow trout (Oncorhynchus mykiss) half-sib families. 80: 143-144-151, 1998.

46. **Vallejo RL, Rexroad CE,3rd, Silverstein JT, Janss LL and Weber GM.** Evidence of major genes affecting stress response in rainbow trout using Bayesian methods of complex segregation analysis. *J.Anim.Sci.* 87: 11: 3490-3505, 2009.

47. **Veerkamp RF and Beerda B.** Genetics and genomics to improve fertility in high producing dairy cows. *Theriogenology* 68 Suppl 1: S266-73, 2007.

48. **Verdugo RA, Farber CR, Warden CH and Medrano JF.** Serious limitations of the QTL/microarray approach for QTL gene discovery. *BMC Biol.* 8: 96, 2010.

4.12 Supplemental Data

Appendix File 4 *Supplemental Fig. S4-1.*

Comparisons of fish weights and lengths between treatment groups (mean, SD).

Appendix File 5 *Supplemental Table S4-1*
Sequences and efficiencies of primers used for qPCR

Appendix File 6 *Supplemental Table S4-2.*
Lists of significantly differentially expressed genes with corresponding fold changes and *p*-values generated by pair-wise comparisons among treatment groups (C = control, I = Intolerant, T = tolerant) then further analyzed using a Venn diagram. A: I vs. T only (i.e., significantly differentially expressed genes from the non-overlapping portion of the Venn diagram comparing Intolerant and Tolerant fish); B: T vs. C only; C: I vs. C only; D: C vs. T and T vs. I (i.e., significantly differentially expressed genes from the segment of the Venn diagram that overlaps the C vs. T and T vs. I pair-wise comparisons); E: C vs. I and T vs. I; F: C vs. I and C vs. T; G: all pair-wise comparisons (i.e., genes from the center section of the Venn diagram).

Appendix File 7 *Supplemental Table S4-3*
The genes tested by qPCR with their corresponding GenBank accession numbers, the microarray gene list that the genes were found on and the corresponding fold change, the results of the qPCR analysis with the average RQ value for each gene tested per treatment group and results (*p* values) of pair-wise t-tests between groups. Results from the qPCR

analysis that are in significant agreement with the microarray analysis are highlighted in yellow, while those showing the same trend as the microarrays are in orange, and those with no trend are left white. Of the 13 genes examined by qPCR, six showed differential expression in the same direction as indicated by the microarray analysis, six showed no trend and one showed significant differential expression in the opposite direction.

5: Ribosomal genes and heat shock proteins as putative markers for chronic, sub-lethal heat stress in Arctic charr: applications for aquaculture and wild fish

Published in: *Physiological Genomics* (2011) Vol. 43, No.18, pp. 1056–1064 ISSN 1056-1064.

Author list: Nicole L. Quinn[1], Colin R. McGowan[2], Glenn A. Cooper[3], Ben F. Koop[3] and William S. Davidson[1]
[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada
[2]Icy Waters Inc. Km. 4.2 Fish Lake Road, P.O. Box 21351, Station Main, Whitehorse, Yukon, Canada
[3]Department of Biology, University of Victoria, Victoria, British Columbia, Canada

Author contributions: NLQ, CRM, BFK and WSD conceived the project. NLQ and CRM performed the temperature trials and dissections. NLQ and GAC conducted the microarray analysis, NLQ conducted qPCR analysis. NLQ and GAC performed statistical analyses. NLQ, CRM, WSD and BFK interpreted the data. NLQ and WSD wrote the manuscript.

5.1 Abstract

Arctic charr thrive at high densities and can live in freshwater year round, making this species especially suitable for inland, closed containment aquaculture. However, it is a cold water salmonid, which both limits where the species can be farmed and places wild populations at particular risk to climate change. Previously, we identified genes associated with tolerance and intolerance to acute, lethal temperature stress in Arctic charr. However, there remained a need to examine the genes involved in the stress response to more realistic temperatures that could be experienced during a summer heat wave in grow-out tanks that are not artificially cooled, or under natural conditions. Here, we exposed Arctic charr to sub-lethal heat stress of 15–18ºC for 72 hours, and gill tissues extracted before, during (i.e., at 72 hrs), immediately after cooling and after 72 hours of recovery at ambient temperature (6ºC) were used for gene expression profiling by microarray and qPCR analyses. The results revealed an expected pattern for heat shock protein (Hsp) expression, which was highest during heat exposure, with significantly reduced expression (approaching control levels) quickly thereafter. We also found that the expression of numerous ribosomal proteins was significantly elevated immediately and 72 hrs after cooling, suggesting that the gill tissues were undergoing ribosome biogenesis while recovering from damage caused by heat stress. We suggest that these are candidate gene targets for the future development of genetic markers for broodstock development or for monitoring temperature stress and recovery in wild or cultured conditions.

5.2 Introduction

Arctic charr (*Salvelinus alpinus*) is an attractive aquaculture species because it features the desirable tissue traits of other salmonids (e.g., salmon and trout), commands a high market value, and is bred and grown at inland freshwater tank farms year round. This circumvents some of the adverse affects of marine net pen aquaculture, the current method used for most species of salmon (1). However Arctic charr is a cold-water species that thrives in water temperatures from 0.1–14ºC, which presents substantial geographical limitations in terms of where it can currently be grown. Tank farms that are otherwise equipped to grow and distribute freshwater fish species, including salmonids such as rainbow trout, often cannot accommodate Arctic charr due to an unsuitable climate during at least part of the year and the high energy cost of maintaining tanks within the optimal temperature range to thrive. In addition, fish forced to live at temperatures higher than their natural range show signs of stress, including reductions in immune function, appetite, growth and reproduction, as well as susceptibility to disease and ultimately death (10). This is a problem of increasing concern, even in temperate regions where the species is currently farmed, as temperatures are rising as a result of climate change (13). Indeed, an Arctic charr aquaculture facility in the Yukon, Canada has recorded tank temperatures as high as 18ºC for several consecutive days during past summer heat waves (unpublished observation, Colin McGowan). As temperatures continue to climb and become less predictable, Arctic charr hatcheries and tank farms throughout the world will be faced with an on-going battle to keep fish alive and healthy, and to keep them growing

and spawning at the optimal rate. These issues have highlighted the need for a temperature tolerant Arctic charr broodstock that can thrive in temperatures outside the natural range for the species, and can withstand fluctuations in temperature that are not normally observed in their natural habitat. This could improve animal health and welfare, while simultaneously reducing costs and resource use for the Arctic charr aquaculture industry. In addition, wild Arctic charr living in native habitats are also expected to experience substantial thermal stress as climates in Arctic regions climb, suggesting the need for biomarkers that can effectively test for temperature stress among wild populations, which could facilitate population-based conservation initiatives as well as studies of population movement, colonization or evolution as habitats change.

We recently identified genes putatively associated with tolerance and intolerance to acute thermal stress in Arctic charr (16). Specifically, we used a combination of QTL association and expression analysis to examine genes differentially expressed in fish that were more and less tolerant to acute exposure to upper lethal temperature (~25ºC). Our results suggested that small heat shock proteins (Hsps) as well as Hsp-90 genes are associated with tolerance to extreme heat, whereas, hemoglobin expression was significantly down-regulated in tolerant compared to intolerant fish. Additionally, QTL analysis and expression profiling identified *COUP-TFII* as a candidate gene transcription factor involved in acute upper temperature tolerance (16). This information is valuable in that it may be used to generate markers specifically associated with upper temperature tolerance and intolerance, which can be integrated into broodstock programs. However, these fish were exposed to lethal temperatures that are far more extreme than those that would normally be experienced by any Arctic charr in an aquaculture or natural setting.

Thus, there remains a need to examine the genes involved in the stress response to more realistic temperature stress situations such as those that may be experienced during a summer heat wave in grow-out tanks that are not artificially cooled, or under natural conditions. It is also of interest to examine genes involved in recovery from chronic, sub-lethal heat stress for purposes such as the identification of populations of Arctic charr that are better able to recover from such stress, or that show different means of recovering.

Here, we set out to identify genes that are differentially expressed during and after exposure to prolonged sub-lethal heat stress in Arctic charr. Specifically, we exposed Arctic charr to temperature stress that mimicked that which has been recorded at an Arctic charr aquaculture facility in the Yukon, Canada for three days, after which the fish were allowed to recover at ambient temperatures for an additional three days. Tissue samples (blood, gill, liver, muscle) were taken before and during (72 hrs) elevated temperatures (15–19ºC), then immediately after and 72 hrs after returning to ambient water temperature (~6ºC). RNA extracted from the gill tissues was reverse transcribed into cDNA, which was used for microarray analysis followed by qPCR to identify genes that were differentially expressed between temperature conditions. Note that this heat challenge is distinctly different from the acute trial (16) with respect to the exposure temperature (lethal vs. sub-lethal), exposure time (acute vs. chronic) and endpoint (phenotypically tolerant and intolerant individuals identified vs. general response at different time points during and after exposure to an elevated temperature); however, the experimental fish population, all environmental conditions except for temperature, and expression profiling analyses were identical, thereby allowing comparisons of results between the two studies.

This study provides a general overview of the genes that are involved in chronic temperature stress and recovery, which allowed us to identify candidate gene targets for the future development of potential genetic markers that can be integrated into an ongoing Arctic charr broodstock development program to generate more robust aquaculture fish and/or that could serve as biomarkers for monitoring temperature stress and recovery in wild or cultured conditions.

5.3 Methods

5.3.3 Fish, temperature profile and tissue sampling

All experiments were conducted according to the Canadian Council for Animal Care Guidelines, and were approved by the Animal Care Committee at Simon Fraser University, Canada. The temperature trial experiments were conducted at Icy Waters Ltd., Whitehorse, Yukon, Canada in September, 2008 using 2006-born Nauyuk Lake Arctic charr. The average weight (± standard deviation) of the randomly sampled fish was 32.05 g (± 12.97 g), while the average length was 16.09 cm (± 2.27 cm). Tanks were set up with a constant flow-through system (0.33 L/s) with fresh spring water at ambient temperature (approximately 6°C) and ambient oxygen levels (10.0–11.0 ppm). Approximately 250 fish were transferred to an experimental tank (diameter: 1.86 m, depth 50 cm) and left to acclimate for 48 h at ambient temperature (~6ºC). Note that this represents a much lower stocking density than the Arctic charr aquaculture industry standard of 100–185 kg/m^3, and therefore that oxygen supply was not an issue. After acclimation, 10 fish were randomly selected to act as a control group (hereafter referred

to as treatment group C), then water that had been diverted through a heat exchanger was added to the flow-through system to increase the water temperature in the tank by approximately 3.0ºC per hour for 4 hours, until the tank reached 18ºC. Note that this represents the maximum ambient temperature (i.e., that in non-temperature controlled grow-out tanks) observed at Icy Waters during a summer heat wave. The observed lethal temperature for these fish is 25–26ºC (16).

We maintained the water temperature for the next 72 hrs (minimum and maximum recorded temperatures during the trial were 15ºC and 19ºC, respectively) and the fish were closely monitored for signs of stress. During the heat exposure, we observed that the Arctic charr, which normally school in tight packs, swam relatively dispersed within the tank, and we interpreted this change in schooling behavior as a response to stress. In addition, ten fish died during the 72-hour trial. These fish were weighed, their fork lengths measured and fin-clips were taken and stored in 95% ethanol, but no tissues were taken for expression analysis because it was assumed that RNA integrity would be compromised. After 72 hours, 10 fish were randomly selected for sampling (treatment group D) and the heat exchanger was turned off. The water was allowed to return to ambient temperature (~6ºC) overnight (12 hrs), and 10 more fish were randomly sampled (treatment group A). After 72 hrs at ambient temperature, 10 additional fish were randomly selected (treatment group R), which enabled us to examine genes involved in recovery from chronic, moderate heat stress. The temperature protocol and definition of treatment groups are shown in Figure 5-1. The sexes of the sampled fish were unknown, but were assumed to be in a 50:50 ratio given random selection. Prior to and throughout the trial, the fish were maintained under 24 hrs of daylight conditions (indoors), which is

the aquaculture industry standard for fish at this life stage. Dissolved oxygen was allowed to fluctuate naturally and decreased from approximately 10.3 ppm to a minimum of 9.4 ppm, during the trial. Fish of this size are normally fed approximately 0.8% of their body weight per day; however, these fish were not fed after transfer to the experimental tank for acclimation to avoid confounding gene expression results due to food metabolism. This scenario reflected what would happen when an actual heat wave is experienced within an aquaculture setting. That is, when temperatures become extreme, the fish are taken the fish off their feed so that oxygen levels in the tank stay high (less metabolism) and the fish can devote all energy to survival, rather than growth. There tends to be loss of growth during these periods, but mortality is substantially less, especially if starving can be induced in advance of the heat wave (unpublished observation, Colin McGowan). Thus, our protocol followed standard husbandry practices. Finally, note that this experimental design does not incorporate time-matched control groups, but rather an initial, untreated control that is used for all pairwise comparisons. Thus, any differential gene expression resulting simply from time spent in the tank (i.e., 144 hrs by the end of the recovery period) without food cannot be filtered out of the data.

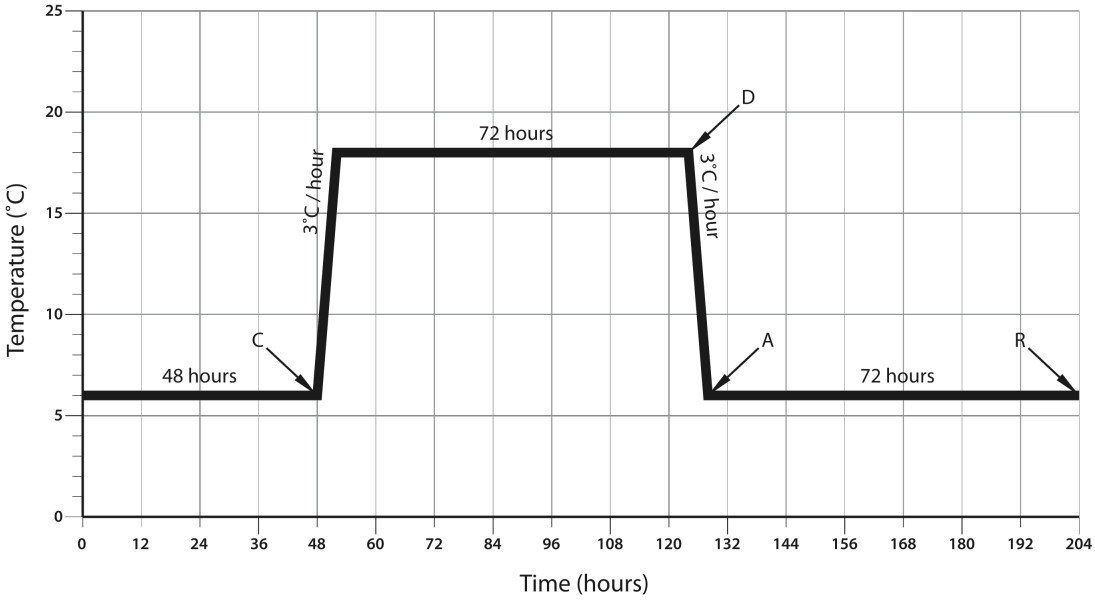

Figure 5-1 Schematic representation of temperature profile for chronic exposure to moderate heat stress.

Tanks were set up with a constant flow-through system (0.33 L/s) with fresh spring water at ambient temperature (approximately 6°C) and ambient oxygen levels (10.0–11.0 ppm). Approximately 250 fish were transferred to an experimental tank (diameter: 1.86 m, depth 50 cm) and left to acclimate for 48 h at ambient temperature (~6ºC). After acclimation, 10 fish were removed to act as a control (treatment group C) group, then water that had been diverted through a heat exchanger was added to the flow-through system to increase the water temperature in the tank by approximately 3.0ºC per hour for 4 hours, until the tank reached 18ºC, after which the water temperature was held at 15–19ºC and the fish were closely monitored for signs of stress. After 72 hours, 10 fish were removed for sampling (treatment group D) and the heat exchanger was turned off. The water was allowed to return to ambient temperature (~6ºC) overnight (12 hrs), and 10 more fish were sampled (treatment group A). After 72 hrs at ambient temperature, 10 additional fish were sampled (treatment group R).

The fish were euthanized by a swift blow to the head, then weighed and their fork lengths were measured. Blood was withdrawn from the caudal vein of the fish (maximum possible volume ~200 µL), the entire lower half of outer-most gill arch was removed, the entire liver was sampled, and an approximately 1 cm^2 section of muscle from above the lateral line and behind the dorsal fin of the fish was removed, in that order. Tissues were placed into RNAlater® (Ambion Inc.) and were stored at room temperature for 24 h to allow RNAlater® to penetrate the tissues, and then moved to -80°C for storage until use as per the manufacturer's instructions.

5.3.4 RNA extraction

RNA isolations and microarray analysis were conducted at the University of Victoria, Canada. Total RNA was isolated from gill, muscle and liver tissue samples. Briefly, tissue samples were removed from RNAlater®, blotted on a clean Kimwipe™ to remove excess solution, and disrupted and homogenized in 1 mL TRIzol reagent using a Mixer-mill (Retch® MM 301) with tungsten carbide beads. Phase separation was conducted using 200 µL chloroform, and RNA was purified using the RNeasy Mini Kit (Qiagen) following the manufacturer's instructions. Purified RNA was treated with 1 µL RNase inhibitor (Invitrogen). RNA integrity was verified by agarose gel with ethidium bromide staining to visualize ribosomal bands and by measuring the 260/280 absorbance ratio (>1.9) using a Nano Drop (ND-1000 Spectrophotometer, Thermo Scientifc) then stored at -80ºC until use.

5.3.5 Generation of cDNA and microarray hybridization

The microarray study followed a reference design format. cDNA was prepared from 1 μg gill RNA from six samples from each treatment group (C, D, A, R; 24 microarray slides in total) using Invitrogen's SuperScript Indirect cDNA labeling system. Treatment groups were compared indirectly against one another using a common reference sample that was hybridized to each microarray alongside the sample cDNA. The reference sample, designed to hybridize to as many spots on the array as possible, was comprised of high-quality RNA isolated from Atlantic salmon gonad, brain and spleen tissues that had been amplified using Ambion's Amino Allyl Message Amp^{TM} aRNA kit, then quantified, combined in equal amounts and divided into per-use aliquots to avoid degradation due to repeated freeze-thawing.

The GRASP 32K cDNA microarray was used (7). Details of the microarray hybridization process can be found at the University of Victoria cGRASP website (http://web.uvic.ca/grasp/microarray/array.html) within the .pdf document entitled *Invitrogen Indirect cDNA Labeling System version 3*. Briefly, microarray slides were post-print processed by rinsing in 0.2% SDS and water and dried by centrifugation, then pre-hybridized in 5 x SSC, 0.1% SDS, 3% BSA, washed with water, dried again and stored in a dry oven at 49ºC until cDNA hybridization. cDNA (300 ng) and aRNA (500 ng) were labeled with Cy5 and Cy3 (Amersham Biosciences), respectively, using Invitrogen's SuperScript Indirect cDNA Labeling System, then combined with 2 x Formamide buffer and LNA dT blocker (Genisphere) to a total volume of 60 μL, which was heated to 80ºC then loaded on to the slide in the dark. Microarrays were incubated for 16 h at 49ºC in a dark, humidified chamber, then underwent a series of washes and were dried by

centrifugation. Slides were scanned at 74 and 72 PMT for Cy3 and Cy5, respectively, using a ScanArray™ Express Microarray Scanner (Packard BioScience BioChip Technologies, model #ASCEX00) and spot intensity was calculated with ImaGene ver. 6.5.1.

5.3.6 Statistical analysis for identifying differentially expressed genes by microarray analysis

All statistical analyses of the microarray data were conducted using Genespring ver. 7.3.1. Spots were identified as per the fully annotated gene ID file from the GRASP website (http://web.uvic.ca/grasp/microarray/array.html) (IE007 onwards; last modified on Nov. 3 2008). Signals were normalized per spot and per chip using an intensity-dependent (LOWESS) normalization, then per gene to normalize to the median. Spots were filtered on flags present, and only spots with signals greater or equal to the average base/proportional value of the raw channel were retained.

Three pairwise student's t-tests were performed: C vs. D, C vs. A and C vs. R with a *p*-value <0.01 (note that this is more stringent than a Bonferroni-corrected p-value of 0.05 with three comparisons), and any genes not meeting a two-fold differential expression between pairs were filtered out in the same step. By comparing each of the D, A and R treatment groups against the Control group, and then compiling the resulting gene lists into a Venn diagram, we were able to compare each of the treatment groups against one another, and to determine which genes were differentially expressed in one, two or all three comparisons. Thus, we could sort the genes in the lists into those whose expression was elevated or down-regulated during the chronic heat exposure, those that remained

differentially expressed (or did not) after the water returned to ambient temperature, and those that were still differentially expressed (or not) after the fish had been at ambient temperature for 3 days. Therefore, the Venn diagram and the gene lists generated provide a very comprehensive and in-depth means of analyzing the gene expression data. Note that the gills were not perfused, and thus these results may reflect transcription taking place within the gill tissues and/or the blood trapped within the gills.

5.3.7 Expression analysis using qPCR

We prepared cDNA for qPCR from 1 µg of total RNA using Invitrogen's SuperScript III Reverse Transcriptase kit following the manufacturer's instructions. The six RNA samples per treatment group that were used for microarray analysis were used along with RNA from three additional fish per treatment group (i.e., nine individuals per treatment group were tested with qPCR). qPCR primer pairs were designed for 24 genes selected based on interest in function as well as the degree of fold change observed in the microarray analysis. qPCR primer pairs were designed from the Atlantic salmon EST sequences used on the GRASP 32K array using Primer 3 version 0.4.0 (http://frodo.wi.mit.edu/primer3/). Primers were tested for amplification efficiency using cDNA generated from a single gill RNA sample using the qPCR conditions described below followed by a dissociation curve analysis to test for a single product for each primer pair and that no primer dimers were generated during the 40 amplification cycles. The 12 primer pairs meeting these criteria and showing the highest efficiencies (range 76–97%; Table 5-1) were used for expression analysis of cDNAs from nine individuals from each treatment group. qPCR was conducted using the ABI 7900HT system with

Sybr green (Quanta Biosciences) under the following conditions: 95ºC for 3 min followed by 40 cycles of 95ºC for 15 s 60ºC for 30 s and 72ºC for 15 s, with one 96-well plate run per individual cDNA sample, which included, in triplicate, all 12 primers with corresponding no-template-controls (NTC), and two primer pairs for the endogenous control gene, $EF1A_A$. Specifically, we used the $EF1A_A$ primers designed by Olsvik et al. (9), which cross exons 5 and 6, and also designed primers that span exons 3 and 4 ($EF1A_A$3to4) (Table 5-1). Having two primer sets within one gene serves as a control for the quality of the cDNA reverse transcription reaction because the expression of $EF1A_A$ should be the same using both primer pairs. The amplification efficiencies of the endogenous control primer sets $EF1A_A$ and $EF1A_A$3to4 were 98.6% and 98.8%, respectively. Each plate also contained a no-reverse-transcriptase control (i.e., RNA from the individual being tested that had gone through the steps of the cDNA preparation but lacking reverse trancriptase) to test for genomic DNA contamination of the cDNA. Also included was a linker sample, a pooled mixture of cDNA that was aliquoted (to prevent repeated freeze-thaws) and amplified in triplicate on each plate with the $EF1A_A$ primers, which was compared across all plates to test for technical variations between plates and was used to calibrate the data across plates (average CT = 22.19 SD = 0.44).

Gene Name	GenBank	Primer Sequences (F and R)	Efficiency (%)
Heat shock protein HSP 90-beta	CA767842	F-GCG TTG CCC ACC ATT AAC	87.3
		R-AAT GGG TAA CCT GGT CAG TGT C	
Heat shock 70 kDa protein	EG865212	F-AAG ATC AGC GAG GAG GAC AA	81
		R-TGC CTG ATC TCC ACA GCA	
Cathepsin D precursor	CA043554	F-TCT TCC AAC TTG TGG GTT CC	76
		R-GTG CAT GTG TCT TGG CTG AG	
Cell cycle control protein 50A	CA064173	F-CCT TTT CCA TCA CTC CTC CA	80.5
		R-TTG GGG GTC AGA ACA ACT TC	
Serpin H1 precursor	CA063723	F-CTG GGA GGC AAA AAC AAC TG	97
		R-TTC CAC CAT TCT TTT CAC CAG	
60S ribosomal protein L23	DW561359	F-AGG GCC AAA CCT TTC ATT TC	94
		R-CTT CCT GGG GTT GTG ATA CG	
60S ribosomal protein L11	CA052515	F-ACG ACG TAG AAG TCC AGT CCA T	79
		R-CTC AAG GTG CGT GAG TAC GA	
Vascular endothelial growth factor C precursor	DW563176	F-ATT GAG TCA GAG TGG AAA AAG ACC	89
		R-TGC TGA TGT AGG AGG TGC TG	
Ubiquitin carboxyl-terminal hydrolase 8	DY733528	F-GAA ATG TTT GCT GGC AAC G	80
		R-CCA TGG AAC AGA GCT ACG ATG	
Heat shock protein 30	CB498291	F-CCA GAG GAG CTG TCT GTC AAG	93
		R-GGA GCC TGG ATC TGT AGC TG	
Heme oxygenase	CB515893	F-GGC TAC ACA GAT CCC CAG AA	78
		R-GAG GGA AAG TGA GCT CAT GC	
Heat shock 70 kDa protein cognate 4	CB485951	F-AGC ATG GCA AGG TTG AAA TC	90
		R-TGT CCG ATT GAA CAA CTC CA	
EF1AA (endogenous control)	N/A	F-CCCCTCCAGGACGTTTACAAA	98.6
		R-CACACGGCCCACAGGTACA	
EF1AA3to4 (endogenous control)	N/A	F-CCTGTGGAAGTTTGAGACTGG	98.8
		R-GAGTCTGCCCGTTCTTTGAG	

Table 5-1 Sequences and efficiencies of primers used for qPCR

5.3.8 Statistical analysis of qPCR results

qPCR results were analyzed using the ΔΔCt method (11) and calibrated for individual primer amplification efficiencies, producing a relative quantification (RQ) compared to the mean CT of the linker cDNA per plate. RQ values were tested for outliers using the Box-Whisker method such that any data points falling outside of the 95% range were eliminated from the analysis. Remaining RQ values were log transformed to meet the assumptions of the statistical tests. To test whether, and the extent to which, the qPCR and microarray results were in agreement, pairwise t-tests were conducted between the treament groups that were originally identified as showing significant differential expression by microarray analysis. We also performed an ANOVA among all four treatment groups (C, D, A and R) followed by a Tukey post-hoc test to determine whether the qPCR analysis revealed any results not seen by the microarray analysis.

5.4 Results and Discussion

5.4.1 Fish sizes

The average length (± standard deviation) of all of the sample fish (groups C, D, A and R) was 16.09 ± 2.27 cm, while the average mass was 32.05 ± 12.97 g, and there were no significant differences between the sizes (length or mass) of the fish between treatment groups (1-way ANOVA, p=0.65 and p=0.67 for length and mass, respectively). However,

the 10 fish that died during the trial were significantly smaller, both in length (13.75 ± 12.56 cm) and mass (13.75 ± 12.56 g), than the sampled fish (p=0.0041 and p=0.0155, respectively; pairwise t-test between all treatment groups combined and the fish that died during the trial). The sizes (weight and length) of the fish for each treatment group, including those that died during the trial, are shown in Fig. 5-2.

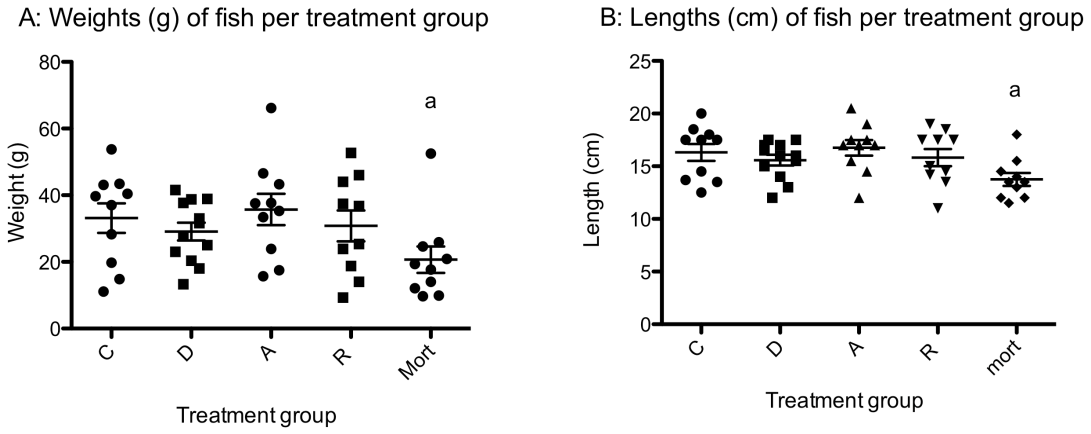

Figure 5-2 Comparison of fish weights and lengths across treatment groups.

The mean ± SD weights (g) for the fish in treatment groups C, D, A, R and the fish that died during the trial (morts) were: 33.14±14.0, 29.09±9.27, 30.83±14.6, 30.83±14.64 and 20.67±12.65, respectively. The mean ± SD lengths (cm) for the fish in treatment groups C, D, A, R and the fish that died during the trial (morts) were: 16.32±2.5, 15.58±1.78, 16.75±2.84, 15.82±2.4 and 13.75±1.93, respectively. a: The fish that died during the trail were significantly smaller than all of the fish that were randomly selected for sampling (1-way ANOVA comparing all C, D, A and R fish combined against morts; P=0.0155 and P=0.0041 for weight and length, respectively).

5.4.2 Genes identified as differentially expressed by microarray analysis

The resulting lists from the Venn diagram (Fig. 5-3) are provided in Supplemental Table S5-1. Note again that we used reverse-transcribed RNA isolated from non-perfused gill tissues for the expression analysis, and thus that any differential expression could reflect transcription occurring within the gill tissue itself and/or the blood trapped within the gills. Further analyses using gill-only and blood-only RNA are necessary to distinguish the location of transcription, which would provide further insight into the biological mechanisms taking place.

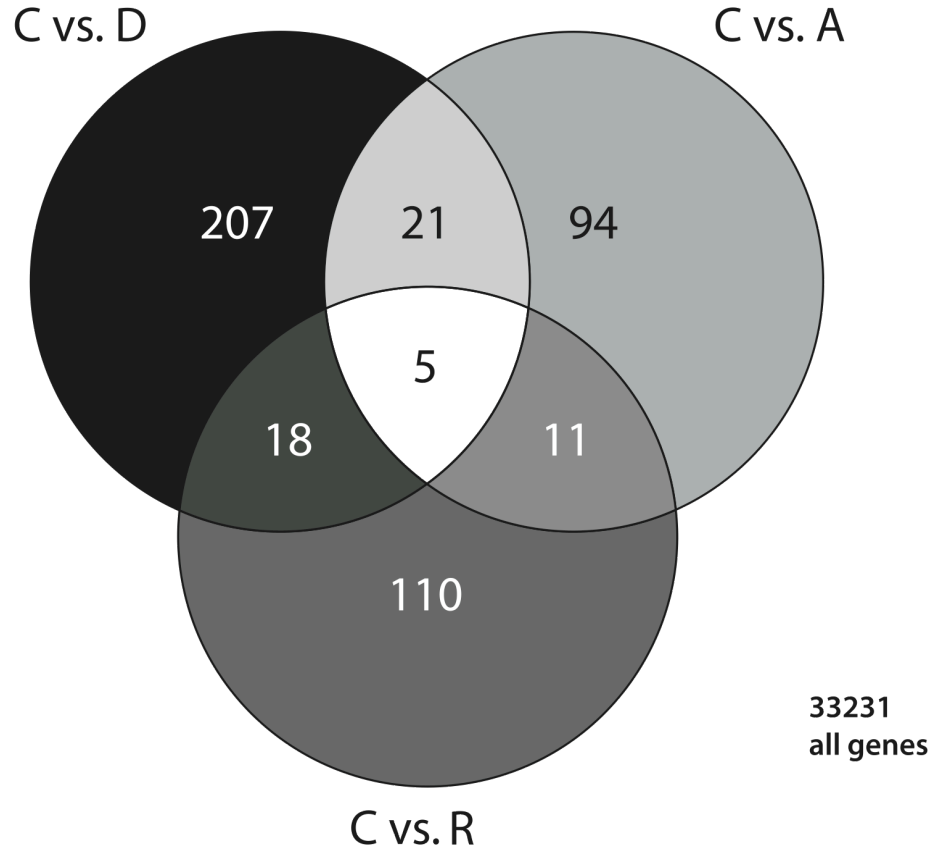

Figure 5-3 Venn diagram of three pair-wise comparisons of treatment groups

Venn diagram of three pair-wise comparisons of treatment groups (i.e., C vs. D, C vs. A, C vs. R; large circles). Numbers refer to the number of genes in that section. Overlapping regions provide lists of genes that were differentially expressed in more than one comparison. See Supplemental Table S5-1 for lists of genes within each section.

A total of 466 genes were identified as differentially expressed by the three pairwise comparisons (C vs. D, C vs. A and C vs. R) at $p<0.01$ with a two fold change threshold. All of the microarray data (normalized as well as raw data) were deposited within Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE29610. Specific genes or gene families of interest are discussed below. Note again that the gene lists contained within the non-overlapping portions of the Venn diagram (Fig. 5-3) include genes that were only differentially expressed at a particular time point (i.e., During, After or Recovery), whereas those contained within the overlapping portions of the Venn diagram were differentially expressed at more than one time point.

There were 251 genes differentially expressed between the C and D treatment groups, 207 of which were only indicated in this pair-wise comparison (i.e., the non-overlapping portion of the Venn diagram). Not surprisingly, the chronic (i.e., 72 hr, 15–18ºC) heat exposure resulted in the up-regulation of several Hsps of various families. These included five 70 kDa Hsps (2.14–14.2 fold), one 71 kDa Hsp (3.81 fold), five Hsp 90 beta genes (3.69–6.7 fold), Hsp30 (7.5 fold) and Hsp beta-11 (3.43 fold). Hsps are molecular chaperones that facilitate proper protein folding under a stress conditions (18). Ubiquitin, a small regulatory protein found in all cells that tags proteins for recycling (6) and is commonly observed to follow the same expression patterns of Hsps (12), was also up-regulated in the D treatment group (3.43 fold; Fig. 5-4). In addition, *78 kDa glucose-regulated protein precursor*, an Hsp70 homolog that is known to be a stress protein and has been reported to protect the cells by suppressing oxidative damage and stabilizing

calcium homoeostasis (2), was up-regulated in D vs. C fish (4.54 fold). We previously found this gene to be associated with tolerance to acute lethal heat exposure (16). Thus, *78 kDa glucose-regulated protein precursor* deserves recognition as a putative gene for identifying biomarkers for either genomics assisted selection or population assessment of Arctic charr under both acute and chronic temperature stress.

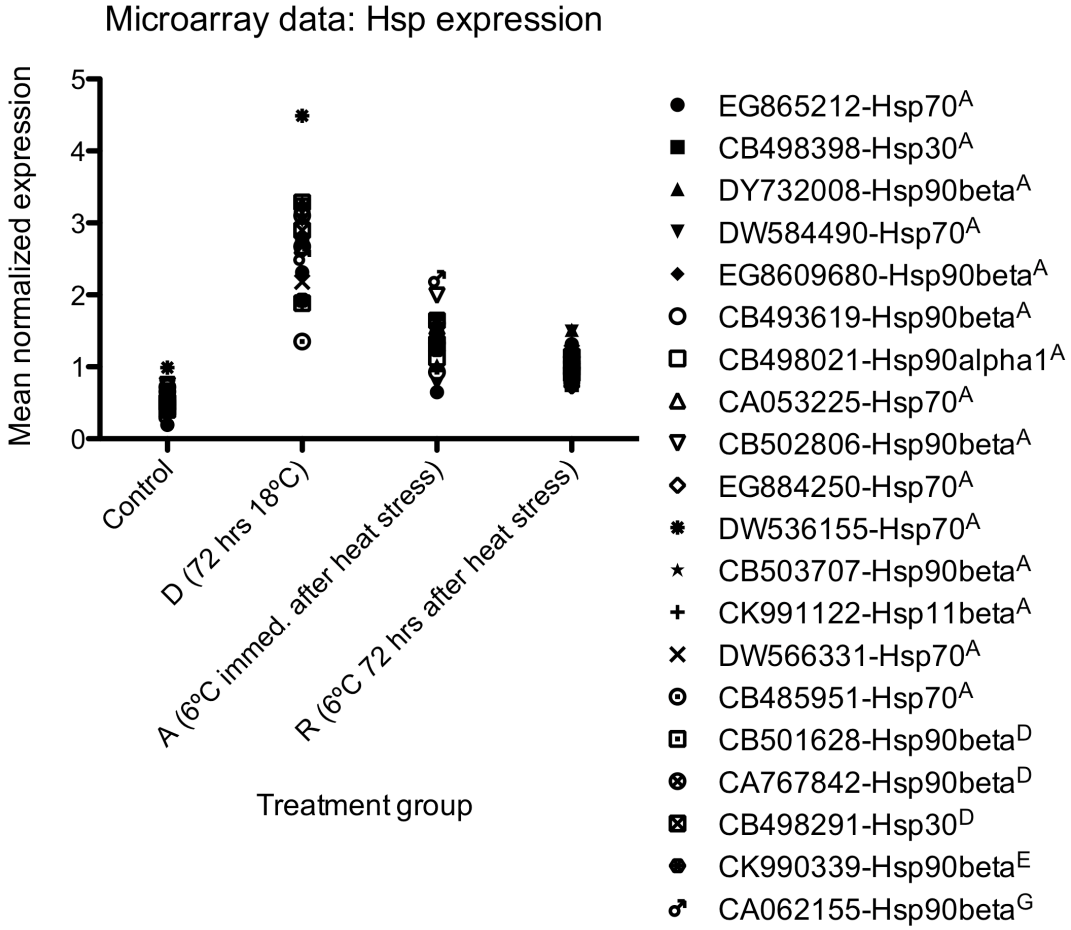

Figure 5-4 Microarray results for Hsps after prolonged exposure to moderate heat stress

Mean normalized expression (n=6 per treatment group) for all Heat shock proteins identified as differentially expressed by microarray analysis. Letters in superscript correspond to the gene list (i.e., pairwise comparison) within which each Hsp was identified as per Supplemental Table S5-1, which lists the fold change value for all genes within these lists. In addition, all expression data for each individual for each gene are available in the GEO database (GSE29610). *results tested and verified by qPCR

There were 130 genes differentially expressed between the C and A (after) treatment groups, 94 of which were only indicated in this pair-wise comparison. Notable genes in this list included eight ribosomal proteins (six 60S, one 40S and one 39S mitochondrial protein; 2–3.23 fold). A total of 144 genes were differentially expressed between the C vs. R groups, 110 of which were specific to this comparison. In total, there were 12 ribosomal proteins (seven 60S and five 40S) up-regulated in R compared to C fish (2.02–5.07 fold). These results are suggestive of increased ribosome biogenesis. It has long been recognized that ribosomes are key ROS (reactive oxygen species) targets, and that they may be more sensitive than DNA is to oxidative damage (3). More recently, ribosome activity has been recognized as a potential biomarker for cellular stress; for example, caged mussels exposed to heavy metal contamination exhibited degradation of ribosomal subunits (14). The expression patterns of all ribosomal genes identified by the microarray results are presented in Fig. 5-5. Although Fig. 5 shows that the elevated expression of ribosomal protein genes started during heat exposure, there were only three ribosomal protein genes that met the two-fold increased expression, $p<0.01$ cut-off in the D group, and none in the overlapping gene lists, whereas ribosomal protein genes accounted for approximately 10% of the genes in both the A and R lists. Thus, it appears that ribosome biogenesis was initiated late in the heat exposure, and continued long after the water temperature was returned to normal, perhaps suggesting that the gill cells were undergoing repair from oxidative damage induced by chronic heat stress.

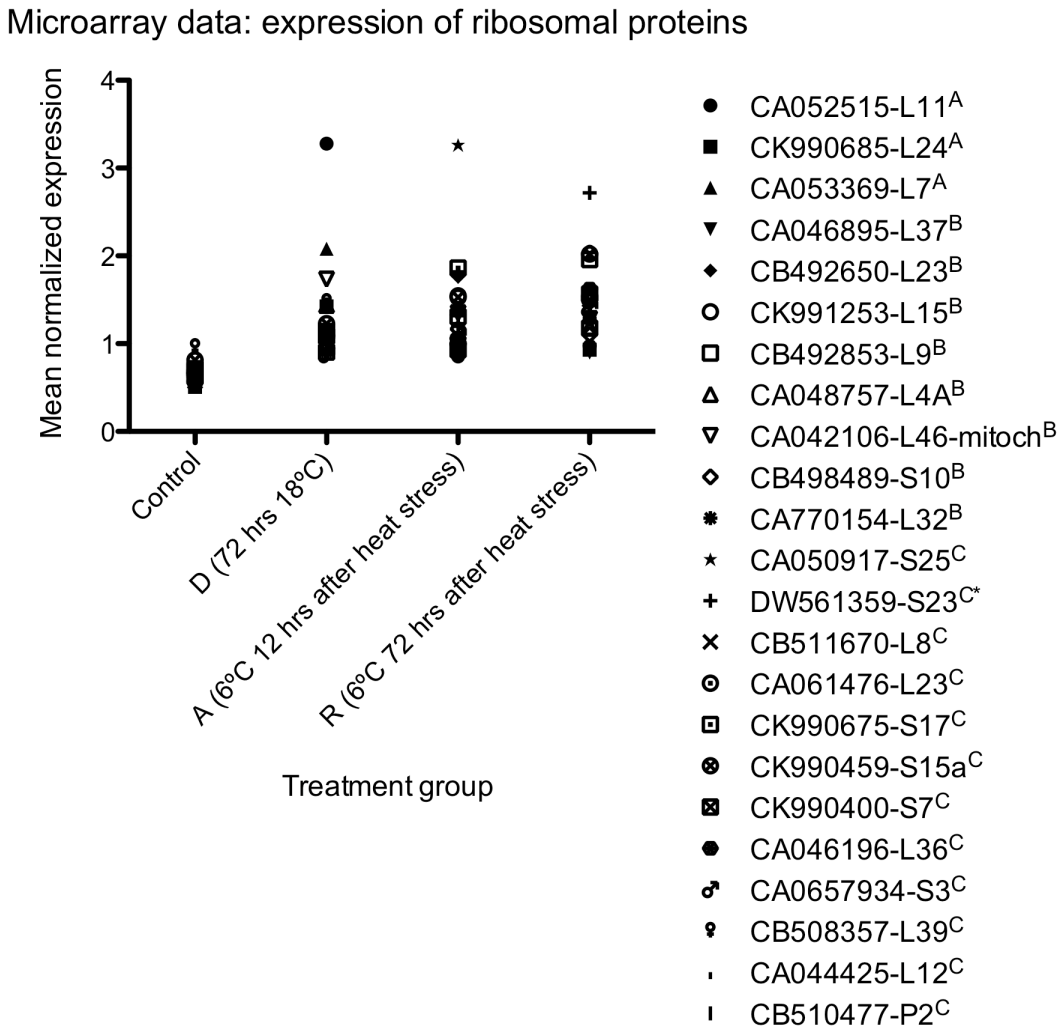

Figure 5-5 Microarray results for ribosomal proteins after prolonged exposure to moderate heat stress

Mean normalized expression (n=6 per treatment group) for all ribosomal proteins identified as differentially expressed by microarray analysis. Letters in superscript correspond to the gene list (i.e., pairwise comparison) within which each ribosomal protein was identified as per Supplemental Table S5-1, which lists the fold change value for all genes within these lists. In addition, all expression data for each individual for each gene are available in the GEO database (GSE29610).

Interestingly, we also found that 11 ribosomal proteins were significantly up-regulated in fish that exhibited tolerance to acute, lethal temperature stress in our previous study, as well as one ribosomal protein elevated in the Intolerant group relative to the controls (see Supplemental Table S4-2 of (16)). Two of the former (CA061476 and CB498489) as well as the latter (CK990675) ribosomal proteins were also indentified as differentially expressed in the present analysis. Thus, it appears that ribosome biogenesis plays a role in tolerance to acute thermal stress as well as the general response to chronic, sub-lethal thermal stress. Future studies of chronic heat exposure could incorporate a longer recovery period as well as finer time-course measurements to obtain a more precise timeline of ribosomal protein gene expression. In addition, examinations of various populations of Arctic charr, which are known to exhibit differences in their ability to withstand, and perhaps recover from, heat stress, should focus on the expression of these ribosomal protein genes as potential bioindicators for susceptibility to extreme temperature fluctuations in the natural environment.

Of the 21 genes that were differentially expressed between both the C vs. D and C vs. A (i.e., overlapping region of the Venn diagram), three were Hsps (two Hsp 90 beta genes and one Hsp 30). In addition, there was one Hsp 90 that was differentially expressed between both the C vs. D and C vs. R groups, and one that was identified in all three pairwise comparisons (middle section of Venn diagram). Interestingly, this Hsp 90 (GenBank accession number CA062155) was also up-regulated in both heat tolerant and intolerant Arctic charr in our previous study (16), with significantly higher expression in tolerant fish. This suggests that this is a particularly sensitive Hsp that is associated with

thermo-tolerance to extreme, acute heat exposure as well as with recovery from chronic, moderate heat exposure. Thus, we propose that this Hsp 90 warrants further attention with respect to its specific roles in heat response, and that it is a particularly strong candidate for incorporation into a genomics-assisted breeding program for Arctic charr or for use as a biomarker for wild populations.

There were no Hsps with elevated expression in the non-overlapping gene lists comparing A or R groups to Controls (vs. 13 in the C vs. D group). This suggests that elevated Hsp expression in the gill tissues of Arctic charr exposed to chronic, moderate temperature stress started to decrease quickly after the water temperature returned to normal. This is in accordance with the results reported by Lewis et al. (2010), in which elevated expression Hsps in rainbow trout blood samples was decreased 24 hours after heat stress (8). The results for all of the Hsps identified by the microarray analysis are shown in Fig. 5-4.

We also found a single hemoglobin beta gene down-regulated in the R group compared with the Controls (7.25 fold). In our recent report that identified genes associated with upper temperature tolerance in Arctic charr, we found that numerous hemoglobin transcripts, both alpha and beta (nine of each) were differentially expressed with highly similar expression patterns (16). We speculated that this reflected cross-hybridization among the hemoglobin transcripts on the cDNA microarray. In addition, we previously reported that the Atlantic salmon (the species from which the microarray was designed) hemoglobins are highly similar to one another (15). Thus, we suspect that the present result of a single differentially expressed beta hemoglobin transcript is a Type 1 error, as we would expect to see similar expression behavior among several hemoglobin

genes (alpha and beta) if hemoglobin was indeed playing a biological role in the heat stress recovery. However, this hypothesis remains to be tested by qPCR, which is not possible at this time given that the Arctic charr hemoglobin repertoire has not been sequenced, a worthy pursuit of future research.

5.4.3 Microarray validation by qPCR

Table 5-2 lists the genes tested by qPCR with their corresponding GenBank accession numbers from which the primers were designed, the microarray gene list that the genes were found on and the corresponding fold change. Also shown are the results of the qPCR analysis (pairwise t-test of original pairs) and those results (*p*-values) of the ANOVA. Results from the qPCR analysis that are in significant agreement with the microarray analysis are highlighted in dark grey, those showing a different trend from the microarray analysis are in light grey and those with no trend are not highlighted. Of the 12 genes selected from the microarray results that were tested with qPCR, six showed significant differential expression ($p<0.001$) in the same direction and between the same two treatment groups as the microarray analysis. One gene (*60S ribosomal protein L11*) showed no trend between the original pair of groups (C vs. D), but showed a slight trend of up-regulation in A vs. D, as determined by the ANOVA ($p=0.0446$). The remaining five genes produced very high p-values and showed no trend of differential expression between any treatments.

	Microarray Results			*qPCR Results (p-values)*	
Gene Name	**Gene list**	**Direction**	**Fold diff.**	**t-test original pair(s)**	**qPCR ANOVA**
Heat shock protein HSP 90-beta	C vs. D and C vs. A	D>A>R>C	C/D=0.282; C/A=0.447	C/D=0.0044; C/A=0.0094	0.0077
Heat shock 70 kDa protein	C vs. D	up in D	C/D=0.0704	< 0.0001	< 0.0001
Cathepsin D precursor	C vs. R	up in R	C/R=0.39	0.0063	0.0024
Cell cycle control protein 50A	C vs. D and C vs. R	D>R>A>C	C/D=0.227; C/R=0.286	C/D=0.0095 C/R=0.254*	0.1791
Serpin H1 precursor	C vs. A	up in A	C/A=0.098	< 0.0001	< 0.0001
60S ribosomal protein L23	C vs. R	up in R	C/R=0.227	0.0088	0.0816
60S ribosomal protein L11	C vs. D	up in D	C/D=0.236	0.6478	0.0446†
Vascular endothelial growth factor C precursor	C vs. D	up in C	C/D=2.943	0.4691	0.695
Ubiquitin carboxyl-terminal hydrolase 8	C vs. D and C vs. R	C>D>R>A	C/D=2.387; C/R=5.613	C/R=0.867; C/D=0.96	0.923
Heat shock protein 30	C vs. D and C vs. A	D>A>R>C	C/D=0.118; C/A=0.223	C/D=0.506; C/A=0.5065	0.2699
Heme oxygenase	C vs. D	up in D	C/D=0.189	0.202	0.3236
Heat shock 70 kDa protein cognate 4	C vs. D	up in D	C/D=0.467	0.3469	0.5227

Table 5-2 Summary of microarray vs. qPCR results

The genes tested by qPCR, the microarray gene list that the genes were found on and the corresponding fold change, the results of the qPCR analysis with the average RQ value for each gene tested per treatment group and results (*p* values) of pair-wise t-tests between groups. Results from the qPCR analysis that are in significant agreement with the microarray analysis are highlighted in dark grey (six genes), while those showing a different trend from the microarrays are in highlighted light grey (one gene), and those with no trend are not highlighted (five genes). *qPCR significant for this comparison only †trend of up-regulation in A vs. D

This level of agreement between microarray and qPCR results is in accordance with our previous results obtained using Arctic charr cDNA with the cGRASP microarray (16), and is also similar to the level of agreement reported by Lewis et al. (2010) (8) for rainbow trout cDNA (from blood RNA) and the 16K cGRASP microarray (19). We discuss possible reasons for this seemingly low level agreement, as well as some of the advantages and drawbacks of the cDNA microarray and qPCR approaches in our previous report (16). Briefly, these include the statistical pitfalls of using large microarrays as well as the potential for cross-hybridization and the added complication of using cDNA from a species other than that from which the microarray was designed. These issues are increasingly relevant when studying complex (i.e., highly repetitive), duplicated genomes such as those of the salmonids, for which there is currently no reference sequence (4). We thus suggest that microarray analysis be regarded as a valuable exploratory tool, and that data analysis and interpretation should focus on genes for which there is supporting data such as families of genes with several members showing similar responses (e.g., the Hsps and ribosomal proteins in this case), genes related through a common pathway or common functions, results that are confirmed by qPCR and/or other types of analyses such as transcriptome sequencing or QTL examinations, and those that are in agreement with other supporting studies.

Finally, it should be noted that the expressions of the genes indentified herein represents, at least in part, the phenotypic response to thermal stress, rather than the genotypic response. Thus, at least for the purposes of identifying genetic markers for marker-assisted selection, or for use as biomarkers in wild populations, further

experimentation and analyses are necessary to identify how this differential expression is achieved at the DNA level. That is, modifiers of gene expression (e.g., promoter regions or transcription factors) can be cis- or trans-acting, meaning that genetic markers associated with differential expression may not necessarily be co-localized with the gene in question. We cover this topic in detail our previous report (16), and refer the reader to that paper for further discussion of the issue and its implications. We do, however, stress that this present lack of knowledge of the relationship between the phenotypic and genotypic responses of these genes to heat stress does not negate their current utility as markers for thermal stress. That is, their differential expression in the gill tissues could be used to identify fish responses, which could feed into traditional breeding programs or for the identification of stressed populations in the wild, or for identification of heat-stress-associated quantitative trait loci (QTL). Indeed, the success of non-lethal gill sampling techniques has been demonstrated (17), suggesting that large quantities of fish can be screened for differential expression of target genes, thus presenting an alternative use of our results should the identification of genetic markers not be immediately feasible.

5.4.4 Conclusions and applications

In our previous study (16), we exposed Arctic charr to acute lethal temperature stress to identify genes that are potentially associated with upper temperature tolerance. The goal was to identify genetic markers for temperature tolerance to feed into an ongoing breeding program at the Icy Waters Inc. breeding facility in Whitehorse, Canada, to generate a more robust broodstock that can withstand fluctuations in temperature and thus be grown at inland tank farms in more diverse regions than present. However, given that

fish are not likely to experience acute exposure to lethal temperatures in a tank farm or natural setting, we also sought to examine transcriptomic responses under more realistic conditions of chronic exposure to moderate temperature stress, such as that exhibited during a summer heat wave in both uncontrolled tank farms as well as natural conditions. Thus, for this study, we conducted a chronic temperature trial, exposing Arctic charr to water temperatures of 15–19ºC (mimicking the maximum tank temperatures observed at Icy Waters Inc., a freshwater Arctic charr tank farm in Whitehorse, Canada) for 72 hours, and gill tissues extracted before, during (i.e., at 72 hrs), immediately after cooling and after 72 hours of recovery at ambient temperature (6ºC) were used for gene expression profiling by microarray analysis. The results revealed an expected pattern for Hsp gene expression, with the highest expression seen during heat exposure, with significantly reduced expression (approaching control levels) quickly thereafter. One Hsp 90 in particular (GenBank accession number CA062155) was differentially expressed in all three pairwise comparisons, and was also up-regulated in both heat tolerant and intolerant Arctic charr in our previous study (16), with significantly higher expression in heat tolerant fish. These results highlight this Hsp as playing a particularly important role in both thermotolerance and recovery from heat stress, and are thus deserving of further attention. We also found that the expression of several ribosomal proteins was significantly elevated immediately after cooling and 72 hrs after cooling, suggesting that the gill tissues were undergoing ribosomal biogenesis while recovering from damage caused by heat stress. Interestingly, several ribosomal proteins were identified as associated with tolerance to acute thermal stress in our previous experiment (see Supplemental Table S4-2 of (16)). We suggest that ribosomal proteins be examined

further as potential biomarkers of tolerance to as well as susceptibility to and/or recovery from heat stress for both wild and farmed populations. Of additional interest was that *78 kDa glucose-regulated protein precursor*, which we previously found to be associated with tolerance to acute lethal heat exposure (16), was up-regulated during exposure to warm temperatures; therefore, this is another target for future examinations of heat stress, tolerance and recovery. Finally, our previous study identified hemoglobin genes as being significantly elevated in fish that exhibited intolerance to acute lethal temperature stress, whereas only one such gene was identified in the present study, which we attribute to a Type I error (see above).

The results of this study, combined with those of our previous study on tolerance and intolerance to acute lethal temperatures have numerous implications for both cultured and wild Arctic charr and, perhaps, other closely-related salmonids such as Pacific salmon. In addition, by comparing results obtained from testing extreme laboratory conditions with a more moderate and realistic temperature regime, we have provided a comprehensive understanding of the genes that govern temperature responses in these fish. Specifically, we have identified several candidate genes that can used to screen for genetic markers for temperature tolerance, which can be integrated into broodstock development programs to develop temperature tolerant strains of Arctic charr, such that they can be grown in regions closer to where they are consumed, perhaps in facilities previously designed to grow other fresh water fish such as rainbow trout. In addition, it was recently shown that populations of Sockeye salmon in the Fraser River in British Columbia, Canada exhibit marked differences in the cardiorespiratory physiology, with fish from more challenging environments showing greater cardiac performance (5), thus suggesting that different

populations of salmonids show differences in their adaptations to environments that may be altered by climate change. The genes identified here as well as in our previous study (16) could be used to develop DNA biomarkers to screen for wild populations of fish (Arctic charr as well as perhaps other closely related salmonids) that are particularly susceptible or not to changes in temperature, as well as to track population migration, colonization and even evolution as natural habitats are altered by climate change. Therefore, future studies should aim to identify genetic markers associated with the genes of interest identified herein as well as screen natural and cultured stocks of Arctic charr for variability in these markers. This will provide the ability to predict responses to acute and chronic temperature stress and to select for optimal performance during temperature stress and under commercial production conditions.

5.5 Acknowledgements

The authors thank S Pavey, as well as the members of the Koop Laboratory, particularly K von Schalburg and B Sutherland, and members of the Davidson laboratory, particularly K Lubieniecki and K Johnstone, for technical assistance and help with data analysis. K Johnstone designed the EF1A_A3to4 primers.

5.6 Grants

This work was supported by funding for the Consortium for Genomic Research on All Salmonids Project (cGRASP) from Genome Canada and Genome BC and by a Strategic Grant from the Natural Sciences and Engineering Research Council of Canada.

5.7 Disclosures

CR McGowan is the Broodstock Development Manager for Icy Waters Inc. He accepts full responsibility for the conduct of the trial, has full access to all the data, and control over the decision to publish.

5.8 References

1. **Ayer NW and Tyedmers PH.** Assessing alternative aquaculture technologies: life cycle assessment of salmonid culture systems in Canada. *J.Cleaner Prod.* 17: 362-363-373, 2009.

2. **Baek HY, Lim JW, Kim H, Kim JM, Kim JS, Jung HC and Kim KH.** Oxidative-stress-related proteome changes in Helicobacter pylori-infected human gastric mucosa. *Biochem.J.* 379: Pt 2: 291-299, 2004.

3. **Burdon RH, Gill VM and Rice-Evans C.** Oxidative stress and heat shock protein induction in human cells. *Free Radic.Res.Commun.* 3: 1-5: 129-139, 1987.

4. **Davidson WS, Koop BF, Jones SJ, Iturra P, Vidal R, Maass A, Jonassen I, Lien S and Omholt SW.** Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biol.* 11: 9: 403, 2010.

5. **Eliason EJ, Clark TD, Hague MJ, Hanson LM, Gallagher ZS, Jeffries KM, Gale MK, Patterson DA, Hinch SG and Farrell AP.** Differences in thermal tolerance among sockeye salmon populations. *Science* 332: 6025: 109-112, 2011.

6. **Kettern N, Dreiseidler M, Tawo R and Hohfeld J.** Chaperone-assisted degradation: multiple paths to destruction. *Biol.Chem.* 391: 5: 481-489, 2010.

7. **Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmbhatt S, Rosner J, Rexroad CE,3rd, McGowan CR and Davidson WS.** A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* 9: 545, 2008.

8. **Lewis JM, Hori TS, Rise ML, Walsh PJ and Currie S.** Transcriptome responses to heat stress in the nucleated red blood cells of the rainbow trout (Oncorhynchus mykiss). *Physiol.Genomics* 42: 3: 361-373, 2010.

9. **Olsvik PA, Lie KK, Jordal AE, Nilsen TO and Hordvik I.** Evaluation of potential reference genes in real-time RT-PCR studies of Atlantic salmon. *BMC Mol.Biol.* 6: 21, 2005.

10. **Pankhurst NW and King HR.** Temperature and salmonid reproduction: implications for aquaculture. *J.Fish Biol.* 76: 1: 69-85, 2010.

11. **Pfaffl MW.** A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29: 9: e45, 2001.

12. **Pratt WB, Morishima Y, Peng HM and Osawa Y.** Proposal for a role of the Hsp90/Hsp70-based chaperone machinery in making triage decisions when proteins undergo oxidative and toxic damage. *Exp.Biol.Med.(Maywood)* 235: 3: 278-289, 2010.

13. **Prowse TD, Furgal C, Bonsal BR and Edwards TW.** Climatic conditions in northern Canada: past and future. *Ambio* 38: 5: 257-265, 2009.

14. **Pytharopoulou S, Sazakli E, Grintzalis K, Georgiou CD, Leotsinidis M and Kalpaxis DL.** Translational responses of Mytilus galloprovincialis to environmental pollution: integrating the responses to oxidative stress and other biomarker responses into a general stress index. *Aquat.Toxicol.* 89: 1: 18-27, 2008.

15. **Quinn NL, Boroevich KA, Lubieniecki KP, Chow W, Davidson EA, Phillips RB, Koop BF and Davidson WS.** Genomic organization and evolution of the Atlantic salmon hemoglobin repertoire. *BMC Genomics* 11: 539, 2010.

16. **Quinn NL, McGowan CR, Cooper GA, Koop BF and Davidson WS.** Identification of genes associated with heat tolerance in Arctic charr exposed to acute thermal stress. *Physiol.Genomics* 2011.

17. **Rees CB, McCormick SD and Li W.** A non-lethal method to estimate CYP1A expression in laboratory and wild Atlantic salmon (Salmo salar). *Comp.Biochem.Physiol.C.Toxicol.Pharmacol.* 141: 3: 217-224, 2005.

18. **Vabulas RM, Raychaudhuri S, Hayer-Hartl M and Hartl FU.** Protein folding in the cytoplasm and the heat shock response. *Cold Spring Harb Perspect.Biol.* 2: 12: a004390, 2010.

19. **von Schalburg KR, Rise ML, Cooper GA, Brown GD, Gibbs AR, Nelson CC, Davidson WS and Koop BF.** Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics* 6: 126, 2005.

5.9 Supplemental Data

Appendix File 8 *Supplemental Table S5-1*

Lists of significantly differentially expressed genes with corresponding fold changes and *p*-values generated by pair-wise comparisons among treatment groups (C vs. D, C vs. A, C vs. R) then further analyzed using a Venn diagram. A: C vs. D only; B: C vs. A only; C: C vs. R only; D: C vs. D and C vs. A (i.e., significantly differentially expressed genes from the segment of the Venn diagram that overlaps the C vs. D and C vs. A pair-wise comparisons, representing genes with prolonged differential expression starting during heat exposure); E: C vs. D and C vs. R (representing genes that were significantly differentially expressed at least 2-fold during heat exposure then again 72 hrs later, but not in between or immediately following heat exposure); F: C vs. A and C vs. R (representing genes differentially expressed immediately and 72 hours after heat exposure, but not during); G: all pair-wise comparisons (genes from the center section of the Venn diagram – i.e., those significantly differentially expressed throughout the trial except during the control period).

6: Genomic organization and evolution of the Atlantic salmon hemoglobin repertoire

Published in: *BMC Genomics* (2010) Vol. 11, No.1, 539 ISSN 1471-2164.

Author list: Nicole L. Quinn[1], Keith A. Boroevich[1], Krzysztof P. Lubieniecki[1], William Chow[1], Evelyn A. Davidson[1], Ruth B. Phillips[2], Ben F. Koop[3], William S. Davidson[1]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada.
[2]Department of Biological Sciences, Washington State University, Vancouver, WA, USA.
[3]Department of Biology, University of Victoria, Victoria, British Columbia, Canada.

Author contributions: NLQ contributed to the study design, identification and isolation of BACs, sequence assembly and hand-finishing, sequence annotation, data analysis and manuscript preparation. KAB and WC conducted the bioinformatics and contributed to sequence annotation and data analysis. KPL assisted with sequence assembly and data analysis. EAD performed the linkage analysis and RBP conducted the FISH analysis. BFK and WSD contributed to the study design, data analysis and manuscript preparation.

6.1 Abstract

6.1.0 Background

The genomes of salmonids are considered pseudo-tetraploid undergoing reversion to a stable diploid state. Given the genome duplication and extensive biological data available for salmonids, they are excellent model organisms for studying comparative genomics, evolutionary processes, fates of duplicated genes and the genetic and physiological processes associated with complex behavioral phenotypes. The evolution of the tetrapod hemoglobin genes is well studied; however, little is known about the genomic organization and evolution of teleost hemoglobin genes, particularly those of salmonids. The Atlantic salmon serves as a representative salmonid species for genomics studies. Given the well documented role of hemoglobin in adaptation to varied environmental conditions as well as its use as a model protein for evolutionary analyses, an understanding of the genomic structure and organization of the Atlantic salmon α and β hemoglobin genes is of great interest.

6.1.1 Results

We identified four bacterial artificial chromosomes (BACs) comprising two hemoglobin gene clusters spanning the entire α and β hemoglobin gene repertoire of the Atlantic salmon genome. Their chromosomal locations were established using fluorescence *in situ* hybridization (FISH) analysis and linkage mapping, demonstrating that the two clusters are located on separate chromosomes. The BACs were sequenced and assembled into scaffolds, which were annotated for putatively functional and pseudogenized

hemoglobin-like genes. This revealed that the tail-to-tail organization and alternating pattern of the α and β hemoglobin genes are well conserved in both clusters, as well as that the Atlantic salmon genome houses substantially more hemoglobin genes, including non-Bohr β globin genes, than the genomes of other teleosts that have been sequenced.

6.1.2 Conclusions

We suggest that the most parsimonious evolutionary path leading to the present organization of the Atlantic salmon hemoglobin genes involves the loss of a single hemoglobin gene cluster after the whole genome duplication (WGD) at the base of the teleost radiation but prior to the salmonid-specific WGD, which then produced the duplicated copies seen today. We also propose that the relatively high number of hemoglobin genes as well as the presence of non-Bohr β hemoglobin genes may be due to the dynamic life history of salmon and the diverse environmental conditions that the species encounters.

Data deposition: BACs S0155C07 and S0079J05 (fps135): GenBank GQ898924; BACs S0055H05 and S0014B03 (fps1046): GenBank GQ898925

6.2 Background

Hemoglobin, one of the most well-studied proteins to date, is responsible for oxygen transport from the lungs or gills to the tissues of vertebrates. The hemoglobin molecule is comprised of two α and two β subunits that non-covalently bond to form a tetramer [1,2].

The genes encoding the hemoglobin subunits are abundantly present and show relatively high similarity in structure (i.e., consisting of three exons and two introns) throughout the vertebrate lineage [3]. These characteristics, combined with the relative ease of isolating and studying hemoglobin proteins and their suspected role in adaptation to variable environmental conditions, have made the hemoglobin genes major targets for evolutionary studies [1,2,3,4].

Examinations of the genomic organization of chromosomal hemoglobin gene regions suggest that all hemoglobin genes evolved from a single monomeric form when gnathostome fish evolved from the more primitive agnathan fish approximately 500–700 million years ago [5,6]. The entire hemoglobin gene region then appears to have undergone a series of tandem duplications and divergence, giving rise to the modern α and β hemoglobin genes. Initially the α and β genes were adjacent on the same chromosome, and expansion of this region, including lineage-specific gene gain and loss, produced the multiple copies of α and β genes seen in Gnathostomata today [4,5,7,8].

The current genomic organization seen in mammals and birds is such that α and β hemoglobin gene clusters are located on different chromosomes and transcribed from the same strand in order of temporal expression [8,9]. The most parsimonious explanation of this arrangement involves a disruption in the α-β linkage by translocation of part of the hemoglobin gene cluster and subsequent gene silencing of α and β hemoglobins on respective chromosomes prior to the lineage leading to birds and mammals approximately 300–350 million years ago [9,10]. Studies of the genomic organization of hemoglobin genes in the mammalian and avian lines examined this hypothesis by looking for evolutionary “footprints” of silenced hemoglobin genes as well as conservation and

divergence patterns of genes surrounding the α and β gene clusters along the mammalian line [9,11].

The disruption in the α and β hemoglobin gene linkage in mammals and birds appears to have occurred after their divergence from the poikilothermic jawed vertebrate taxa. Rather, with the exception of some extreme cold-adapted Antarctic icefish that retain only remnants of α hemoglobin genes and have completely lost the β hemoglobin genes, and thus do not express hemoglobin [12,13], the fish and amphibians studied to date exhibit intermixed α and β hemoglobin genes on the same chromosome. For example, within the amphibian line, the genomes of both *Xenopus laevis* and *X. tropicalis* exhibit linked α and β hemoglobin genes [10].

The teleosts, or the ray-finned fish, are a diverse group that comprises most living species of fish, including more than 20,000 extant species covering more than 40 orders [14]. Despite significant differences in the number of hemoglobin genes and within-chromosome arrangements, model teleosts whose genomes have been studied to date, including the Japanese pufferfish (*Fugu rubripes*) [15], the zebrafish (*Danio rerio*) [16] and medaka (*Orzias latipes*) [17] are reported to exhibit two hemoglobin gene clusters located on distinct chromosomes. These observations support the hypothesis that the teleost lineage experienced a whole genome duplication (WGD) event subsequent to the divergence from tetrapods [18].

The Salmonidae, a family of teleosts that includes the salmon, trout, charr, grayling and whitefish are of considerable environmental, economic and social importance. Indeed, more is known about the biology of salmonids than any other fish group [19]. The common ancestor of salmonids underwent a WGD event between 20 and

120 million years ago [20, 21]. Thus, the extant salmonid species are considered pseudo-tetraploids whose genomes are in the process of reverting to a stable diploid state. The Atlantic salmon (*Salmo salar*) has been chosen as a representative salmonid for genomics studies, and an international collaboration to sequence the Atlantic salmon genome has been established [22].

Wolff and Gannon [23] provided the first sequence of an Atlantic salmon α hemoglobin from a kidney cDNA library. Subsequently, reports of the organization of the Atlantic salmon hemoglobin gene cluster described six lambda phage genomic clones comprising two sets of α and β hemoglobin genes oriented 3' to 3' on opposite strands [24,25,26]. This was the first evidence of this type of hemoglobin gene arrangement for any vertebrate species. The six clones comprised four unique α hemoglobin gene sequences and six unique β hemoglobin genes, including a β hemoglobin containing the characteristic amino acid changes that eliminate the Bohr effect, as well as one partial β hemoglobin gene (GenBank accession numbers X97284–X97289) [26]. It remained unknown, however, whether these represented all hemoglobin-like genes within the Atlantic salmon genome. In addition, it was not known whether the clusters were on separate chromosomes, or whether, as would be predicted by the salmonid-specific 4R WGD hypothesis [27], there were actually four hemoglobin gene clusters in salmon. Furthermore, the relative locations and orders of the clones to one another were not established, and the sequences of intergenic regions as well as the genes surrounding the hemoglobins were not determined. Finally, these investigations were not able to identify putative pseudogenes, incomplete hemoglobin genes or footprints of historical hemoglobin genes within the hemoglobin gene clusters or elsewhere in the genome.

Thus, a full characterization of the Atlantic salmon hemoglobin gene repertoire is needed to provide insight to the evolution of the organization and function of the Atlantic salmon hemoglobins, particularly in light of the teleost and salmonid-specific WGD events.

We used oligonucleotide probes specific for Atlantic salmon α, β and non-Bohr β hemoglobin genes as well as probes designed from rainbow trout (*Oncorhynchus mykiss*) embryonic hemoglobin cDNAs [28] to locate these genes within the Atlantic salmon bacterial artificial chromosome (BAC) library, CHORI-214 [29]. Four BACs, representing two genomic locations on different chromosomes and comprising the entire Altantic salmon hemoglobin gene repertoire were sequenced and annotated. Fluorescence *in situ* hybridization (FISH) and linkage analyses were performed to assign these BACs to chromosomal locations within the Atlantic salmon genome. Here we present the first description of an entire salmonid α and β hemoglobin gene repertoire. We also discuss our results in terms of the fate of the hemoglobin genes during and after the salmonid WGD event, and how this fits into the evolution of the hemoglobin gene family in teleosts.

6.3 Results

6.3.0 Identification and tiling paths of Atlantic salmon hemoglobin-containing BACs

All ^{32}P-labelled 40-mer probes for α, β, non-Bohr β and embryonic hemoglobins (probe and primer sequences are provided in Additional file 1, Table S1) hybridized to Atlantic salmon BACs belonging to two fingerprint scaffolds (fps), within the Atlantic salmon

physical map [30, 31]. Fps1046 contains 21 BACs and spans an estimated 458.8 kb; fps135 is comprised of 391 BACs spanning approximately 3.473 Mb. PCR was used to confirm the hybridization results and narrow down the regions within the fps that contained hemoglobin genes by screening all BACs surrounding the hybridization-positive BACs for the presence of hemoglobin genes. PCR primers were designed for sequence tag sites (STS) within the BAC-end sequences (SP6 and T7 ends) of suspected overlapping BACs spanning the hemoglobin gene region, and overlaps were checked by PCR amplification of the STS within the putative overlapping BACs. The overlapping BACs S0014B03 and S0055H05 were determined to span the hemoglobin gene region of fps1046, while S0155C07 and S0079J05 spanned that of fps135, thus creating BAC tiling paths for the hemoglobin regions of these fps. Individual shotgun libraries were generated for all four BACs and sequenced.

6.3.1 Sequence assemblies and annotation

The CHORI-214 BAC library was made from a diploid male Atlantic salmon individual, meaning that BAC inserts originated from either maternal or paternal chromosomes and therefore, overlapping BACs could exhibit allelic differences. This appeared to be the case for BACs S0014B03 and S0055H05, for which the overlapping region covered the hemoglobin genes within fps1046. Thus, although this overlapping section assembled into one contiguous section, reads containing allelic differences assembled into independent contigs that aligned to homologous regions along the solid contig. Nevertheless, the full BACs assembled very well, and only three contigs >1000 bp and two gaps remained after hand finishing. These gaps presumably span repetitive regions in

the Atlantic salmon genome [32]. Furthermore, given that the entire hemoglobin gene region was assembled with no gaps, it can be assumed that the six putatively functional α and six putatively functional β hemoglobin genes, two of which were defined as non-Bohr β hemoglobins, along with the two putative α hemoglobin pseudogenes and three putative β hemoglobin pseudogenes that were annotated within fps1046 represent all hemoglobin genes within that cluster (note that the solid contig was used for sequence annotation; Figure 6-1A). The total size of the assembly for the two BACs, not including allelic contigs (i.e., the non-redundant sequence), was 242,883 bp, with approximately 49,000 bp of overlap between them and the hemoglobin genes spanning approximately 87,000 bp.

Sequence reads from the overlapping region between the BACs S0155C07 and S0079J05 of fps135 assembled into one sequence contig with no apparent allelic differences. However, the remainder of the fps135 BACs proved much more difficult to assemble, and unfortunately, the repetitive nature of the sequences made it impossible to further improve the assembly by sequencing PCR products to fill gaps because we were unable to design specific PCR primers that would amplify a single product (i.e., numerous bands, or smears on agarose gels were obtained). Thus, a total of 23 sequence contigs >1000 bp remained after hand-finishing of the assembly. The relative orders of some of the sequence contigs of fps135 were determined by matching paired-ends of sequence shotgun clones. Six sequence contigs contained hemoglobin genes, and the relative order of these fps135 contigs with respect to one another was estimated by aligning the contigs against the completed assembly of the two BACs from fps1046. This was based on the assumption, given the highly similar nature of the non-coding regions,

as well as that of the genes flanking the globins, that there has not been a major disruption in the form of an inversion to either of the hemoglobin regions (i.e., that of fps135 or fps1046). Our sequence annotation identified seven putatively functional α hemoglobin genes and one putative α hemoglobin pseudogene, as well as eight putatively functional β hemoglobin genes, four of which were defined as non-Bohr β hemoglobins, and three putative β hemoglobin pseudogenes within the fps135 hemoglobin BACs (Figure 6-1B). This, however, must be considered a minimum estimate of hemoglobin genes within this region given the possibility that gaps between sequence contigs could contain additional hemoglobin genes. The total size of the assembled sequence contigs for the two BACs was 421,907 bp, with approximately 33,000 bp of overlap. The hemoglobin genes spanned approximately 130,000 bp not including gaps between contigs.

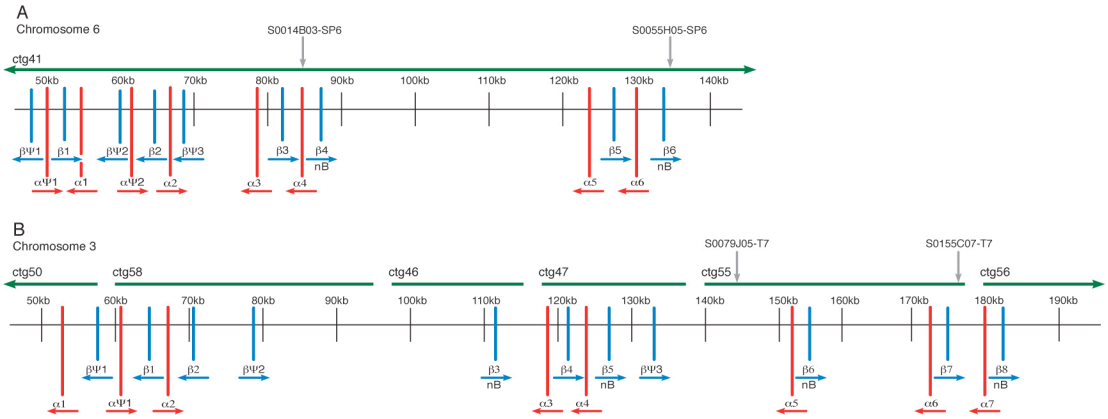

Figure 6-1 Genomic organization of the Atlantic salmon hemoglobin gene clusters.

A) Schematic representation of the region of Atlantic salmon chromosome 6 containing the hemoglobin genes. Sequence reads for this region assembled into one solid sequence contig (ctg 41). B) Schematic representation of the region of Atlantic salmon chromosome 3 containing the hemoglobin genes. Sequence contigs are indicated by horizontal green lines. β hemoglobin genes are indicated in blue; α hemoglobin genes are indicated in red. Arrows indicate strand of transcription. All hemoglobin gene names begin with SsaChr6 or SsaChr3 for chromosome 6 and chromosome 3, respectively, followed by α or β and a number indicating the order of the genes. SP6 and T7 ends of overlapping BACs are indicated by grey arrows. Thus, the regions between the arrows indicate BAC overlapping regions. Nb: Non-Bohr β hemoglobins.

All sequences were deposited in the NCBI GenBank database with the assembled sequence contigs for BACs S0155C07 and S0079J05 (fps135) under the accession number GQ898924 and those for BACs S0055H05 and S0014B03 (fps1046) under the accession number GQ898925.

All previously published Atlantic salmon hemoglobin sequences [25] were identified within the annotated hemoglobin clusters; however, there were two examples of possible allelic differences. Specifically, Clone 3 α hemoglobin (GenBank accession number X97286.1) exhibited 99% similarity at the nucleotide level to SsaChr6α6, which resulted in one amino acid change from a methionine to a leucine at amino acid 32, and the Clone 5 and Clone 6 β hemoglobins (Bohr; X97288 and X97289) showed 99% similarity at the nucleotide level with SsaChr6ß3, which resulted in one amino acid change from valine to leucine at amino acid 143. However, as both of these changes were caused by single nucleotide substitutions, each resulting in a single amino acid change, and given the depth of sequencing coverage of the Atlantic salmon BACs (i.e., >18x coverage in both cases), it is more probable that they reflect sequencing errors in the published clones rather than allelic differences.

To provide further evidence that the two Atlantic salmon hemoglobin gene clusters encompassed all previously identified Atlantic salmon α and β hemoglobin genes, we compared all identified putatively functional Atlantic salmon hemoglobin genes against all full-length Atlantic salmon cDNA clones [33]. Indeed, all unique full-length cDNA clones annotated as β hemoglobins were accounted for within the identified

putatively functional β hemoglobin genes. This was also true for the α hemoglobins, with the exception of the cDNA clones with accession numbers BT046755.1 and BT046550.1, which are highly similar to one another but not to the identified hemoglobins. An alignment of these clones using BLASTn [34] against the nr/nt database revealed similarity to hemoglobin subunit α-D, a distinct type of hemoglobin present in birds, mammals and reptiles that is predicted to have arisen via duplication from a gene that had larval/embryonic function [35]. This gene is apparently found in Atlantic salmon given the presence of the ESTs, but is not in the regions of the α and β hemoglobin genes.

Additional file 2, Table S6-2 lists all annotated α hemoglobin genes (Table S6-2A) and β hemoglobin genes (Table S6-2B) with the source chromosome, strand of transcription, start location, whether the entire hemoglobin gene matches one of the Atlantic salmon hemoglobin clones published by McMorrow et al. [25] at the amino acid level. It also identifies which genes are non-Bohr β hemoglobins, and lists the top hemoglobin EST cluster hit, if any, with the percent identity from the salmonid EST database [33,36] and whether the hemoglobin matches one of the full-length cDNA clones with the corresponding NCBI accession number. Table S6-2C (Additional file 2) lists all putative pseudogenes with the chromosome name, strand of transcription, start location and a description of each exon, with explanations of why the gene was classified as a pseudogene.

The identified putatively functional α hemoglobin genes SsaChr6α2 SsaChr3α2 as well as the β genes SsaChr6ß1, SsaChr6ß2, SsaChr6ß3, SsaChr6ß6, SsaChr3ß1, SsaChr3ß2 and SsaChr3ß8 did not have matching EST clones. This could mean that these represent newly identified hemoglobin genes, or that these genes are rarely or never

transcribed and thus are not represented in the Atlantic salmon cDNA libraries. Interestingly, several of these genes lie in regions where the tail-to-tail alternating order of the hemoglobin genes is disrupted and they would be transcribed on opposite strands than expected (see below for more details). Future studies using expression profiling of the Atlantic salmon transcriptome at various time points throughout the species' life cycle will provide further insight to this.

6.3.2 Conservation of gene order and strand of transcription

Wagner et al. [24] first reported the tail-to-tail orientation and alternating order of the Atlantic salmon α and β hemoglobin genes. We found that this orientation was fairly well conserved, with the α hemoglobins transcribed on the negative strand and β hemoglobins transcribed on the positive strand, and the alternating α-β order was mostly maintained in both chromosomes with some notable exceptions. On chromosome 6 (fps1046), the alternating α-β pattern is conserved throughout (including putative pseudogenes), but there are some apparent disruptions to the strand of transcription at the 5' end of the cluster. Specifically, SsaChr6ß2 as well as all putative β hemoglobin pseudogenes, including SsaChr6ßψ1, SsaChr6ßψ2, SsaChr6ßψ3, were predicted as being transcribed from the negative strand, whereas all putative α hemoglobin pseudogenes (SsaChr6αψ1, SsaChr6αψ2) as well as SsaChr6α2 would be transcribed on the positive strand (Figure 6-1A). On chromosome 3 (fps135), the β hemoglobin pseudogene SsaChr3ßψ1 as well as the putatively functional genes SsaChr3ß1 and SsaChr3ß2 were predicted to be transcribed from the negative strand, and SsaChr3ßψ2 and SsaChr3ßψ3 disrupt the otherwise conserved alternating α-β order and orientation of the genes, SsaChr3ß2 and

SsaChr3ß3 being adjacent (Figure 6-1B). In terms of α hemoglobin genes on chromosome 3, SsaChr3αψ1 and SsaChr3α2 were predicted as transcribed in the positive direction, whereas all others are transcribed on the negative strand. Note again that, for chromosome 3, the order and orientation of the sequence contigs was predicted based on homology with that of chromosome 6 (see dot plot in Additional file 3, Figure S6-1). Thus, it is possible that inversions or rearrangements may have taken place, and that the resulting predicted order and orientation of the hemoglobin genes is incorrect.

As indicated above, it is interesting to note that all of the putatively functional α hemoglobin genes predicted to be transcribed from the positive strand and all β hemoglobin genes predicted to be transcribed from the negative strand are lacking a corresponding EST at this time. It is possible that the apparent rearrangements have contributed to a global shutdown of transcription in these regions of the genome, allowing several of the hemoglobin genes to degenerate into obvious pseudogenes and silencing the remainder. This should be further explored using expression profiling by qPCR across all life stages of Atlantic salmon.

6.3.3 Linkage analysis and karyotyping

Microsatellite marker Ssa10067BSFU, representing fps1046 was informative in both the Altantic salmon SALMAP families (Br5 and Br6) [37, 38] and was mapped to linkage group 4. Microsatellite Ssa0516BSFU was informative in the Br6 family and mapped to linkage group 11 (Figure 6-2). FISH analysis revealed that fps1046 is found within Atlantic chromosome 6 and fps135 is within chromosome 3 (see [38] for chromosome nomenclature). The FISH and linkage mapping of chromosomes 6 and 3 to linkage

groups 4 and 11, respectively, contributed to the integration of the Atlantic salmon karyotype and linkage map [38]. Primer sequences for the microsatellite markers used for linkage analysis are provided in Additional File 6-1, Table S6-1 and within ASalbase, the Atlantic salmon genomic database [31].

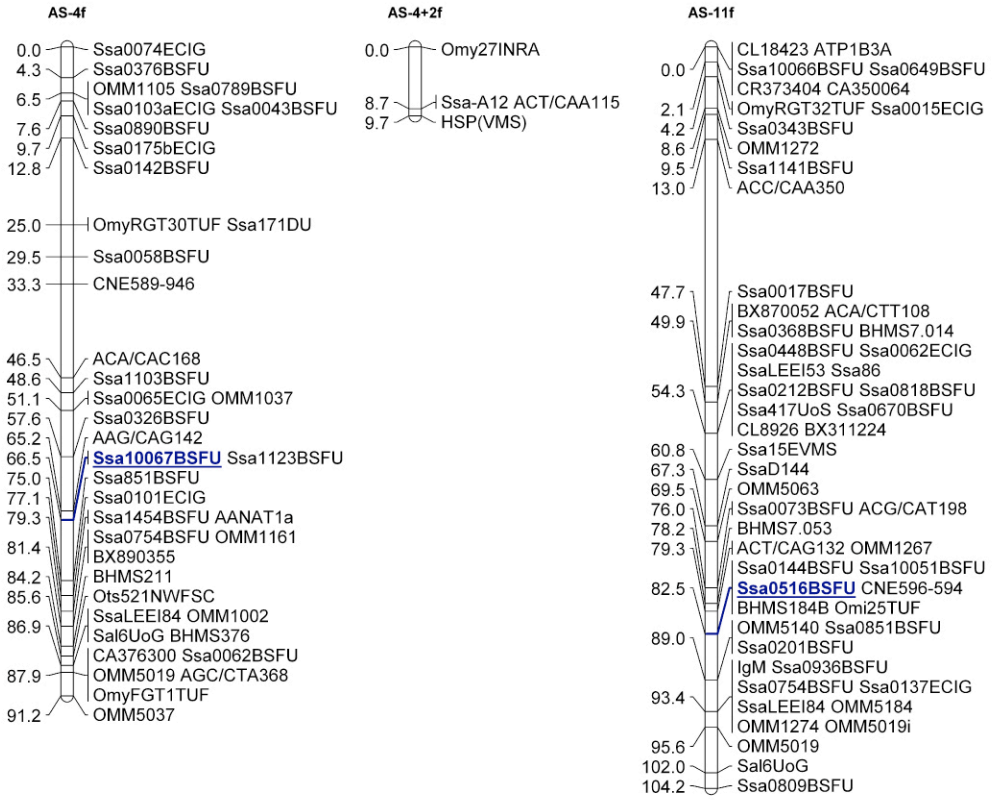

Figure 6-2 Merged female linkage maps for Atlantic salmon SALMAP families Br5 and Br6 showing linkage groups 4 and 11.

Microsatellite marker Ssa10067BSFU (underlined), representing fps1046 was informative in both the Altantic salmon SALMAP families (Br5 and Br6) and mapped to linkage group 4. Microsatellite Ssa0516BSFU (underlined) was informative in the Br6 family and mapped to linkage group 11.

6.3.4 Comparative genomic analysis of hemoglobin gene regions in other teleosts

We examined the regions surrounding the hemoglobin gene clusters in the available teleost genomes and compared them against one another as well as to those predicted to surround the Atlantic salmon hemoglobin gene clusters to gain insight to the nature of the teleostean hemoglobin gene containing chromosomes. We found that the genes surrounding the hemoglobin gene clusters are well conserved. Figure 6-3 shows a schematic diagram of the hemoglobin gene regions and the predicted surrounding named genes in medaka, zebrafish, stickleback and tetraodon compared to those of Atlantic salmon. Note that, for the hemoglobin gene containing BACs within Atlantic salmon chromosome 3, only the order and orientation of the sequence contigs that aligned with those from the BACs within chromosome 6 could be predicted. That is, given the extensive overlap between the two BACs that cover fps1046 (chromosome 6), the total sequenced region is much shorter than that of fps135 (chromosome 3); therefore, any sequence contigs from fps135 that did not fall within the coverage of fps1046 could not be ordered or oriented. Thus, for any sequence contigs that fell outside of this region, we were only able to establish their relative location compared to those that aligned with chromosome 3 based on their source BAC. Within Figure 6-3, solid lines between predicted genes indicate that the order and orientation of the predicted gene relative to those neighboring it is known, whereas a single black dot between predicted genes indicates that the relative location of the predicted genes compared to those joined by

solid lines is known, but their order and orientation (i.e., that of the sequence contigs on which they reside) relative to one another is not. Arrows in Figure 6-3 indicate the direction of transcription of the gene relative to the location of the hemoglobin gene cluster; lack of an arrow indicates that the relative direction of transcription cannot be determined.

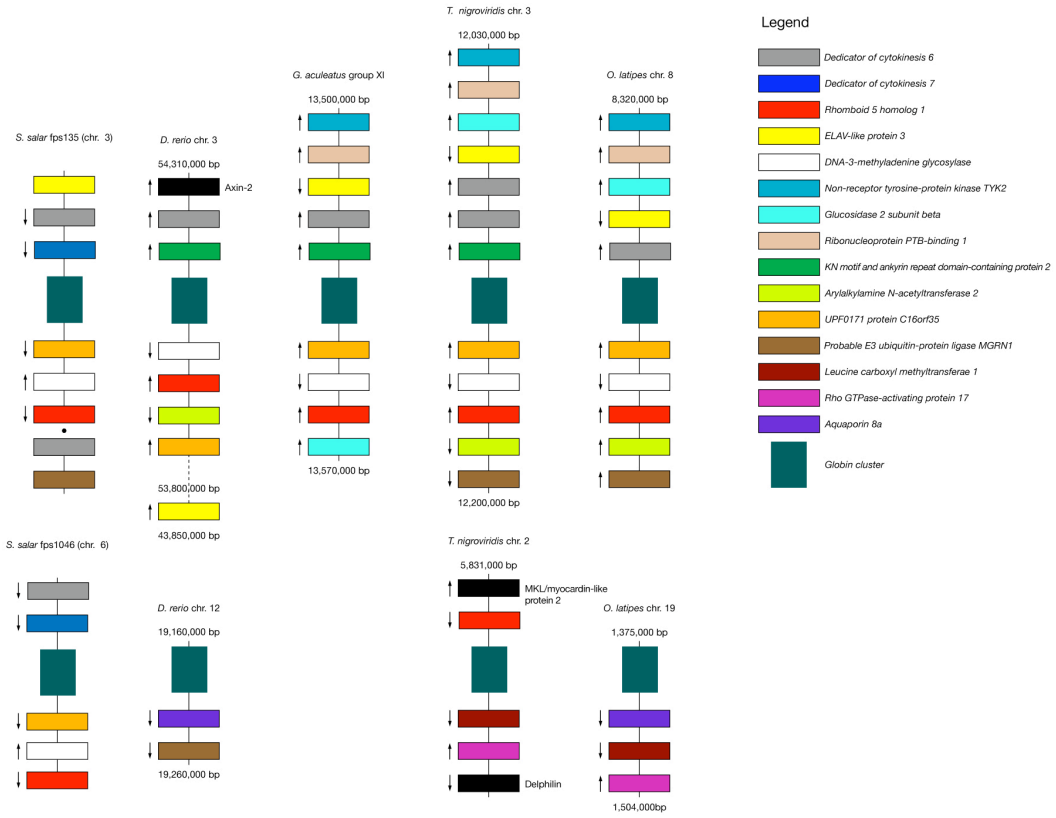

Figure 6-3 Comparative synteny of hemoglobin gene clusters among sequenced teleost species.

Schematic representation of annotated genes within the regions surrounding the hemoglobin gene clusters for Atlantic salmon and four annotated teleost genomes (*O. latipes, D. rerio, G. aculeatus, T. nigroviridis*). Colored blocks indicate shared or common genes as specified in the Figure legend. Black blocks indicate genes that are not shared within the indicated regions of any other species. Distances between genes vary (i.e., figure is not to scale); the start and end of the chromsome/group region is shown in base pairs (bp) for each of the annotated teleost genomes. Solid lines between predicted genes indicate that the order and orientation of the predicted gene relative to those neighboring it is known, whereas for Atlantic salmon fps 135 (chromosome 3), a single black dot between predicted genes indicates that the relative location of the predicted genes compared to those joined by solid lines is known, but their order and orientation (i.e., that of the sequence contigs on which they reside) relative to one another is not. Arrows indicate the direction of transcription of the gene relative to the location of the hemoglobin gene cluster; lack of an arrow indicates that the relative direction of transcription cannot be determined. For *D. rerio* chromsome 3, the gene for *ELAV-like protein* was found distantly downstream of the nearest common gene (*Arylakylamine N-acetyltransferase 2*), as indicated by the distance shown, with numerous predicted genes in between.

Briefly, medaka, zebrafish and tetraodon and Atlantic salmon exhibit two distinct hemoglobin gene clusters on separate chromosomes or linkage groups, whereas stickleback has only one. Although there are some rearrangements in terms of the positioning of genes relative to the hemoglobin genes and direction of transcription as well as some apparent gains, losses and duplications of genes, all of the organisms possess one similar cluster (hereafter Cluster 1) that contains, among others, the shared genes *UPF0171 protein C16orf35, Rhomboid family member 1*, *Dedicator of cytokinesis protein 6*, *ELAV-like protein 3* and *DNA-3-methyladenine glycosylase* (*MPG*; see Figure 6-3). Note these results are consistent with those of Patel et al. [11], who report that *MPG* and *C16orf35* surround the α hemoglobin gene cluster in frog, chicken and human, and one of the α and β hemoglobin clusters in platypus and opposum. However, whereas this cluster appears twice in Atlantic salmon, the second cluster in zebrafish, tetraodon and medaka (hereafter Cluster 2) is characterized by a different set of shared genes; specifically, the presence of *Aquaporin-8* and *Rho-GTPase-activating protein*, although tetraodon is lacking the former and zebrafish is lacking the latter. In addition, tetraodon exhibits a copy of *Rhomboid family member 1* on Cluster 2 as well as Cluster 1. Stickleback and Atlantic salmon, however, appear to have lost Cluster 2 entirely. Instead, the stickleback genome only has one hemoglobin cluster (Cluster 1), whereas that of Altantic salmon shows two copies of Cluster 1.

A dot plot generated using the JDotter software [39] comparing the sequenced BACs from Atlantic salmon chromosomes 3 and 6 showed that the regions surrounding the hemoglobin genes are >95% similar between the two chromosomes, with variations

only within the hemoglobin gene regions (Additional file 3, Figure S1). This further suggests that the two Atlantic salmon hemoglobin gene containing chromosomes or regions are homeologous (i.e., represent duplicated copies of the same cluster as the result of a WGD event). Thus, we hypothesize that the WGD at the base of the teleost lineage produced Cluster 1 and Cluster 2, which remain in the zebrafish, medaka and tetraodon lineages, that Cluster 2 was lost in the stickleback lineage, and that Cluster 2 was lost within the salmonid lineage prior to the WGD, which yielded two copies of Cluster 1. This hypothesis is also supported by the fact that the Atlantic salmon chromosome arms 3q and 6q (where the hemoglobin gene clusters are located) share nine duplicated genetic markers [38].

6.3.5 Phylogenetic analysis of teleostean hemoglobin genes

The results of the phylogentic analysis (Figures 6-4 and 6-5 for α and β genes, respectively) suggest that the hemoglobin genes cluster according to functional similarity, which corresponds to sequence similarity. This is expected given the high sequence similarity and short nature of the hemoglobin genes. Specifically, in Figure 6-5, all of the non-Bohr β hemoglobin genes (SsaChr3β3, SsaChr3β5, SsaChr3β6, SsaChr3β8, SsaChr6β4 and SsaChr6β6) form a distinct clade with no other hemoglobin genes, further supporting that there are no β globin genes lacking the Bohr effect in the other fish species examined (see Discussion). Additionally, many genes that were annotated as embryonic within Ensembl (identified with “emb” following the species name) clustered closely, which provides some suggestion as to candidate Atlantic salmon embryonic hemoglobin genes (see Discussion).

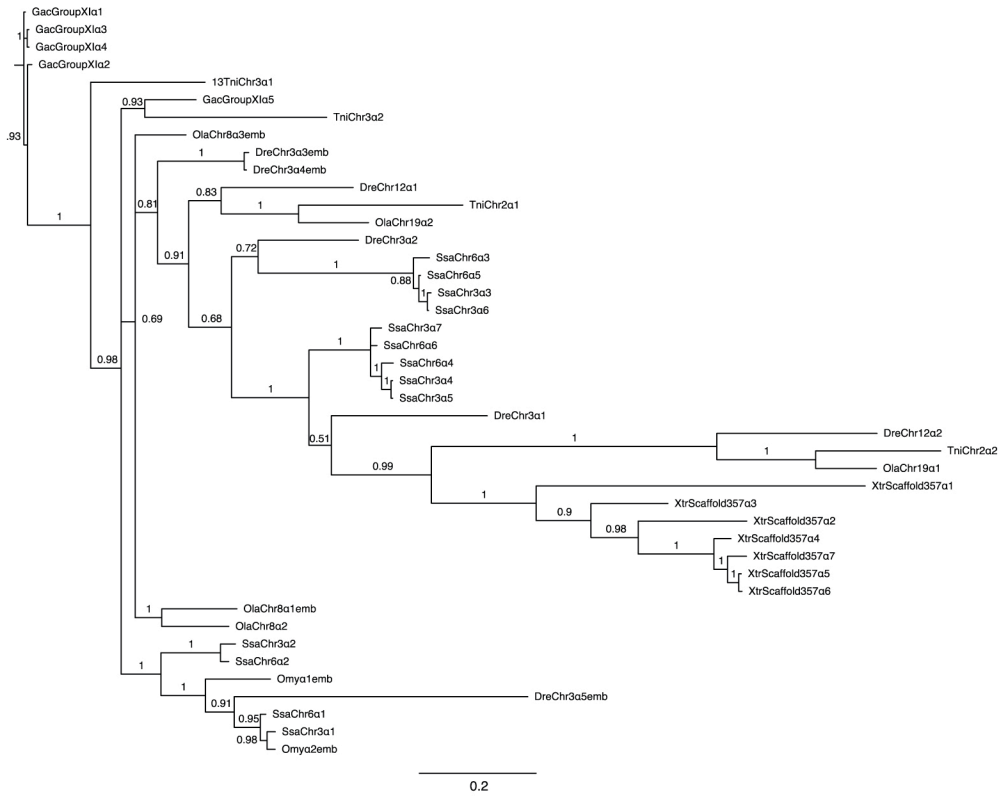

Figure 6-4 Phylogenetic tree of teleost α hemoglobins

Phylogenetic tree of teleost α hemoglobins. The α and β hemoglobin cDNAs (exclusive of untranslated regions) annotated within the Ensembl 54 database for medaka, zebrafish, tetraodon and stickleback, as well as those identified in Atlantic salmon here and the hemoglobin genes identified as embryonic within rainbow trout [28] were independently aligned using EBioX [70]. Phylogenetic trees were constructed using the a Bayesian approach with (5 runs, 100,000 generations, 40% burn-in period) within the TOPALi V.2 software package [71] running the MrBayes program [72] under the best selected model (SYM). For simplicity, as well as to clearly indicate the source chromosome of the gene, the teleostean hemoglobin genes were named using the same system used to name those of Atlantic salmon. That is, an abbreviated three letter (genus species) name followed by chromosome/linkage group name followed by α or β followed by a number indicating the sequential order of the genes from 5' to 3' as defined by Ensembl (Additional file 4, Table S6-3). Hemoglobin genes that were previously identified via expression analysis as being expressed exclusively during embryogenesis, and that are identified as embryonic within the Ensembl 54 database are denoted with "emb" following the assigned gene name. Branch numbers indicate posterior probabilities.

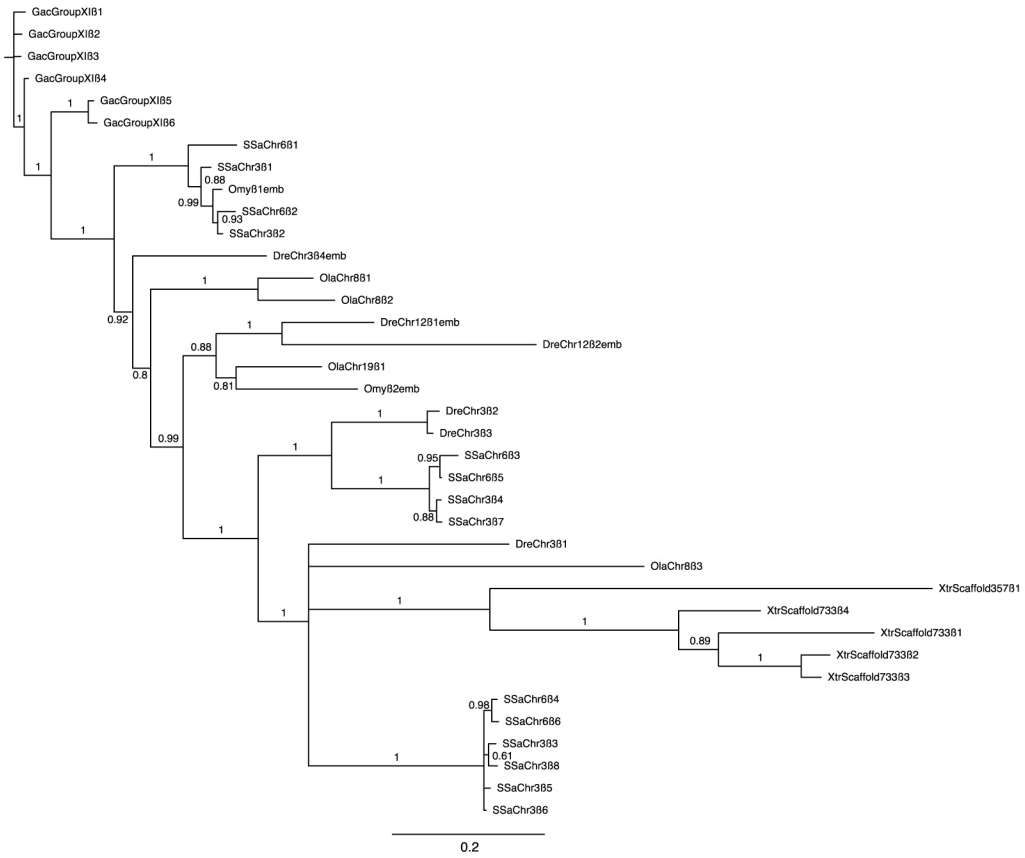

Figure 6-5 Phylogenetic tree of teleost β hemoglobins.

See legend to Figure 4 for details.

In both trees, the *X. tropicalis* hemoglobin genes formed their own clade, although they did not form distinct outgroups, which again, may be a function of the high similarity between the hemoglobin genes across species. All annotated medaka, zebrafish, stickleback and tetraodon α and β hemoglobin genes that were used to generate the phylogenetic trees (i.e., all that were identified within the Ensembl 54 database) are provided in Additional file 4, Table S6-3 by species with the corresponding name assigned by us for comparison purposes (see Methods), as well as the Ensembl gene ID, chromosome/linkage group, start and stop location and strand of transcription.

6.4 Discussion

6.4.1 Number of hemoglobin gene clusters and whole genome duplications

Notably, there are not four clusters of hemoglobin genes in Atlantic salmon even though the hemoglobin clusters were already duplicated within the teleost lineage prior to the salmonid-specific WGD event that took place between 20 and 120 million years ago [20, 21]. Furthermore, our comparative genomic analysis, as well as the high similarity in the non-coding regions of the two fps, suggests that Atlantic salmon exhibit a duplicated copy of one cluster (Cluster 1), and are missing the second cluster (Cluster 2), which is still seen in medaka, zebrafish and tetraodon (see Figures 6-3 and 6-6). Figure 3 shows a schematic diagram of the predicted genes surrounding the hemoglobin genes in Atlantic salmon as well as four annotated teleost genomes, while Figure 6-6 depicts the phylogenetic relationships of the studied teleost fishes (adapted from [40]), and illustrates

our hypothesis of the evolutionary events that took place to produce the observed chromosomal arrangements of the teleost hemoglobin genes. With respect to the other teleosts studied, subsequent to the teleost WGD, which produced Clusters 1 and 2, zebrafish and medaka and tetraodon appear to have maintained both hemoglobin gene clusters, whereas stickleback and Atlantic salmon have lost Cluster 2. In addition, zebrafish exhibits an apparent inversion in Cluster 1 such that *ELAV-like protein 3* is located on the opposite side of the hemoglobin genes, far downstream from *Rhomboid family member 1* with several unshared genes between them. Within the tetraodon genome, Cluster 2 also exhibits some shuffling compared to those of the other genomes, and, interestingly, contains *Rhomboid family member 1*, which is also found on Cluster 1. These relationships will be clarified by further analysis and in-depth annotation of the full-length hemoglobin gene repertoires of other teleost species as more of them undergo full genome sequencing.

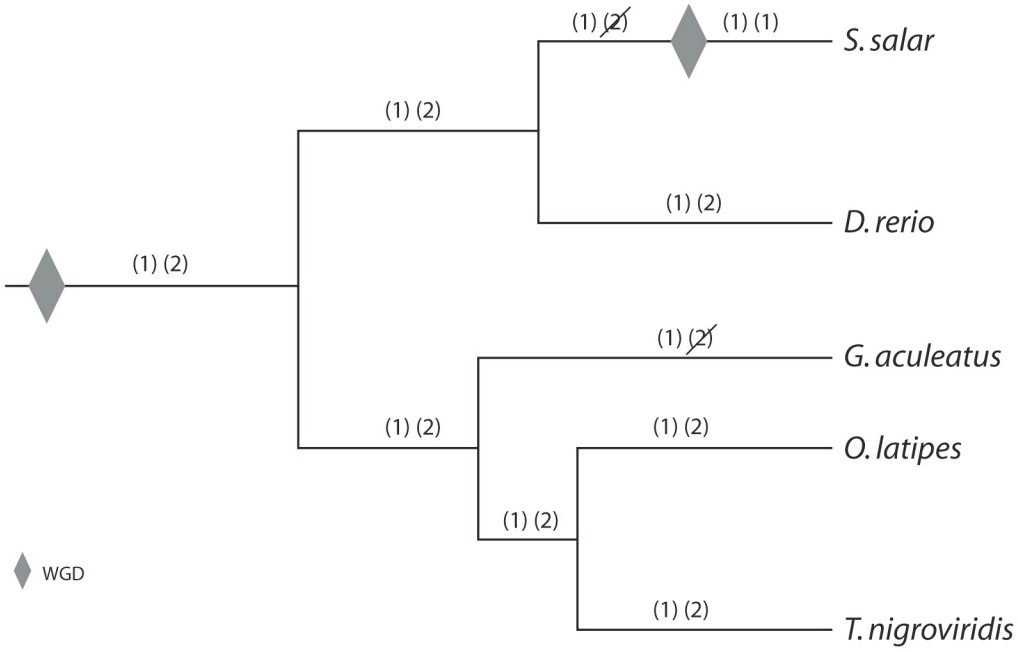

Figure 6-6 Schematic representation of the evolution of teleostean hemoglobin gene clusters

Whole genome duplication (WGD) events are indicated by grey diamonds. The two hemoglobin gene clusters resulting from the teleost WGD are represented as (1) and (2) for Cluster 1 and Cluster 2, respectively (see text). Loss of a hemoglobin gene cluster by excision is indicated by a diagonal slash across that cluster. Although the genome sequence is available for the pufferfish, *T. rubripes*, the fugu genome was not included in this analysis because the published hemoglobin arrangement of two hemoglobin gene clusters, one containing only α hemoglobin genes and one containing both α and β hemoglobin genes [15] did not agree with the annotation results of the latest fugu genome assembly reported within the Ensembl database.

With respect to the salmonid lineage, we propose that Atlantic salmon lost Cluster 2 prior to the salmonid-specific WGD, which duplicated Cluster 1, thus producing the two copies of Cluster 1 and lack of Cluster 2 seen within the Atlantic salmon genome today. We also recognize the possibility of an alternatively pathway, which involves tetraploidization of hemoglobin gene Clusters 1 and 2 as predicted by the salmonid-specific WGD (i.e., producing two copies of each cluster), followed by subsequent loss of both copies of Cluster 2 within the salmonid lineage. However, this involves two separate excision events subsequent to the salmonid WGD, whereas the former hypothesis only involves one such event prior to the salmonid WGD, and is thus the more parsimonious route. Note that the loss of the two clusters must have taken place by excision of the entire regions or chromosome loss as opposed to gradual degradation of the hemoglobin genes or gene silencing, or we would have expected our hemoglobin probes to hybridize to footprints of old hemoglobin clusters in other regions of the genome.

6.4.2 Conservation of order and orientation of α and β hemoglobin genes

The tail-to-tail orientation and alternating order of the α and β hemoglobin genes were fairly well conserved throughout both hemoglobin gene clusters, although both hemoglobin gene clusters exhibited some apparent disruptions to these patterns (see Results). However, given that no hemoglobin gene footprints or ghost genes could be found within these regions, we predict that these changes took place via gradual shuffling, gene loss via excision and pseudogenization over time. Such lineage-specific

gains and loses of hemoglobin genes have also been reported throughout the mammalian lineage [8,35]

6.4.3 Number of hemoglobin genes in Atlantic salmon

Our results revealed that the Atlantic salmon genome contains substantially more paralogous α and β hemoglobin gene copies than previously published [24,25,26]. Furthermore, there are more copies of the α and β hemoglobin genes within the Atlantic salmon genome than in other teleosts whose genomes have been sequenced. Specifically, of the teleost genomes examined, that of zebrafish contains the most hemoglobin genes, with six β hemoglobin genes and seven α hemoglobin genes, compared to 13 and 14 putatively functional α and β hemoglobin genes, respectively, in Atlantic salmon. Note, however, that some of the putatively functional hemoglobin genes do not have an EST associated with them, perhaps suggesting that the actual number of functional genes should be reduced accordingly. However, this would still leave the Atlantic salmon with more α and β hemoglobin genes than any of the other teleosts examined to date.

Numerous reports suggest that mammalian hemoglobin levels are implicated in the increased oxygen affinity of blood in situations of adaptation to altitude-induced hypoxia (reviewed by [41]), thus suggesting that hemoglobin levels contribute to survival in low oxygen environments. Further, Hoffman et al. [35] suggest that variation in hemoglobin gene copy number may be a source of regulatory variation affecting physiological differences in blood oxygen transport and aerobic energy metabolism in mammals. Indeed, it has been proposed that the capacity of fish to colonize a wide range of habitats is directly related to their hemoglobin systems [42]. In addition, a study using

real-time quantitative PCR demonstrated that hypoxic conditions induce complex responses in hemoglobin gene expression in zebrafish [43]. Thus, the extensive array of hemoglobins in Atlantic salmon may reflect the diverse range of environmental conditions that an individual salmon must endure throughout its lifecycle as it migrates from freshwater streams to open ocean and back.

6.4.4 Identification of β hemoglobins lacking the Bohr effect

The Bohr effect is the phenomenon that the affinity of hemoglobin for O_2 is affected by pH. Specifically, an increase in the blood pCO_2 shifts the oxygen dissociation curve to the right, resulting in the release of O_2 and thereby enabling more efficient gas exchange between blood and tissues. This is the result of the oxy-deoxy conformational change and allosteric interactions between O_2 and $H+/CO_2$ binding sites of the hemoglobin molecule. Hemoglobin molecules lacking the Bohr effect are able to retain O_2 under conditions of elevated acidity caused by increased oxygen consumption [44]. Therefore, the non-Bohr hemoglobin may function as an emergency oxygen supplier when an organism is exercising vigorously, such as when a fish is escaping a predator, catching prey, or swimming against a current.

The Bohr effect depends on the intricate arrangement and interactions of all cation and anion binding sites in the hemoglobin molecule and involves a number of contributing amino acid groups that have not yet been fully elucidated [45]. Indeed, a comparison of mammalian, avian and teleost fish hemoglobins suggested that several different histidine and non-histidine sites contribute to the Bohr effect in different species to varying degrees [45]. It is widely accepted, however, that a greater overall histidine

content in the hemoglobin molecule correlates with an increased Bohr effect [46], with the C-terminal histidine residue accounting for up to 50% of the effect [45,47,48]. In Atlantic salmon, the non-Bohr β hemoglobin exhibits phenylalanine at this position [25]. In addition, in Atlantic salmon, the non-Bohr β hemoglobin has 147 amino acids (vs. 146 in the Bohr molecule), including the initiator methoinine, and the amino acid at position 93 is alanine [26,49]. We used these three characteristics to identify a total of six putatively functional non-Bohr β hemoglobin genes within the Atlantic salmon genome (Figure 6-1; Additional file 2, Table S6-2B). Conversely, no β hemoglobin genes identified within the medaka, zebrafish, stickleback or tetraodon genomes exhibited all three of these hallmarks of the non-Bohr β hemoglobin. In addition, a recent PCR-based exploration of the Atlantic cod genome found no β hemoglobin genes exhibiting these characteristics [50], implying an absence of the non-Bohr hemoglobins in this species. All of the Atlantic salmon non-Bohr β hemoglobins formed a distinct clade in the phylogenetic analysis, which reflects their common structural elements and therefore highly similar sequences, and lends further support to the finding that no other fish species examined possesses non-Bohr hemoglobin genes (Figure 6-5).

This apparently high number of non-Bohr β hemoglobin gene copies within the Atlantic salmon genome may be attributable to the Atlantic salmon life history, with its extensive migratory range and the need to swim upstream into fresh water habitats to spawn. In contrast, all of the model teleosts studied so far inhabit relatively consistent environments with little variation in depth, temperature or salinity. Additionally, although Atlantic cod, a non-model teleost that does not appear to possess non-Bohr β hemoglobin genes, inhabit depths from the surface up to 600 m and undertake seasonal migration, at

no point in their lifecycle do they inhabit freshwater [50]. Thus, future studies of the hemoglobin gene repertoires of other migratory salmonids, such as the Pacific salmon species, as well as land-locked freshwater salmonids, in addition to expression profiling of hemoglobin genes at different life stages will provide further insight to this phenomenon.

6.4.5 Embryonic hemoglobin genes

To date, there has been no published study examining temporal expression of hemoglobin genes in Atlantic salmon. Indeed, the most closely related species for which this has been done is rainbow trout [28], for which there is no published genome sequence as well as no comprehensive examination of the rainbow trout hemoglobin repertoire such as this one. Given the complex life history of salmon and the lack of available expression data, we could not confidently assign the title of embryonic to any Atlantic salmon hemoglobins at this time. However, it is noteworthy that SsaChr3α1 and SsaChr6α1 form a clade with Omyα2emb, as well as Omyα1emb and an embryonic *Danio rerio* α globin, DreChr3α5emb (Figure 6-4), and that SsaChr3α2 and SsaChr6α2 cluster closely with this clade. Figure 6-5 shows that SsaChr3β1, SsaChr6β1, SsaChr3β2 and SsaChr6β2 form a clade with Omyβ1emb. These phylogenetic relationships suggest that these Atlantic salmon hemoglobin genes may be embryonic. Also worth noting is that all of these are the first genes in the 5'–3' direction on their respective chromosomes. In mammals, temporal expression of hemoglobin genes correlates with spatial location on the chromosome, with the first upstream hemoglobin gene being the first expressed [3]. This further suggests that these genes (i.e., SsaChr3α1 and SsaChr3α2, SsaChr6α1,

SsaChr6α2, SsaChr3β1 SsaChr3β2, SsaChr6β1 and SsaChr6β2) encode candidate embryonic hemoglobins. However, further analysis, in particular, detailed expression profiling of hemoglobin genes during all life stages, is required to examine these hypotheses.

6.5 Conclusions

We found that, despite the Atlantic salmon genome having gone through at least two WGD events relative to tetrapods, which would result in four predicted hemoglobin gene clusters, only two such clusters were present. Furthermore, the Atlantic salmon genome appears to exhibit two copies of one of the duplicated ancestral teleost hemoglobin gene clusters, and has presumably lost the other cluster. We also found that the Atlantic salmon genome harbors substantially more hemoglobin genes than the other teleosts for which the hemoglobin gene repertoires have been identified, and that they possess several hemoglobin genes that appear to encode non- Bohr β hemoglobins. We suggest that these characteristics of the Atlantic salmon hemoglobin genes reflect the dynamic life history of Atlantic salmon.

6.6 Methods

6.6.1 Identification of Atlantic salmon hemoglobin BACs

As part of the Genomic Research on All Salmonids Project (GRASP), an Atlantic salmon BAC library was produced from a partial EcoRI restriction enzyme digest of DNA from a

Norwegian aquaculture strain male fish (CHORI-214 segments 1–3). There are 312,000 BAC clones in the library with an average insert size of 190,000 bp, which have been arrayed onto nylon membranes, thus representing an 18.8-fold coverage of the Atlantic salmon genome [29]. BACs were fingerprinted using HindIII and arranged into contigs to create the first physical map of a salmonid genome [30]. Approximately 210,000 BAC end-sequences have been determined, corresponding to approximately 3.5% of the Atlantic salmon genome. Information on the Atlantic salmon BACs and physical map can be found at [31].

To identify the Atlantic salmon BACs containing the hemoglobin genes, oligonucleotide probes (~40-mers) were designed from the published Atlantic salmon Clone 6 (GenBank accession number X97289) for α, β and non-Bohr β hemoglobins. PCR primers sets were also designed to span intron 1 of all three hemoglobin types. In addition, primer sets were designed to span intron 1 of the embryonic α and β hemoglobins of rainbow trout, with a ~40-mer forward primer that was used for hybridization probing (GenBank accession numbers: α: AB015448; β: AB015450). All primers and probes were designed using Primer3 ver. 0.4.0 [51] and are provided in Additional file 1, Table S6-1. The oligonucleotide probes were end-labeled with ^{32}PγATP using T4 polynucleotide kinase and hybridized to six BAC filters at a time as described by Johnstone et al. [52]. Briefly, prehybridization was carried out in 5x saline-sodium citrate buffer (SSC), 0.5% sodium dodecyl sulfate (SDS), and 5 x Denhardt's solution at 65°C. The filters were washed three times for 1 hr at 50°C, in 1 x SSC and 0.1% SDS. Filters were exposed to phosphor screens that were scanned using the Typhoon Imaging System and visualized using ImageQuant software, giving an image of the ^{32}P-labeled

hybridization-positive BACs containing the hemoglobin markers. The hybridization-positive BAC clones were picked from the library, cultured in 5 mL LB media containing chloramphenicol (50 μg/mL) overnight at 37ºC shaking at 250 rpm and made into glycerol stocks for subsequent PCR verification that they indeed contained hemoglobin genes. Hybridization and PCR-positive BACs for the hemoglobin genes were matched to two fingerprint scaffolds (fps) within the Altantic salmon physical map (fps135 and fps1046; [30,31]).

6.6.2 BAC shotgun library generation and sequencing

Using a combination of hybridization probing and PCR (see above) to screen all BACs within the suspected hemoglobin gene containing regions, we identified two overlapping BACs from each of fps1046 (BACs S0055H05 and S0014B03) and fps135 (BACs S0155C07 and S0079J05) spanning the entire Atlantic salmon hemoglobin gene repertoire. That is, all primers amplified hemoglobin gene products within these BACs, and no additional BACs that were not contained within the four BACs as determined by the Atlantic salmon physical map yielded PCR products using the hemoglobin gene primers. The four BACs were sequenced using standard Sanger sequencing of a shotgun library as previously described [53]. Briefly, BAC DNA was isolated from each of the hemoglobin-containing BACs using Qiagen's Large Construct kit as per the manufacturer's directions (Qiagen, Mississauga, Ont. Canada). The kit includes an exonuclease digestion step to eliminate *E. coli* genomic DNA. The purified BAC DNA was sheared by sonication and blunt-end repaired. The sonicated DNA was size

fractioned by agarose gel electrophoresis and 2–5 kb fragments were purified using the QIAquick Gel Extraction Kit (Qiagen, Mississauga, Ont. Canada). DNA fragments were ligated into pUC19 plasmid that had been digested with SmaI and treated with shrimp-alkaline phosphatase to produce de-phosphorylated blunt ends. The ligation mixture was used to transform supercompetent *E. coli* cells (XL1-Blue; Stratagene, La Jolla, CA. USA). Transformed cells were cultured overnight at 37°C on LB/agar plates supplemented with ampicillin (200 µg/mL) and 1,920 (5 × 384 well plates) clones were sent to the Michael Smith Genome Sciences Centre, Vancouver, BC Canada, for sequencing. The sequences were analyzed for quality using PHRED [54], assembled using PHRAP [55], and viewed using Consed version 15.0 [56]. BAC assemblies were complicated by the repetitive nature of the Atlantic salmon genome [32]. Assemblies were hand-finished to fill gaps (i.e., join sequence contigs) as best as possible using primer walking; however, primers could not be designed to join some sequence contigs that ended in repetitive sequence, or often primers amplified multiple products (i.e., showed a multiple bands or a smear on an agarose gel).

6.6.3 Linkage analysis and chromosome assignment

The sequences of BACs S0055H05 and S0155C07 (representing fps 1046 and 135, respectively) were screened for microsatellite markers that were variable (i.e., informative) within the two Atlantic salmon SALMAP mapping families, Br5 and Br6, each of which contains two parents and 46 offspring [37]. Markers Ssa10067BSFU and Ssa10051BSFU were identified within S0055H05 and S0155C07, respectively. PCR primers were designed to amplify the region containing the microsatellite. The forward

primer for each pair contained an M13 sequence tag that was used for genotyping analysis. Genotyping results were analyzed with LINKMFEX ver. 2.3 [57].

A single end-sequenced BAC containing the α, β and non-Bohr β hemoglobin genes was chosen from each of fps135 and fps1046 (S0155C07 and S0055H05, respectively) to be used for chromosome assignment. Approximately 1 μg of BAC DNA was purified (Qiagen mini-prep kit; Qiagen, Mississauga, Ont., Canada) and used for FISH analysis to identify the Atlantic salmon chromosomes containing the hemoglobins. Comparison of the results of the linkage and FISH analysis of the Atlantic salmon hemoglobin BACs contributed to the recent integration of the Atlantic salmon linkage map and karyotype [38].

6.6.4 BAC sequence annotation and identification of putatively functional and pseudogenized hemoglobin genes

All sequence contigs >1,000 bp within the assembled sequences were analyzed using a variety of sequence similarity searches and gene prediction algorithms that have been incorporated into an in-house computational pipeline and database [58] described previously [53]. Briefly, sequences entering this pipeline were screened (masked) for repetitive elements using RepeatMasker 3.2.6 [59] and were searched against the NCBI nr (non-redundant) and Atlantic salmon EST [33] databases using BLAST [34]. A GENSCAN gene model prediction algorithm [60] was used to predict introns and exons, and the resulting predictions were searched against the Uniref50 (clustered sets of sequences from UniProt Knowledgebase) database [61]. Finally, a rps-BLAST search against the NCBI CDD [62] was conducted to provide additional information with

respect to the predicted genes. Any sequence contigs that were identified as containing hemoglobin-like genes by this pipeline were put through an additional series of annotation steps to ensure consistent calling of predicted open reading frames (ORFs) and that we did not miss any putatively functional or dysfunctional hemoglobin-like genes. Specifically, the masked and unmasked sequences were analysed using the ab initio gene prediction programs GENSCAN [60], GeneMark [63], FGENESH [64] and HMMGene [65], and the results of each prediction program were compared. In an attempt to identify putative psuedogenes or hemoglobin gene remnants, HMMer (v1.8.5) [66] was used to scan the sequence with hemoglobin exon-specific HMMs. Hemoglobin genes were labeled as putatively functional if they were predicted to be intact hemoglobins by our annotation procedures, and met all of the following criteria:

1) The genes were predicted to contain three exons and two introns.

2) The predicted exons were of the appropriate sizes, meaning that predicted splice junction sites aligned with those of known functional hemoglobins and start and stop codons were present in the appropriate places.

3) The final predicted protein included 147 or 148 amino acids for β hemoglobins and 143 amino acids for α hemoglobins.

The sequences of any predicted ORFs that aligned to hemoglobins but failed to meet any one of the above criteria were examined by eye for potential miss-calling by our annotation procedures. Specifically, we looked for historical footprints of missing exons that were not recognized by the pipeline, interruptions to splice sites as well as insertions and deletions of stop and start codons, potential sequencing errors and frame-shift

mutations caused by insertions or deletions. We also examined by eye any putative three-exon ORFs identified by our pipeline that were not recognized by a BLAST search as encoding hemoglobins to determine whether they may be remnant hemoglobin genes or previously undefined hemoglobin-like genes. If, after this hands-on annotation, predicted proteins still did not meet the above criteria, the sequences were defined as putative pseudogenes. Furthermore, any regions for which the predicted orientation (i.e., α hemoglobin genes transcribed on the negative strand and β hemoglobin genes on the positive) and alternating order of the α and β hemoglobin genes was disrupted were examined by eye for putative remnant hemoglobin exons and introns. All such regions were aligned against intact hemoglobin genes using ClustalW2 [67], and predictions were made as to whether these regions represented footprints of historical hemoglobin genes.

All annotated hemoglobin genes were assigned an Ssa (*Salmo salar*) name followed by Chr3 for fps135 and Chr6 for fps1046 to denote its chromosomal location, then α or β to identify the gene encoded, and finally a number corresponding to its order relative to the other α or β genes on that chromosome from 5' to 3'.

6.6.5 Identification of β hemoglobins lacking the Bohr effect

We defined β hemoglobin genes exhibiting three hallmarks of a lack of the Bohr effect were defined as putative non-Bohr β hemoglobin genes. These hallmarks include: 1) the non-Bohr β hemoglobin has 147 amino acids, including the initiator methionine; 2) the C-terminal amino acid is phenylalanine; 3) the amino acid at position 93 is alanine [48, 27].

6.6.6 Identification of genes surrounding hemoglobin gene clusters in Atlantic salmon and other teleosts

Atlantic salmon BAC sequences surrounding the hemoglobin gene clusters were annotated using our in-house annotation pipeline described above. This provided a preliminary prediction of the genes lying within the sequenced regions. However, note that different components of the pipeline can differ in their gene predictions, and that a full, comprehensive annotation of these regions as well as the rest of the Atlantic salmon genome will be completed with sequencing of the whole genome.

The genes surrounding the hemoglobin clusters in four annotated teleost genomes, medaka, zebrafish, tetraodon (*Tetraodon nigroviridis*) and stickleback (*Gasterosteus aculeatus*), were identified using the Pfam ID for the hemoglobin protein family (PF00042) [68] available within Biomart [69]. Specifically, the Ensembl 54 Genes database was searched using the appropriate genome-specific dataset for hemoglobins. Once the genomic locations of the hemoglobin genes were determined, the region surrounding the hemoglobin gene clusters was expanded until at least five predicted genes were identified on either side of the hemoglobin gene cluster, or until no additional common or shared genes could be identified. This allowed us to examine the synteny of the regions surrounding the hemoglobin genes, and thereby generate hypotheses of hemoglobin gene evolution in teleost fishes. Note that the fugu genome was not included in this analysis because the published hemoglobin arrangement of two hemoglobin gene clusters, one containing only α hemoglobin genes and one containing both α and β hemoglobins [15] did not agree with the annotation results of the latest fugu genome assembly reported within the Ensembl database. Instead, only one apparent hemoglobin

cluster containing both α and β hemoglobin genes could be identified on fugu scaffold 3, and when the genes surrounding this cluster were compared to those of the other genomes examined, no shared genes (i.e., no conserved synteny) could be found.

6.6.7 Phylogenetic analyses

The α and β hemoglobin cDNAs (exclusive of untranslated regions) annotated within the Ensembl 54 database for medaka, zebrafish, tetraodon and stickleback, as well as those identified in Atlantic salmon here and the hemoglobin genes identified as embryonic within rainbow trout [28] were independently aligned using EBioX [70]. We examined the relationships among the gene products by constructing phylogenetic trees using the a Bayesian approach with (5 runs, 100,000 generations, 40% burn-in period) within the TOPALi V.2 software package [71] running the MrBayes program [72] under the best selected model (SYM). For simplicity, as well as to clearly indicate the source chromosome of the gene, the teleostean hemoglobin genes were named using the same system used to name those of Atlantic salmon. That is, an abbreviated three letter (genus species) name followed by chromosome/linkage group name followed by α or β followed by a number indicating the sequential order of the genes from 5' to 3' as defined by Ensembl (Additional file 4, Table S6-3). Note that hemoglobin genes of medaka [73], zebrafish [16] and rainbow trout [28] that were previously identified via expression analysis as being expressed exclusively during embryogenesis, and that are identified as embryonic within the Ensembl 54 database are identified within the phylogenetic trees (denoted with "emb" following the assigned gene name) as well as within Additional file 4, Table S6-3.

6.7 Acknowledgements

We are grateful to the Michael Smith Genome Sciences Centre for sequencing the BAC shotgun libraries. This work was supported by funding from Genome Canada, Genome BC, and the Province of British Columbia (WSD and BFK) and the United States Department of Agriculture Grant # 2006-04814 (to RBP) as well as graduate scholarships from Weyerhaeuser Corporation (Weyerhaeuser Molecular Biology Scholarship) and Simon Fraser University (Molecular Biology Graduate Fellowship and President's Research Stipend; NLQ).

6.8 References

1. Hardison R: **Hemoglobins from bacteria to man: Evolution of different patterns of gene expression.** *J Exp Biol* 1998, **201**:1099–1117.

2. Strandberg B: **Chapter 1: Building the ground for the first two protein structures: Myoglobin and Haemoglobin.** *J. Mol. Biol.* 2009, **932**:2–10.

3. Fromm G, Bulger, M: **A spectrum of gene regulatory phenomena at mammalian b globin gene loci.** *Biochem Cell Biol* 2009, **87**: 781–790.

4. Goodman M, Moore W: **Darwinian evolution in the genealogy of haemoglobin**. *Nature* 1975, **253**:603–608.

5. Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE: **Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes**. *Nature* 1982, **298**:297–300.

6. Lanfranchi G, Pallavicini A, Laveder P, Valle G: **Ancestral hemoglobin switching in lampreys.** *Dev Biol* 1994, **164**:402–408.

7. Hoffman FG, Opazo JC, Storz JF: **Rapid rates of lieage-specific gene duplication and deletion in the α hemoglobin gene family**. *Mol Biol Evol* 2008, 25:591–602.

8. Opazo JC, Hoffman FG, Storz JF: **Differential loss of embryonic hemoglobin genes during the radiation of placental mammals.** *Proc Nat Acad Sci USA* 2008, **105**:12950–12955.

9. Wheeler D, Hope R, Cooper JB, Gooley AA, Holland RAB: **Linkage of the β-like, ω-like gene to the α-like hemoglobin genes in an Australian marsupial supports the chromosome duplication model for separation of hemoglobin gene clusters.** *J Mol Evol* 2004, **58**:642–652.

10. Jeffreys AJ, Wilson V, Wood D, Simons PJ: **Linkage of adult α and β-hemoglobin genes in *X. laevis* and gene duplication by tetraploidization.** *Cell* 1980, **21**:555–564.

11. Patel VS, Cooper SJB, Deakin JE, Fulton B, Graves T, Warren WC, Wilson RK, Graves JAM: **Platypus hemoglobin genes and flanking loci suggest a new insertional model for β-hemoglobin evolution in birds and mammals**. *BMC Biol* 2008, **6**:34 July 25.

12. Near TJ, Parker SK, Detrich HW 3rd: **A genomic fossil reveals key steps in hemoglobin loss by the antarctic icefishes.** *Mol Biol Evol* 2006, **23**:2008–2016.

13. Giordano D, Russo R, Coppola D, di Prisco G, Verde C: **Molecular adaptations in haemoglobins of notothenioid fishes.** *J. Fish Biol* 2010, **76**:301–318.

14. Nelson JS: *Fishes of the World*. 4th Ed. New York: John Wiley and Son; 2006.

15. Gillemans N, McMorrow T, Tewari R, Wai AW, Burgtorf C, Drabek D, Ventress N, Langeveld A, Higgs D, Tan-Un K, Grosveld F, Philipsen S: **Functional and comparative analysis of hemoglobin loci in pufferfish and humans.** *Blood* 2003, **101**:2842–2849.

16. Brownlie A, Hersey C, Oates AC, Paw BH, Falick AM, Witkowska HE, Flint J, Higgs D, Jessen J, Bahary N, Zhu H, Lin S, Zon L: **Characterization of embryonic hemoglobin genes of the zebrafish.** *Dev Biol* 2003, **255**:48–61.

17. Maruyama K, Shigeki Y, Iuchi I: **Evolution of hemoglobin genes of the medaka *Orzias latipes* (Euteleostei; Beloniformes; Oryziinae).** *Mech Develop* 2004, **121**:753–769.

18. Taylor JS, Van de Peer Y, Braasch I, Meyer A: **Comparative genomics provides evidence for an ancient genome duplication event in fish.** *Phil Trans R Soc Lond* 2001, **356**:1661–1679.

19. Thorgaard GH, Bailey GS, Williams D, Buhler DR, Kaattari SL, Ristow SS, Hansen JD, Winton JR, Bartholomew JL, Nagler JJ, Walsh PJ, Vijayan MM, Devlin RH, Hardy RW, Overturf KE, Young WP, Robison BD, Rexroad C, Palti Y: **Status and opportunities for genomics research with rainbow trout.** *Comp Biochem Physiol B Biochem Mol Biol* 2002, **133**:609–646.

20. Ohno, S: *Evolution by Gene Duplication*. New York: Springer-Verlag; 1970.

21. Allendorf FW, Thorgaard GH: **Tetraploidy and the evolution of salmonid fishes.** In *Evolutionary Genetics of Fishes*. Edited by Turner BJ. New York: Plenum Press; 1984:55–93.

22. **Consortium for Genomics Research on All Salmonids Program** [www.cgrasp.org]

23. Wolff JP, Gannon F: **cDNA and deduced amino acid sequence of the *Salmo salar* (Atlantic salmon) adult hemoglobin α chain.** *Nucleic Acids Res* 1989, **17**:4369.

24. Wagner A, Deryckere F, McMorrow T, Gannon F: **Tail-to-tail orientiation of the Atlantic salmon α- and β-hemoglobin genes.** *J Mol Evol* 1994, **38**:28–35.

25. McMorrow T, Wagner A, Deryckere F, Gannon F: **Structural organization and sequence analysis of the hemoglobin locus in Atlantic salmon.** *DNA Cell Biol* 1996, **15**:407–414.

26. McMorrow T, Wagner A, Harte T, Gannon F: **Sequence analysis and tissue expression of a non-Bohr β-hemoglobin cDNA from Atlantic salmon.** *Gene* 1997, **189**:183–188.

27. Moghadam HK, Ferguson MM, Danzmann RG: **Evidence for Hox gene duplication in rainbow trout (*Oncorhynchus mykiss*): a tetraploid model species.** *J Mol Evol* 2005, **61**:804–818.

28. Maruyama K, Shigeki Y, Iuchi I: **Characterization and expression of embryonic hemoglobin in the rainbow trout, *Onchorhynchus mykiss*: Intra-embryonic initiation of erythropoiesis.** *Develop Growth Differ* 1999, **41**:589–599.

29. Thorsen J, Zhu B, Frengen E, Osoegawa K, de Jong PJ, Koop BF, Davidson WS, Høyheim B: **A highly redundant BAC library of Atlantic salmon (*Salmo salar*): an important tool for salmon projects.** *BMC Genomics* 2005, **6**:50.

30. Ng SH, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJ, de Jong PJ, Koop BF, Krzywinski MI, Lubieniecki K, Marra MA, Mitchell LA, Mathewson C, Osoegawa K, Parisotto SE, Phillips RB, Rise ML, von Schalburg KR, Schein JE, Shin H, Siddiqui A, Thorsen J, Wye N, Yang G, Zhu B: **A physical map of the genome of Atlantic salmon, Salmo salar.** *Genomics* 2005, **86**:396–404.

31. **Asalbase: Atlantic salmon genomics database** [www.asalbase.org]

32. de Boer JG, Yazawa R, Davidson WS, Koop BF: **Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids.** *BMC Genomics* 2007, **8**:422.

33. Leong JS, Jantzen SG, von Schalburg KR, Cooper GA, Messmer AM, Liao NY, Munro S, Moore R, Holt RA, Jones SJM, Davidson WS, Koop BF: ***Salmo salar*** **and** ***Esox lucius*** **full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidiztion genome.** *BMC Genomics* 2010, **11**: 279.

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–10.

35. Hoffmann FG, Storz JF: **The αD-hemoglobin gene originated via duplication of an embryonic α-like hemoglobin gene in the ancestor of tetrapod vertebrates.** *Mol Biol Evol* 2007, 24:1982–90.

36. Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmbhatt S, Rosner J, Rexroad CE 3rd, McGowan CR, Davidson WS: **A salmonid EST genomic study: genes, duplications, phylogeny and microarrays.** *BMC Genomics* 2008, **9**:545.

37. Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Koop BF, Hoyheim B, Lien S, Lubieniecki KP, Moghadam HK, Park J, Phillips RB, Davidson WS: **Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon).** *BMC Genomics* 2008, **9:**557.

38. Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF, Danzmann RG, Davidson WS: **Assignment of Atlantic salmon** ***(Salmo salar*****) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout** ***(Oncorhynchus mykiss*****).** *BMC Genetics* 2009, **10**:46**.**

39. Brodie R, Roper RL, Upton C: **JDotter: a Java interface to multiple dotplots generated by dotter.** *Bioinformatics* 2004, 20:279–81.

40. Steinke D, Salzburger W, Meyer A: **Novel relationships among ten fish model species revealed based on phylogenomic analysis using ESTs.** *J Mol Evol* 2006, **62**:772–784.

41. Samaja M, Crespi T, Guazzi M, Vandegriff KD: **Oxygen transport in blood at high altitude: role of the hemoglobin-oxygen affinity and impact of the phenomena related to hemoglobin allosterism and red cell function.** *Eur J Appl Physiol* 2003, **90**:351–359.

42. Verde C, Parisi E, di Prisco G: **The evolution of thermal adaptation in polar fish.** *Gene* 2006, **385**:137–145.

43. Roesner A, Hankeln T, Burmester T: **Hypoxia induces a complex response of hemoglobin expression in zebrafish (*Danio rerio*).** *J Exp Biol* 2006, **209**:2129–2137.

44. Jensen FB: **Red blood cell pH, the Bohr effect, and other oxygenation-linked phenomena in blood O_2 and CO_2 transport.** *Acta Physiol Scand* 2004, **182**:215–227.

45. Berenbrink M: **Evolution of vertebrate haemoglobins: Histidine side chains, specific buffer value and Bohr effect.** *Respir Physiol Neurobiol* 2006, **154**:165–184.

46. Jensen FB: **Hydrogen ion equilibria in fish haemglobins.** *J Exp Biol* 1989, **143**:225–234.

47. Riggs A: **The Bohr effect.** *Annu Rev Physiol* 1988, **50**:181–204.

48. Lukin JA, Ho C: **The structure-function relationship of hemoglobin in solution at atomic resolution.** *Chem Rev* 2004, **104**:1219–1230.

49. Brunori M: **Molecular adaptation to physiological requirements: The hemoglobin system of trout.** *Curr Topics Cell Regul* 1975, **9**:1–39.

50. Halldórsdóttir K, Árnason E: **Multiple linked β and α hemoglobin genes in Atlantic cod: A PCR based strategy of genomic exploration.** Mar Genomics 2009, **2:**169–181.

51. Rozen S, Skaletsky HJ: **Primer3 on the WWW for general users and for biologist programmers.** In *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Edited by S Krawetz, S Misener. New Jersey: Humana Press; 2000:365–386.

52. Johnstone KA, Ciborowski KL, Lubieniecki KP, Chow W, Phillips RB, Koop BF, Jordan WC, Davidson WS: **Genomic organization and evolution of the vomeronasal type 2 receptor-like (OlfC) gene clusters in Atlantic salmon, *Salmo salar*.** *Mol Biol Evol* 2009, **26**:1117–1125.

53. Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, Davidson, WS: **Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome.** *BMC Genomics* 2008*,* **9**:404.

54. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175–185.

55. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186–194.

56. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, 8:195–202.

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

58. **GRASP: Genomic Research on Atlantic Salmon Project** [grasp.mbb.sfu.ca]

59. **Repeat Masker** [www.repeatmasker.org]

60. Burge CB, Karlin S: **Finding the genes in genomic DNA.** *Curr Opin Struct Biol* 1998, **8**:346–354.

61. **UniProt Knowledgebase** [www.pir.uniprot.org/database.nref]

62. **NCBI CDD (Conserved Domain Database)** [www.ncbi.nlm.nih.gov/Structure/cdd/cdd]

63. Lomsadze A, Ter-Hovhannisyan V, Chernoff Y, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucleic Acids Res* 2005, **33**:6494–6506.

64. Salamov A, Solovyev V: **Ab initio gene finding in Drosophila genomic DNA.** *Genome Res* 2000, **10**:516–522.

65. Krogh A: **Two methods for improving performance of an HMM and their application for gene finding.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:179–186.

66. **HMMer** [hmmer.janelia.org]

67. **ClustalW2** [www.ebi.ac.uk/Tools/clustalw2/]

68. **Wellcome Trust Sanger Institute Pfam database** [pfam.sanger.ac.uk/]

69. **Biomart** [www.biomart.org/index.html]

70. **EbioX** [http://www.ebioinformatics.org/index.html]

71. Milne L, Linder D, Bayer M, Husmeier D, McGuire G: **TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops.** *Phylogenetics* 2009, **25**: 126–127.

72. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754–755.

73. Maruyama K, Yasumasu S, Iuchi I. **Characterization and expression of embryonic and adult globins of the teleost *Oryzias latipes* (medaka).** *J Biochem* 2002, **132:**581–589.

6.9 Additional Files

Appendix File 9 *Additional file 1, Table S6-1*

Primer and probe sequences. [a] ~40-mer forward primers were also used as hybridization probes.

Appendix File 10 *Additional file 2, Table S6-2A–C*

Identified Atlantic salmon putatively functional and pseudogenized hemoglobin genes. S2A) Identified putatively functional Atlantic salmon α hemoglobin genes with chromosome, sequence contig number and approximate location (kb), strand of transcription, most highly similar Atlantic salmon EST cluster (if any), whether the gene has a corresponding full-length EST, whether the gene matches any of the previously published Atlantic salmon hemoglobin genes at the amino acid level and whether the gene is identical to any of those identified on the other Atlantic salmon chromosome. S2B) Identified putatively functional Atlantic salmon β hemoglobin genes with chromosome, sequence contig number and approximate location (kb), strand of transcription, most highly similar Atlantic salmon EST cluster (if any), whether the gene has a corresponding full-length EST, whether the gene matches any of the previously identified Atlantic salmon hemoglobin genes at the amino acid level, whether the gene is identical to any of those identified on the other Atlantic salmon chromosome, and whether the β hemoglobin gene possesses the hallmarks of lacking the Bohr effect. S2C) Putatively identified Atlantic salmon hemoglobin pseudogenes with chromosome, sequence contig, location (kb), direction and descriptions of each exon.

Appendix File 11 *Additional file 3, Figure S6-1:*

Dot plot comparing the sequenced BACs from Atlantic salmon chromosomes 3 and 6. Regions surrounding the hemoglobin genes are >95% similar. The dot plot was generated using the software JDotter [39]. The shared non-hemoglobin genes [*Dedicator of cytokinesis 6* (*DOCK6*), *Dedicator of cytokinesis 7* (*DOCK7*) and *Rhomboid 5 homolog 1*] within these regions are indicated. For chromosome 3, seven parts (P1–P7) are shown (bottom axis), representing seven sequence contigs, the first of which (sequence contig 49) does not contain any hemoglobin genes and is therefore not shown in Figure 6-1.

Appendix File 12 *Additional file 4, Table S6-3*

Putative α and β hemoglobin genes from other teleosts and *Xenopus tropicalis* used to generate phylogenetic trees. The table lists all predicted intact α and β hemoglobin genes indentified within Biomart [69] for teleost genomes that have been sequenced and annotated (medaka, zebrafish, tetraodon, danio) and *Xenopus tropicalis*, which was used as an outgroup. For each hemoglobin gene identified, the table lists the species, chromosome or scaffold, start and stop positions, strand of transcription, Ensembl gene ID and our assigned gene name used in the phylogenetic trees.

7: Conclusions

7.1 Summary

For my PhD thesis, I set out to identify genes involved in temperature tolerance and the response to thermal stress in Arctic charr, a species that stands to play a substantial role in the aquaculture industry, but for which there are few genomics resources. Taken together, the chapters presented herein have accomplished this goal, and in doing so, have also provided extensive insight into the nature of the Atlantic salmon genome as well as those of salmonids in general. Specifically, Chapter 2 describes an attempt at using the genomics resources for Atlantic salmon to conduct *in silico* skim sequencing of a UTT QTL by aligning BAC-end sequences with the fully sequenced and annotated medaka genome. The results of this work, which were inconclusive and highlighted numerous limitations of this approach, suggested that more in-depth and comprehensive approaches are required to identify candidate UTT genes given the complexity of the Atlantic salmon genome and the lack of a suitable reference genome. Chapter 3 presents another attempt at exploring one of the previously-identified UTT QTL, this time using Roche/454's GS-FLX pyrosequencing, a next-generation sequencing technology that, at the time, had yet to be tested for its ability to handle sequencing and assembling complex genomic sequences, such as those of the Atlantic salmon genome, *de novo*. The conclusions of this chapter state that, although the Roche/454's GS-FLX assembly was adequate for identifying the genes within the QTL region, the numerous gaps in the final sequence scaffold suggested that read length was still insufficient to provide a desirable assembly

quality. For Chapters 4 and 5, I conducted expression profiling of Arctic charr exposed to acute, lethal and moderate, prolonged temperature stress, respectively. Chapter 4 reports that *COUP-TFII*, hemoglobins and heat shock proteins appear to play a role in the responses to short term lethal temperature exposure, and that these genes may help differentiate between thermo-tolerant and intolerant fish. Chapter 5 implicates ribosomal proteins and heat shock proteins in the response to chronic, moderate temperature stress. Finally, for Chapter 6, I identified, sequenced and annotated the full hemoglobin repertoire of Atlantic salmon. This work confirmed that Atlantic salmon have two hemoglobin gene clusters, and revealed that there are more hemoglobin genes in Atlantic salmon than in any other fish studied to date. In addition, Atlantic salmon harbour several putative non-Bohr hemoglobin genes, which may act as emergency oxygen suppliers during periods of stress or prolonged exertion.

7.2 Implications of advances in genomics technologies

Since my research was conducted, several advances have been made in the field of genomics. Perhaps the most significant advancement in terms of this thesis is the ongoing improvement of next generation sequencing technologies. Currently, even though 454 pyrosequencng remains the only next generation sequencing technology that can reliably achieve read lengths greater than 150 bp, advances in the ability to generate paired-end libraries of various insert sizes have enhanced the ability to assemble short-reads, and thus the ability of these technologies to tackle more complex *de novo* sequencing projects. For example, a draft sequence of the Panda (*Ailuropoda melanoleura*) genome was recently generated using the Illumina Genome Analyzer sequencing technology, which, with an average read length of 52 bp, achieved an N50 contig size of 40 Kbp and

an N50 scaffold size of 1.3 Mb by sequencing libraries constructed with 500 bp, 2 Kb, 5 Kb and 10 Kb insert sizes (1). Another short-read technology, ABI's SOLiD sequencing system (read length ~25–50 bp), is advertised at targeting *de novo* sequencing of small organisms, or for filling in sequence gaps left after an initial low-coverage assembly by capillary sequencing (www.appliedbiosystems.com). Therefore, although the primary focus of these short-read next generation sequencing technologies is the re-sequencing of individual genomes, they are showing clear potential for playing an integral part in *de novo* sequencing projects.

454 Life Sciences, who still claims the only currently available long-read next generation sequencing technology, has largely focused their research and development on improving read length and assembly capabilities for the purpose of *de novo* sequencing of large, complex genomes. This push towards read length improvements was likely fuelled by the results of our research (i.e., Chapter 3), which clearly showed that Sanger-like read lengths are required for such complex sequencing projects. In late 2008 (after our manuscript was published), 454 Life Sciences launched the GS FLX Titanium platform, which uses a new set of reagents on the original GS FLX instrument. Reported read lengths for the Titanium technology are in the 400–500 bp range, and the 454 Life Sciences website claims that achievable read lengths with the Titanium platform rival those of the Sanger technology (i.e., 800–1000 bp) (www.my454.com). Recently, a draft sequence of the Atlantic cod genome was generated using the GS FLX Titanium system, with a total read data set consisted of 63.6 million shotgun reads (peak read length 503 bases) and 20.2 million paired-end reads (peak read length 389 bases, including the linker sequence; average pair half length 81 bases), with paired end jumping distances of 1 to 2

Kb, 3 Kb, 8 Kb and 20 Kb (2).

The implications of the improvements of next generation (and now "third generation") sequencing towards *de novo* genome sequencing, including those of read length, paired-end library construction and assembly algorithms (3), are extensive. The drastic reduction in cost and resources needed to sequence a full genome surely mean that we will increasingly see the full genome sequences of rare, endangered, or otherwise environmentally or medically relevant species. This, of course, has synergistic consequences given that as more genomes can be used as reference sequences, more genomes can be sequenced, perhaps eventually filling in the evolutionary tree of life with fully sequenced genomes on every branch. This is certainly the case for salmonids, as the anticipated sequences of Atlantic salmon and rainbow trout could facilitate the sequencing the genomes of additional salmonids, such as that of Arctic charr, or of other closely related non-salmonid species. At the very least, these genomes will greatly enhance the capacity for Arctic charr genomics research, which could build on the findings presented within this thesis.

Aside from their implications towards *de novo* sequencing, the ongoing advances in sequencing technologies have enabled the development and advancement of other genomics techniques, such as SNP detection and the identification of RAD (restriction site associated) markers, which facilitate genetic mapping and enable large-scale genome screening of organisms. In addition, RNA-seq, or whole transcriptome sequencing, provides in-depth expression profiling of individual organisms. These technologies have started to replace, or at least compliment more traditional genomics approaches such as

microarrays and qPCR, and could certainly provide more insight into the results presented herein.

7.3 Future work

The primary goal of this research was to provide a foundation for identifying genetic markers for tolerance to elevated temperatures that can be used for marker assisted selection for an Arctic charr broodstock that can withstand higher than normal temperatures, and that is relatively robust to temperature fluctuations. It is also possible that any UTT markers developed for aquaculture could be used for screening wild populations of Arctic charr, and perhaps other salmonids, for conservation purposes. Indeed, the data presented herein go a long way towards accomplishing these goals. Several genes and gene families were identified for their putative involvement in UTT, and thus should be further examined in subsequent studies. Specifically, the next steps for this project include examining these genes for differences among temperature tolerant and intolerant Arctic charr. Firstly, we have developed PCR primers that span the *COUP-TFII* genes, and have begun sequencing these genes in the temperature tolerant and intolerant fish identified in Chapter 3. These sequences will be examined for genetic differences that are transcribed into different protein products, and that thus may play a role in determining the phenotypes of the fish. Secondly, we are currently designing additional temperature trials with larger sample sizes and that include fish with known pedigrees. These trials will involve longer exposure times, thus enabling more in depth analyses of the temperature stress response. Additionally, we intend to develop a live sampling assay, so the expression responses of individual fish can be monitored by qPCR

throughout the exposure. Finally, we are pursuing the development of an assay for using expression levels of given genes as markers of temperature tolerance, which could avoid the need for genetic profiling of individuals, and thus serve as a tool for mass screening of fish families and populations. These ongoing projects will be developed and conducted by the Davidson lab group in collaboration with Icy Waters Ltd.

7.4 References

1. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463(7279):311-317.

2. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzen A, Winer R, Knight J, Vogel J, Aken B, Andersen O, Lagesen K, Tooming-Klunderud A, Edvardsen RB, Tina KG, Espelund M, Nepal C, Previti C, Karlsen BO, Moum T, Skage M, Berg PR, Gjoen T, Kuhl H, Thorsen J, Malde K, Reinhardt R, Du L, Johansen SD, Searle S, Lien S, Nilsen F, Jonassen I, Omholt SW, Stenseth NC, Jakobsen KS. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011;477(7363):207-210.

3. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y: Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet. 2011;56(6):406-414.

Appendix

The CD-ROM, attached, forms part of this work. File details are provided in the table below and the legends/descriptions of the files are provided at the end of each chapter throughout the text. .xlsx files can be opened with MSExcel or other spreadsheet programs; .pdf and .jpg files can be opened with Adobe Acrobat or another viewing program.

Appendix Data Files