Identification and Annotation of Full-length Genes in Atlantic Salmon (*Salmo salar*)

by

Jong S. Leong
B.Sc., McGill University, 1997

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

MASTER OF SCIENCE

in the Department of Biology

Supervisory Committee

Identification and Annotation of Full-length Genes
in Atlantic Salmon (*Salmo salar*)

by

Jong S. Leong
B.Sc., McGill University, 1997

Supervisory Committee
Dr. Ben F. Koop (Department of Biology)
Supervisor

Dr. John Taylor (Department of Biology)
Departmental Member

Dr. Christopher Upton (Department of Biology and Microbiology)
Outside Member

Abstract

Supervisory Committee
Dr. Ben F. Koop (Department of Biology)
Supervisor

Dr. John Taylor (Department of Biology)
Departmental Member

Dr. Christopher Upton (Department of Biochemistry and Microbiology)
Outside Member

Large-scale expressed sequence tags (ESTs) in Atlantic salmon (*Salmo salar*) are examined to answer questions regarding salmonid transcriptomes. ESTs represent raw and incomplete gene sequences that need to be read, assembled and analyzed with computer software. The goal of this thesis was to develop an automatically curated and publicly accessible set of annotated full-length genes, representing a near-complete transcript set for *Salmo salar*. In turn, these genes provide the framework for studies in gene expression, conservation, and molecular evolution. The work presented here also touches on the results of a molecular evolution study, as an example of how full-length gene identification can be used to answer biological questions.

Previous to this study, a limited number of Atlantic salmon cDNA libraries and ESTs were known. To further the goal of determining complete gene sequences, highly enriched full-length cDNA libraries and full-length libraries were created and sequenced, resulting in the ability to identify a large number of full-length reference genes. Together, all libraries represent a diverse pool of transcriptome sequences for *Salmo salar*.

The goal of producing an accurate large-scale full-length gene set on a duplicated genome is not trivial. Complete systems for this objective do not readily exist. EST sequencing, EST assembly, and data storage, are just a few of the initial computational issues that are addressed. Once these issues are resolved, the multi-step workflow of full-length gene determination is described. The final challenge involving the development of a concise and universally accessible system for visualization is discussed. The resulting computational framework that has been developed is shown to be able to handle the intricacies and the size of a duplicated salmonid genome. It has been largely accepted that Atlantic salmon have undergone a recent genome duplication. Gene paralogs provide one source of evidence for this event. Analysis of paralogs revealed signatures of asymmetric evolution possibly due to relaxation of selective pressure.

This thesis provides a complete Bioinformatics analysis pipeline to analyze and to visualize a set of full-length reference genes for Atlantic salmon. Using full-length genes as a framework, the topic of molecular evolution was addressed to show evidence of asymmetrical evolution among gene duplicates. The full-length reference genes, along with ESTs and all putative transcripts, have been made publicly available. These results serve as a valuable genomic resource for next-generation sequencing and for all other salmonid research endeavours.

Table of Contents

Supervisory Committee ii
Abstract iii
Table of Contents v
List of Tables vii
List of Figures viii
List of Acronyms ix
Acknowledgments x
Dedication xi
Chapter 1 Introduction 1
Thesis Overview 1
Competition for natural resources 4
Dilemma 8
Role of molecular biology 9
Role of Bioinformatics 11
Mutations 12
Mobile elements 14
Segmental duplications 17
Genome duplications 19
Implications of genome duplication 23
Atlantic salmon as a target for transcriptome sequencing 26
Chapter 2 Gene Identification and Analysis of Salmonid Expressed Sequence Tags 31
Summary 31
Introduction 33
Methods and Materials 36
Tissues, RNA, Aquaculture and Sampling 36
cDNA libraries 36
Sequencing, Sequence Analysis, and Contig Assembly 37
Gene phylogenetic analysis 40
Microarray Clone selection 41
Microarray hybridizations 44
Microarray analyses 45
Results and Discussion 46
cDNA libraries 46
Transcript analysis: sequence and assembly 48
Salmonid comparisons 53
Transcriptome representation 55
Full-length analysis 55
Salmonid EST, assembly, ORF and annotation database 56
Salmonid phylogeny and gene duplication 58
Salmonid 32 K microarray 65
Chapter 3 Gene Identification and Full-Length Reference Genes in Salmonids 67
Summary 67
Introduction 68
Material and Methods 72

Tissues, RNA, and Sampling 72
cDNA Libraries 72
Sequencing, Sequence Analysis, and Contig Assembly 73
FLcDNA contig identification 73
Reference FLcDNA identification 75
Reference FLcDNA identification using individual clone assembly 77
Reference FLcDNA assessment 78
Selection of homologous genes 78
Determination of alignment regions 79
Sequence alignment 80
d_N/d_S estimation 80
Gene Ontology analysis 81
Results 82
Full-length cDNA library construction and analysis 82
Identification of *S. salar* Full-Length cDNA contigs from existing EST assemblies 83
E. lucius ESTs 85
Reference Full-Length cDNA identification using individual clone assembly 85
Reference Full-Length cDNA assessment 87
S. salar and *E. lucius* alignments 89
Pairwise comparisons to determine d_S and ω 90
d_S and ω for tree segments 93
Gene Ontology analysis 94
Discussion 96
Chapter 4 Conclusions 102
Bibliography 105
Appendix A Atlantic Salmon Contig Viewer (Figure 7) - Website Access Statistics 119
Appendix B Publication List 120
Appendix C Presentation List 122
Appendix D Full-Length Reference Gene Data 124

List of Tables

Table 1. Salmonid cDNA libraries, sequencing and assembly summary statistics for data provided in this study 47
Table 2. Summary of salmonid ESTs and contig assemblies 50
Table 3. Cross-species comparisons of contig transcripts 52
Table 4. Gene sets used in phylogenetic analysis..... 60
Table 5. Cross species hybridization results for the salmonid 32 K cDNA microarray 66
Table 6. Summary of confirmed and unique reference FLcDNAs in contig sets for *S. salar* and *E. lucius* 89
Table 7. Proportions of genes in GO categories 95

List of Figures

Figure 1. World human population with moderate estimates 5
Figure 2. Worldwide fisheries capture 6
Figure 3. Worldwide fish consumption shows an increase in fish consumption consistent with population growth 6
Figure 4. Global aquaculture production by major species groups shows major efforts in aquaculture expansion 7
Figure 5. Molecular evolutionary tree with 5 proposed genome duplications 25
Figure 6. Number of aligned contigs (y-axis) out of 81,398 total contigs is plotted against percent similarity of alignments (x- axis) 53
Figure 7. Screen shot of Atlantic salmon contig viewer 57
Figure 8. Summary of 78 gene set consensus (70%) trees depicting the relationships among the major groups of Salmonidae 63
Figure 9. Schematic of *S. salar* FLcDNA contig identification and reference FLcDNA identification 84
Figure 10. Schematic of *S. salar* reference FLcDNA identification through individual clone assemblies 86
Figure 11. Distributions and means of ORF, 5' and 3' UTR sizes in reference FLcDNAs for (A) *S. salar* (B) *E. lucius* 88
Figure 12. Frequencies of d_S and ω values for comparisons within *S. salar* and *E. lucius* gene trios 91

List of Acronyms

AS	Atlantic salmon
cDNA	complimentary DNA
CDS	coding sequence
d_n	non-synonymous site
d_s	synonymous site
EST	expressed sequence tag
FL	full-length
FLcDNA	full-length cDNA
GO	Gene Ontology
HSPs	high-scoring segment pairs
MYA	millions of years ago
MH	Major Histocompatibility complex
ML	maximum likelihood
NJ	neighbour-joining
ORF	open read frame
RT	rainbow trout
UTR	untranslated region
WGD	whole genome duplication
WGS	whole genome sequencing

Acknowledgments

I would like to thank Dr. Ben Koop, Dr. John Taylor, and Dr. Chris Upton of my supervisory committee. I am also extremely grateful for the encouragement and support from all the researchers, past and present, in my group.

Dr. Ben Koop is inspirational with his research, and is always pushing me to think of the larger overall picture. He is supportive of my goals, and keeps me focused. Dr. John Taylor's patient explanations about genome duplication theory and research, are always greatly appreciated.

Since the day I joined the Koop lab, Glenn Cooper, Dr. Gordon Brown, and Dr. Kris von Schalburg have been helpful and extremely patient getting me up to speed on projects. Glenn and Kris took the time to describe wet-lab experiments and protocols, and Gord provided support on the vast software infrastructure that I inherited from him.

The staff and fellow colleagues at the Centre for Biomedical Research have brought a level of professionalism and humour to the office. Marjorie Wilder has been very knowledgeable and eager to assist me during the rare times that I need to struggle with paperwork.

During my time at the Koop lab, the calibre of work has always been at the highest level. Internationally our work has commanded attention, leading the way for other researchers in our field. This result is no coincidence and I believe that it speaks highly of our team. I could not have asked for a better group of people to work with.

Lastly, I would like to thank the funders of my work. Without Genome Canada, Genome British Columbia, and the cGRASP consortium, none of this research would have been possible.

Dedication

I would like to dedicate my work to Suk Chun Yee, my grandmother, the most amazing person I know.

Chapter 1
Introduction

Thesis Overview

Large-scale expressed sequence tags (ESTs) in the pseudopolyploid Atlantic salmon (*Salmo salar*) are examined *in silico*. Lack of Atlantic salmon molecular information prompted the need to establish a foundation from which all other biological studies can benefit from. First and foremost, a basic gene set for Atlantic salmon using EST data is established. Using this gene set, the determination of full-length genes is carried out. Lastly, once full-length genes are determined, the question of the characteristics of early evolution in duplicated genes is examined. Accomplishing the goals of this study involved many novel challenges, most notably the lack of a complete Bioinformatics pipeline that automatically integrated all of the individual analyses. Features that were required of such a pipeline included the ability to deal with the duplicated genome and unambiguous full-length cDNA identification. Ultimately, the ability to accurately visualize and share both analysis results and raw data while utilizing an efficient relational database backend was required.

In the first chapter of my thesis, I discuss the motivations and the benefits of Atlantic salmon molecular biology research. I suggest that Bioinformatics is essential with large-scale genomics projects. As the Atlantic salmon has a genome complicated by a recent autotetraploidization event, mobile elements, as well as basal mutation events, the subject of accounting for extensive genome remodelling is explored. The state of current salmonid genomic resources as well as the steps required to achieve the goals of my thesis are outlined.

In the chapter 2 of my thesis, I begin my work by establishing the resources needed to determine full-length genes in Atlantic salmon. These included the creation of EST databases and analysis of large sequence sets from seven separate salmonids, and one non-salmonid species. Using the processed sequences from these eight species, a molecular phylogeny study was performed. The Bioinformatics for a new cDNA microarray used these new sequence sets. My role in this work was as follows: raw EST sequence handling, cDNA statistical analysis, querying the processed EST sequences for assembly, EST assembly and optimization, MySQL database creation, database table design, database table population using assembled EST data (contigs), contig annotation, contig statistical analysis, contig open reading frame predictions, full-length gene analysis using TargetIdentifier, Apache webserver setup, web-based contig viewer creation, assistance with the 32 K microarray probe selection, data formatting for NCBI dbEST submission, NCBI dbEST data submission, NCBI dbEST data updating, and help writing the manuscript. All data population scripts and analyses were done using PERL and Python. 354,061 processed EST sequences from eight fish species were produced and analyzed in this chapter.

In the third chapter of my thesis, I performed detailed characterization of full-length cDNA in Atlantic salmon and northern pike. The work was a continuation of the EST work from the previous chapter, and included three new full-length Atlantic salmon EST libraries, and 29,221 new EST sequences for northern pike. Detailed full-length cDNA characterization required a novel prediction algorithm, so that the 5' and 3' untranslated regions, the coding sequence, and the polyA tail could be identified correctly. Once full-length sequences were identified, a gene duplication study of Atlantic salmon was carried

out. Results from this study dealt with the detection of selection pressures before and after the proposed 4R salmonid gene duplication event, using northern pike as a non-duplicated outgroup. Because EST assemblies may represent the combination of multiple unique transcripts stemming from different alleles, recent duplicates and sequencing errors, reference full-length cDNAs from single completely sequenced cDNA clones were determined. Reference full-length cDNAs unambiguously represent a single allele of a single gene. Existing full-length cDNAs as well as new EST libraries were examined to determine reference full-length cDNAs for both Atlantic salmon and northern pike. My role in this work was as follows: raw EST sequence handling, cDNA statistical analysis, querying the processed EST sequences for assembly, EST assembly and optimization, MySQL database creation, database table design, database table population using assembled EST data (contigs), contig annotation, contig statistical analysis, contig open reading frame predictions, creation of a novel full-length gene prediction algorithm, characterization of full-length gene regions, Apache webserver setup, adding additional web-based contig viewer functionality to support detailed full-length gene annotation, data formatting for NCBI dbEST and core nucleotide submission, NCBI dbEST and core nucleotide data submission, and NCBI dbEST and core nucleotide data updating, translation of all gene coding sequences, and writing the manuscript. All data population scripts and analyses were done using PERL and Python.

The final chapter of my thesis summarizes the work that has been carried out on each of the eight distinct species of fish. The implications of the results of the gene duplication study are also reiterated. As the contig viewer is the main gateway to view the complete information of the work presented here, the pattern of web traffic was analyzed to reveal

access patterns. For a one year period (March 15, 2009 – March 7, 2010) the information on the website has been accessed 238,447 times. Of those visits, 7,222 were from a unique IP address. Therefore, each address represents one or more uninterrupted sessions. The contig viewer is a highly accessed platform, and has been refined through feedback from users throughout the world. A complete list of my publications and presentations are listed in Appendix B and Appendix C.

My thesis presents a useful Bioinformatics analysis pipeline that can be used to determine full-length genes in Atlantic salmon. By establishing fundamental molecular knowledge of a keystone species that is of biological and economical importance, I have created an important resource to aid in all future salmonid research.

Competition for natural resources

Seafood represents an important source of food for the world's population. Due to human consumption, wild fish stocks are becoming depleted. These finite fish stocks cannot continually absorb demand from the rapidly expanding human population. The global population currently sits at almost 6.8 billion residents, compared to only 3 billion in 1960 (Department of Economic and Social Affairs, Populaton Division, United Nations 'World Population Prospects' 2009). This population is only expected to increase, with moderate estimates predicting over 9.1 billion by 2050 (Figure 1).

By this time, it is estimated that food production will have to increase by 70% globally to feed the world's population (Food and Agriculture Organization, United Nations 'How to Feed the World in 2050' 2009). The latest report carried out by the Food and Agriculture Organization of the United Nations shows that the world's catch of fish has plateaued since the 1990s (Figure 2). At the same time, population growth has been

steady, while fish consumption per capita has held constant (Figure 3). Implications for these continued trends are clear; without an alternative source of fish at current consumption levels, wild fish stocks will sustain serious, if not permanent, damage. Extinction of a species is a very real possibility for resources that are over-exploited. In North America alone, there has been decimation of many aquatic stocks in the recent past. The west coast of North America once supported large abalone fisheries.

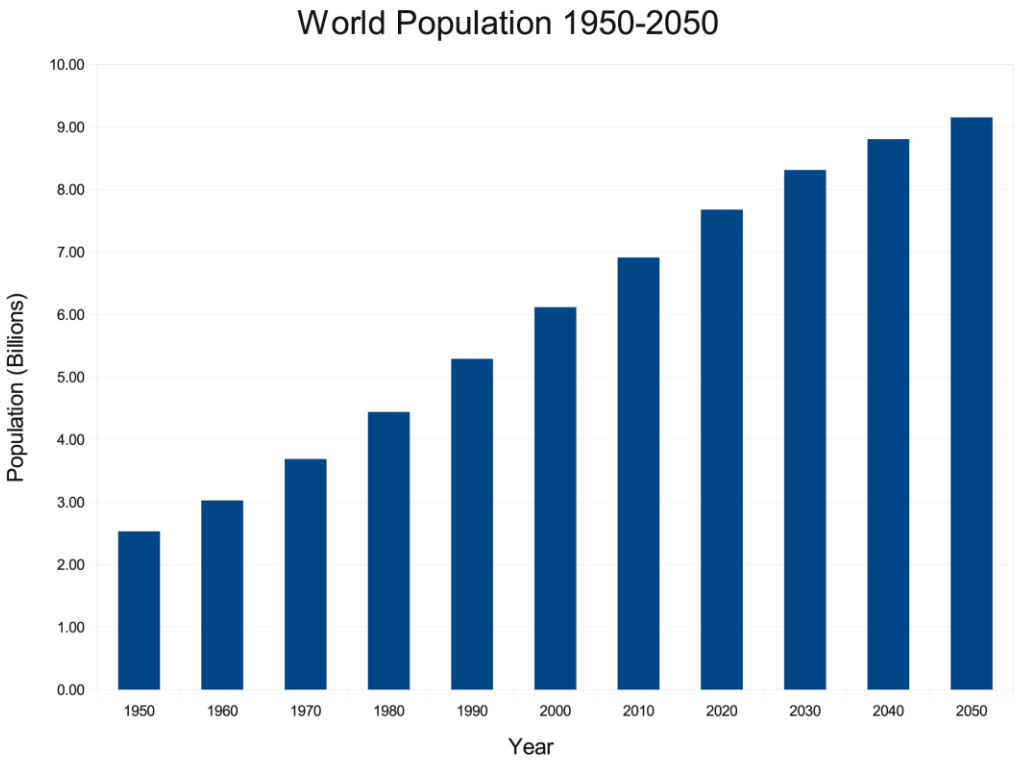

Figure 1. World human population with moderate estimates. The global population is expected to be over 9.1 billion by 2050. Data Source: ('World Population Prospects' 2009).

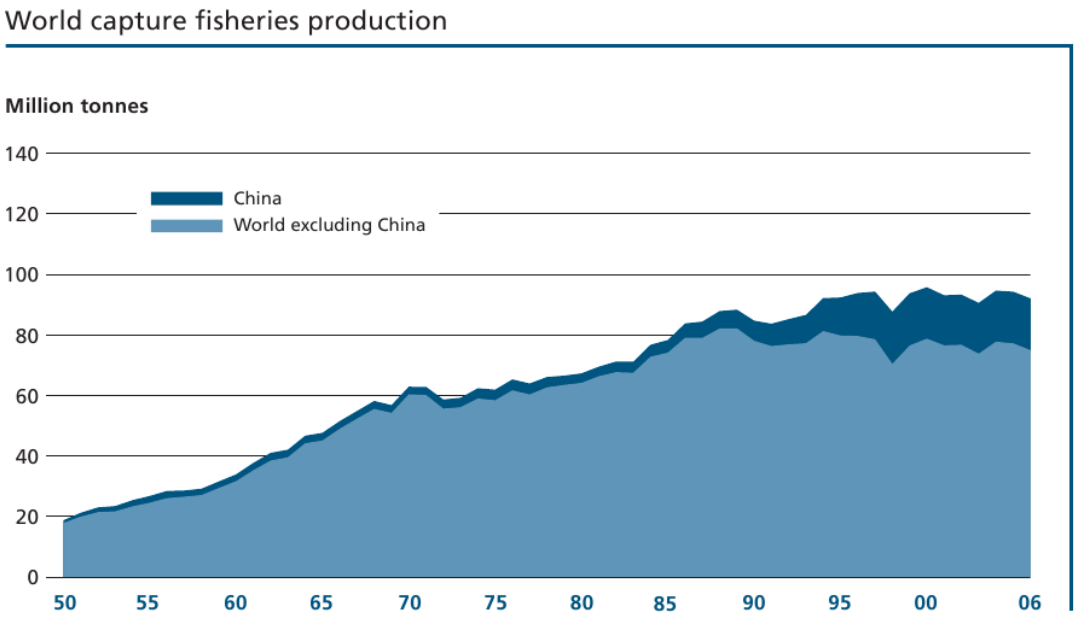

Figure 2. Worldwide fisheries capture. Capture has plateaued in recent years. Source: ('The State of World Fisheries and Aquaculture 2008' 2008).

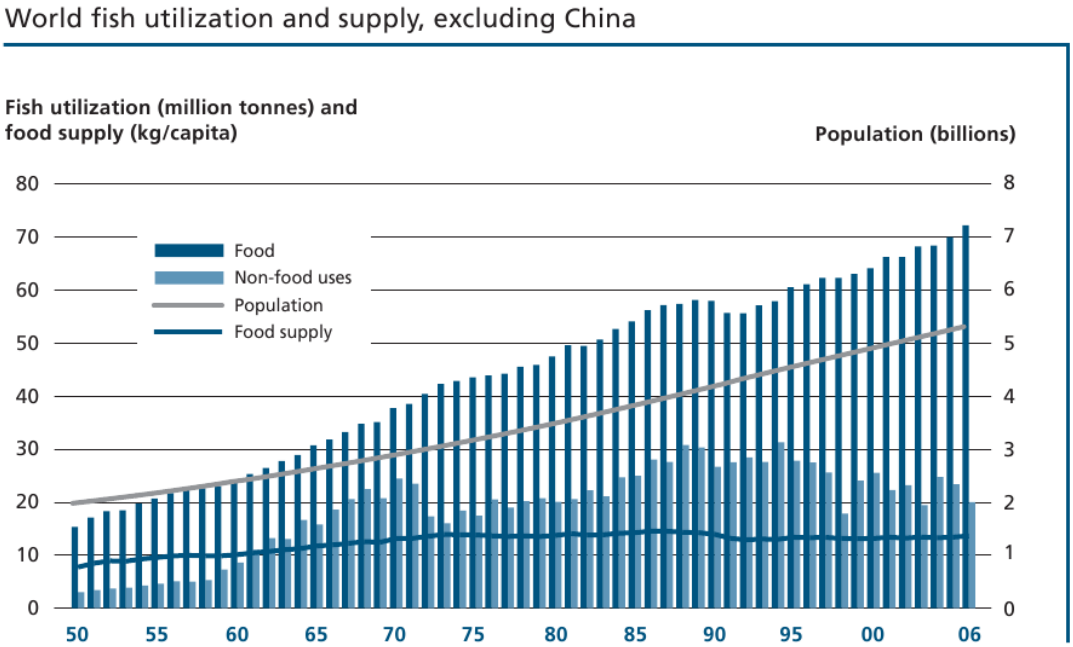

Figure 3. Worldwide fish consumption shows an increase in fish consumption consistent with population growth. Source: ('The State of World Fisheries and Aquaculture 2008' 2008).

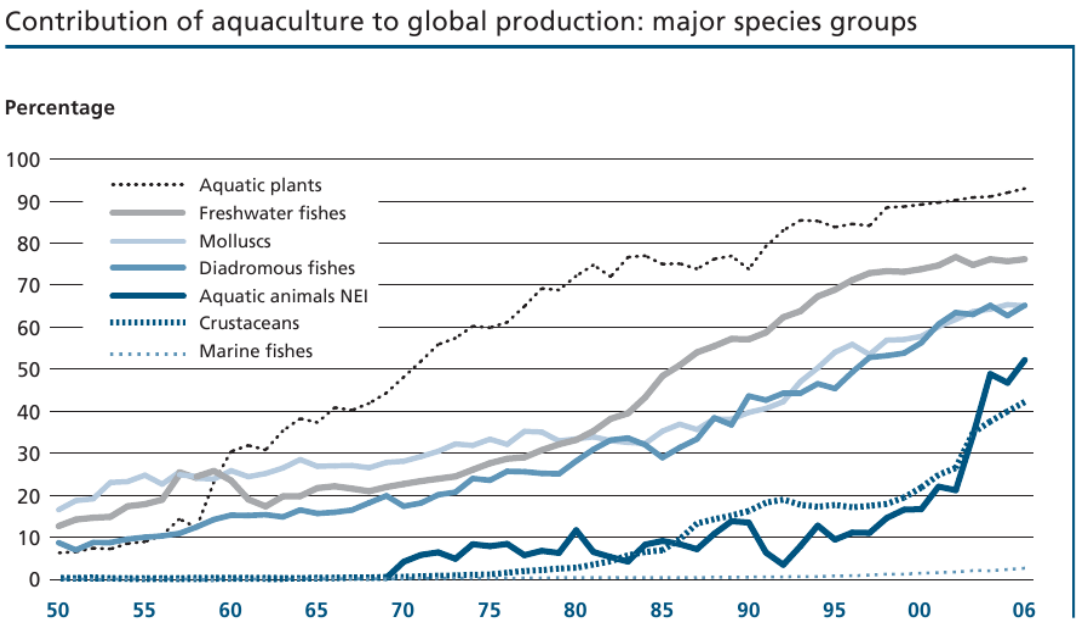

Figure 4. Global aquaculture production by major species groups, shows major efforts in aquaculture expansion. Source: ('The State of World Fisheries and Aquaculture 2008' 2008).

As a result of poor fishing practices, abalone stocks in these areas have now collapsed (Hilborn et al. 2005). In the 1960s and the 1970s, Atlantic Canada witnessed the initial collapse of its cod stocks from overfishing (Lilly 2008). These stocks have still not recovered despite a decades-long fishing moratorium. If protection of wild fish stocks is not made a priority, this pattern is likely to repeat itself on any actively harvested fish species.

There are different possibilities that exist to address the well-being of wild fish stocks. Aquaculture is one approach that has been taken to manage production of a food source that occurs naturally in lakes, rivers, and oceans. It is in fact undergoing a phenomenal expansion, and this growth is faster than any other animal-based food sector (Figure 4). Aquaculture is an immediate and viable option towards increased food production. On the other hand, the health of existing resources must also be considered. Conservation of

wild stocks requires detailed genetic information to accurately distinguish between discrete populations. Using this information, populations can be characterized by the development and use of genetic markers. By understanding the behaviour and size of wild populations, appropriate geographic quotas on harvesting can be implemented.

One thing seems certain, there is no further room for increasing wild marine fish catches. If the world wishes to sustain its current consumption of fish, it must proceed in a more sustainable manner. Before any species is brought to the brink of extinction, its environmental impact and ideally its preservation should be studied in detail. Genomic research is a step towards providing insight on a species at a basic level.

Dilemma

The world's growing population (Figure 1) will naturally result in an increased need for food. Harvesting from wild marine sources has arguably already exceeded their sustainable limits. Therefore, the ability to increase this type of food source will be largely dependent on aquaculture. However, farming is not without impact on the environment. Fish farms often require chemicals to control parasites (Lees et al. 2008) with the hope of increasing yields. Controlling these parasites with drugs can lead to eventual reduced efficacy, which has been documented in various countries (Lees et al. 2008; Saksida et al. 2010).

While drug resistance is of concern, the issue of environmental impact is also important. All types of food production inevitably require varying amounts of fresh water, and associated parcels of land. The amounts of resources that need to be diverted to large-scale food production are usually significant. Resource consumption in turn impacts the environment in terms of the energy needed to manufacture and ultimately

distribute products (e.g., feed, drugs, fuel for machinery). In addition to the long-term effects from production waste and by-products, large-scale production may also lack the properly implemented environmental safeguards. Research is needed to fully understand all levels of environmental impact due to modern high-throughput food production practices.

It is unlikely that demand for fish will diminish; therefore sustainability will be a key to fish production's long-term viability. A clear understanding of the interactions that a species has with its surroundings is essential. Identifying and achieving one particular step towards this ambitious goal is not without its technical challenges. The approaches and their roles and will be discussed in detail.

Role of molecular biology

It is clear in modern biology that in order to begin asking the most fundamental questions, such as how diversity first arose, species should be examined at all levels including the molecular level. Detailed study of genetic codes may reveal insights into the origins and history of diversity.

Research in vertebrate phylogeny has traditionally been studied through a complex system of morphological categorizations. In this method, species are studied through careful observation of physical traits (e.g., shape, size, colour of body parts), or behaviours, and both qualitative and quantitative records of an individual or a group of individuals are compared. Unfortunately these characteristics are often not suitable for directly studying the underlying genetic causal factors that give rise to phenotypic diversity. When studying Earth's biodiversity, variation is important to understanding

adaptation to various environments. Knowledge of these differences is greatly enhanced by fields such as genetics and genomics, that allow for studies in the evolution of genes.

A species' genetic makeup can be considered a fluid entity, changing every generation over millions of years. Those species that reproduce sexually have offspring that represent all genes inherited from their parents. Mutations (i.e., insertions, deletions, duplications), and random genetic drift are also forces that act in a non-adaptive manner (Lynch 2007) on a population. Through any of these events, if any amount of genetic variation can be passed on to subsequent generations, these changes are recorded at the molecular level. As will be illustrated, genetic variation is important to a species long-term survival and adaptation to changing environments.

To understand the importance of genetic variation, species with larger or smaller amounts of genetic diversity can be compared. Heterogeneous populations are likely to be more adaptable to environmental changes. A recent study on the annual return of actively exploited sockeye salmon (*Oncorhynchus nerka*) observed over a period of five decades was completed (Schindler et al. 2010). In this study, the sockeye salmon in Bristol Bay, Alaska were known to be composed of several hundred discrete populations originating from nine major rivers. An important economic source for commercial sockeye salmon fisheries, Bristol Bay has demonstrated 2.2 times less variability in salmon returns when compared to a system of a single homogenous species population. The wide amount of genetic diversity within the single species produces what is known as a 'portfolio effect' providing a level of stability, measured by annual returns. In essence, a diversified portfolio provides a buffering effect against fish that perform poorly. In the case of Bristol Bay sockeye, each of its river systems contained tens to

hundreds of locally adapted populations, providing this overall diversified portfolio. The reduction in salmon return fluctuations for this actively harvested species helps to illustrate the importance of genetic heterogeneity.

In order to study genetic diversity, a comprehensive amount of sequence data must be collected and analyzed. This data, representing an enormous amount of genetic information, requires the use of computation tools as they cannot be manually processed in a timely manner.

Role of Bioinformatics

The flood of raw high-throughput data presents challenges in any genomic analyses. For each piece of datum, analysis must be performed in a consistent fashion to maintain reproducibility. As an illustration of the scope of raw transcript datasets, one can look at the typical number of EST sequences publicly available for a particular organism. NCBI's dbEST database (Boguski et al. 1993) serves as a public central repository for all EST data. As of this writing (August 1, 2010), 66,792,597 ESTs are available for download at dbEST. For example, zebrafish (*Danio rerio*) has 1,481,936, African clawed frog (*Xenopus laevis*) has 677,806, Atlantic salmon has 498,212, and rainbow trout (*Oncorhynchus mykiss*) has 287,967 ESTs. The top 98 large-scale projects have EST datasets measured in the hundreds of thousands. With such an enormous number of sequences, it becomes obvious that data curation and analysis is neither practical nor desirable without the use of computers.

Bioinformatics is the general term given to the use of data/computer information systems to manage and analyze biological data. The explosion of biological sequence data in the last decade necessitates the use of computers to store and process these results.

A common strategy for storing large amounts of data is through the use of relational databases. Relational databases provide a way to store large amounts of data from disparate sources. Once stored, relationships between data sources can be defined. All information is arranged in organized tables, and if designed to represent the data correctly, will optimize data access. Scripting languages such as PERL and Python can be used to automate data access requests, as well as to perform high-throughput data analyses.

There are a number of dynamic processes that shape and remodel genomes. The roles of mutations, mobile elements, segmental and genome duplications will be discussed.

Mutations

Genomes are in a constant state of change, giving rise to diversity. Given that DNA is constantly replicating and repairing itself due to growth cycles, it is not difficult to imagine that mistakes are a possible occurrence. Errors during these cycles can give rise to the incorporation of different sequences, which are called mutations. In molecular evolution, mutations are only relevant if they can be passed on to subsequent generations. Therefore, only non-somatic cells are usually discussed when studying inherited mutations. Plants are a notable exception, as they do not require sexual means for reproduction.

Point mutations affect only a single nucleotide, while segmental mutations affect two or more adjacent nucleotides of a chromosomal region. There are different types of change a mutation is responsible for. Substitution simply replaces a nucleotide with a different one. If a protein-coding region experiences a substitution mutation, its translated product may be affected. Since most codons are degenerate, there is the

possibility that there will be no changes to the product of translation if a single nucleotide is changed. In this case, this is a synonymous mutation. However, when the amino acid a codon encodes for is changed as a result of the substitution mutation, it is known as a nonsynonymous mutation.

Homologous recombination is another mechanism by which mutation occurs. It begins by the formation of an intermediate structure between four strands of DNA, known as a Holliday junction. This structure, once resolved, results in the formation of mismatched double-stranded DNA known as heteroduplexes. Enzymes in the cell will recognize these mismatches and perform excision followed by strand repair. Three outcomes are possible depending on which strand is chosen for excision and subsequent repair. There can be 1) no recombination at all, 2) crossing-over (an equal exchange of homologous sequence), or 3) gene conversion (unequal sequence exchange, resulting in the loss of one of the sequence variants).

Mutations that involve either insertion or deletion mutations are referred to as indels. Indels are of particular concern when they take place in coding regions, as they may cause a frameshift, altering the composition of the amino acid residues and the function of the gene product. Whether an insertion or deletion has occurred, the mechanisms by which they are possible are known. Unequal crossing over results in the deletion of one chromosomal segment, with reciprocal insertion in the other. Intrastrand deletions result when repeated sequences on the same chromatid, pair with each other in the same orientation resulting in a segmental deletion on the chromatid while producing an extrachromosomal element. This mechanism is often used by transposable elements and serves to reduce the number of tandem repeats. Replication slippage is a common event

and occurs in regions of contiguous short repeats, resulting in either the deletion or insertion of a repeat unit. Finally, inversion mutations usually occur in long stretches of DNA and are a result of chromosomal breakage and rejoining, or crossing over between homologous sequence segments that are oriented in opposite directions.

Mobile elements

Variation in genome sizes among similar species has been shown to be attributed to nongenic repetitive sequences (Graur and Li 2000). They occur many times in a genome, either in a localized tandem array or in a dispersed manner. The sequences of the tandem repeats are very uniform and upon genomic DNA fractionation and separation by density, show up as thick bands that are heavier or lighter than non-repetitive sequence. The DNA contained in these bands are called satellites, and can be either extremely G+C- or A+T-rich. Satellite DNA makes up a significant part of most genomes, however they are devoid of any known function. In almost all species, these localized tandemly repeated sequences have not been shown to either increase or decrease the fitness of an individual, and therefore are not maintained by natural selection forces (Graur and Li 2000).

Dispersed repeats are present as simple tandem or interspersed and occur in introns, regions that flank genes, intergenic, and nongenic regions. Simple tandem repeats are further classified according to repetition size. Satellites can be up to 2,000 bp in length, minisatellites are 9-100 bp, short tandem repeats are 3-5 bp, and microsatellites are only 1-2 bp. The most common microsatellite in humans is CA, of which there are roughly 50,000 copies (Hudson et al. 1992). Interspersed (dispersed) repeats also contain different categories. Additionally in humans, there are short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs).

The idea that noncoding DNA elements are able to self-replicate or expand to a point where they will negatively affect host fitness has long been debated. The term 'selfish-DNA' is used to describe the idea of these proliferating sequence elements that can insert themselves randomly into host genomes (Doolittle and Sapienza 1980; Orgel and Crick 1980). Frameshifts, truncations through the interruption of coding sequence, up or downregulation of gene expression from regulatory region insertions, and even chromosomal rearrangements are some of the 'selfish' acts that these mobile elements can carry out. Retrotransposons and transposons represent a large proportion of mobile elements that exhibit this selfish quality. In fact, mobile elements represent nearly half of the human genome (Lander et al. 2001). Therefore, these elements represent a significant topic in molecular biology that must be understood.

Retrotransposons are widely classified according to their physical attributes. Those that contain long terminal repeats (LTR), and those that do not (non-LTR). These classes of retrotransposons have common mechanisms in their mobility in that they both utilize an RNA intermediate, as well as leaving the parental copy intact. Non-LTR retrotransposons use a process known as target-primed reverse transcription. This whole process is rather inaccurate, leading to potential errors in the flanking regions of the insertion, or truncation. As a result of these errors, many retrotransposon events produce dead-on-arrival (nonautomomous) elements. It is interesting to note that these incapacitated retrotransposon elements can still mobilize, by hijacking other independent protein-producing elements, once they are transcribed.

Long terminal repeats flank LTR retrotransposons. These flanking sequences have a key role in proliferation, and are similar to the mobility mechanisms of retroviruses. The

mechanisms with which new LTR elements are produced ensure that the flanking LTRs are 100% identical to their parental LTRs (baring any evolutionary changes). The reproducibility of LTRs is important because it allows for molecular evolution studies of these elements in a given host. LTRs are expected to diverge neutrally because they do not possess any mechanisms for maintaining their homogeneity after host integration. Therefore, the individual LTRs are expected to diverge at the mutation rate. These assumptions mean that the age of insertion can be approximated based on the magnitude of divergence between its two LTRs (Lynch 2007).

The other major group of mobile elements are classified as transposons. Unlike retrotransposons, they do not depend on RNA as an intermediate for insertions. Rather, their movement depends on excision of genomic DNA. Their copy numbers are still able to increase despite this excision step, because double-strand breaks are repaired. This repair is usually carried out through homologous recombination using a sister chromatid, leaving the original transposon and either a single or double daughter chromosome inheriting the new element. Evidence suggests that transposon activity is linked to large-scale genomic reorganization events. A recent study has been able to support this correlation by showing that bursts of transposition replication activity coincided with Salmoninae speciation events (de Boer et al. 2007).

To maintain a presence, mobile elements must continue to protect active autonomous copies. These elements are prone to mutations that can render them inactive. New copies must retain autonomous functionality yet not negatively impact host fitness. To adversely affect its host would result in the mobile elements being subjected to the effects of purifying selection. This type of selection would be avoided to ensure the survival of

future generations of elements. Few studies exist to quantify the rates of insertions of mobile elements and subsequent fitness effects. Of the ones that do exist using a *Drosophila melanogaster* model (Maside et al. 2000; Maside et al. 2001), it was determined that rates of excision (1.3×10^{-4} per element per generation) averaged two orders of magnitude less than the rate of insertion (3.23×10^{-6} per element per generation). Insertions run the risk of being lethal to host fitness. Of the hosts that survive element insertion events, it was determined that the insertions reduced host fitness by as much as 0.5%-1.5%. Insertion is clearly the driving force behind mobile element activity, which in turn drives genome expansion. While the purpose of such activity is not definitively understood, host and element have evolved in such a manner as to accommodate this situation.

It is apparent that autonomous mobile elements, segmental, and genome duplications all serve to dramatically alter a genome. Contrary to popular belief, genome size does not necessarily reflect a higher order of organism. While it can be generalized that the increased complexity of an organism is correlated with a larger genome, this observation is not always correct. The story of a genome is the result of many events. It is known that mobile elements are 'selfish' and they increase in copy number, even at the possible expense of host fitness. While often serving no obvious phenotypic role, mobile elements have the immediate effect of increasing genome size, the consequence of which is not fully understood.

Segmental duplications

Segmental duplications are essentially isolated to small chromosomal regions, and arise through multiple pathways (Lynch 2007). Many newly-arisen gene duplicates are

arranged in tandem (tail-to-head or inverted tail-to-tail/head-to-head), suggesting that they arose from local chromosomal events such as replication slippage or nonhomologous unequal crossing-over. Segmental duplications can also occur through a transcription event. Often resulting in non-functional gene duplications, the sloppy transcription of non-LTR retrotransposons can lead to downstream gene replication. These transcription products can then be reinserted into the genome. Key regulatory or coding regions must be incorporated in the transcripts to allow these duplicates to be functional. If these inserted regions are incorporated next to regulatory regions, there is a chance that new expression patterns can arise (Lynch 2007).

It has been shown that DNA fragments can be captured and inserted into double-strand breaks during the repair process (Ricchetti et al. 1999; Yu and Gabriel 1999; Lin and Waldman 2001; Lin and Waldman 2001). Common sources of DNA fragments are mitochondrial DNA or retrotransposon-derived cDNA. Lastly, double-strand breaks have protruding ends that may invade largely non-homologous sites, using short regions of homology. The invaded chromosome will serve as a template, allowing the broken strand to extend. The extension is followed by a reattachment of the two free ends (Gorbunova and Levy 1997). This type of duplication event has an equal chance of containing an insertion containing functional or non-functional genes.

Regardless of the mechanism of segmental duplication, it has been shown that the majority is gene-sized. One such study has been carried out in *Caenorhabditis elegans*, where spans of duplication were identified using existing genes (Katju and Lynch 2003; Katju and Lynch 2006; Thomas 2006). A distribution of these duplication-span lengths showed a highly L-shaped curve. The length of these spans had an average of 1.4 kb.

The average length of *C. elegans* coding regions is about 1.9 kb. Analysis done on the size distribution of duplication-spans in other species (Fischer et al. 2001; Bensasson et al. 2003; Thomas et al. 2004; Zhang et al. 2005) exhibits the same L-shaped length distribution of the *C. elegans* genome, and supports the idea that the majority of segmental duplications are in fact gene-sized.

Genome duplications

While a segmental duplication is a common event, a whole genome duplication (WGD) is rare (Lynch 2007). However, a WGD represents the most dramatic source of potential genetic variation as this large-scale event effectively doubles all existing genes. Over time, duplicate genes can mutate. In the case of a duplicated gene set (e.g., 2N -> 4N), there is the possibility of more tolerance towards mutations. For example, if one gene experiences mutation and loses its functionality, the remaining duplicate gene could compensate for the loss of function. If none of the genes change, the duplicated gene set may then serve to alter an existing function through a dosage affect. There is also a small possibility that a gene can experience a mutation in its regulatory region so as to provide specialization to an existing function, or provide a totally novel function. Compared to plants that can produce offspring through self-fertilization, vertebrates have less of a tolerance towards WGDs. Vertebrates must be able to produce viable gametes for sexual reproduction. Zygote viability is not very tolerant of polyploidy; a triploid offspring (even if able to survive to maturity) produced from the mating of a normal diploid and a tetraploid will experience problems during meiosis (Lynch 2007). Therefore, it is rare to observe many WGD events in vertebrates, as the survival rate for offspring is low.

WGD events are recognized to have occurred throughout evolutionary history (Figure 5). Genome duplication arises through two main mechanisms. Endogenous genome duplications, or autopolyploidy, result in all alleles at a given locus arising from the same species. Allopolyploidy arises though species hybridization events. In either case, several factors make polyploidization events difficult to analyze. First, the majority of duplicated genes are eventually lost, and in the case of an ancient polyploidization event, large gaps would appear in areas that were initially continuous spans of genes or entire chromosomes. Second, chromosomal rearrangements such as inversions and translocations, combined with ongoing segmental duplications will further obscure gene copy numbers. It is interesting to note that genes produced via polyploidization have a longer half-life than those produced by segmental duplications. This longevity may be explained by stoichiometrics (Lynch 2007). Duplicated genomes have gene expression levels that remain in the same stoichiometric balance with all other interacting genes. The maintenance of this dosage balance may be considered evolutionarily favourable. A single segmental duplication on the other hand, will most likely only involve a single gene. This increase in gene function may throw off the established stoichiometric balance of its interacting genes and hence be unfavourable and selected against.

Hox cluster analysis is one example in which the signatures of polyploidization have been studied in ray-finned fishes. Hox genes are responsible for segmental placement of the body plan during embryonic development. It was discovered that zebrafish have seven Hox clusters, while only four are found in tetrapods (Amores et al. 1998). Detailed studies have suggested that this polyploidization event occurred prior to branching off of the ray-finned fish lineage (Taylor et al. 2001) (Figure 5 – Diamond 3). The evidence

would suggest that the polyploidization event was ancient; a single Hox cluster was eventually lost following the polyploidization event.

Relative to evolutionary time, genes are rapidly lost through non-functionalization following a duplication event [1-1/(2N)] (Kimura and Ohta 1969). However, gene duplicate survival requires its fixation and its subsequent positive selection. The mechanism for the preservation of gene duplicates and their specialization is fundamental to the understanding of molecular evolution and is a subject that will be addressed in my work. The most common gene preservation mechanism may be neofunctionalization. In this model, one copy of a duplicated gene receives a favourable mutation. This mutation results in a novel function. Either the ancestral or the newly duplicated gene could receive this mutation. The reasoning in this model is that gene redundancy allows one copy to be released from selective constraints. In these conditions, one copy can more readily undergo mutational changes (Ohno 1970). There are other variations to this model, but the implications are the same - the neofunctionalization model is completely driven by the idea of positive selection, and disregards other types of mutational possibilities.

The neofunctionalization model states that the majority of gene duplicates become non-functional relatively quickly. However, this model does not adequately explain why there are still a large number of gene duplicates retained in descendants of ancient polyploids. For example, salmonids have retained roughly 50% of their genes over the course of 100 million years (Allendorf et al. 1975). If neofunctionalization were the only gene preservation model at work here, one would not expect to see such a high retention rate of duplicates. As well, one would expect to see large numbers of new gene

functions, which one does not. It turns out that neofunctionalization is not the only model that can describe gene preservation in eukaryotic multicellular species.

Genes in multicellular eukaryotes are often comprised of one or more regulatory regions. The purpose of these regulatory domains is to interact with and control the coding regions of a gene. More specifically, these regulatory regions may control tissue or developmental stage specific expression of the gene. Under the Duplication-Degeneration-Complementation (DDC) model proposed by Force et al. (1999), degenerative mutations act to partition ancestral gene functions as a result of complementary loss-of-function in gene duplicates. This type of complementary loss is termed subfunctionalization, and opsins are an example where duplication followed by functional diversity has acted to preserve the original ancestral gene function (Briscoe 2001; Spaethe and Briscoe 2004). Unlike neofunctionalization, the DDC model is driven by degenerative mutations, rather than by favourable mutations. Moreover, this model explains how duplicate genes could be preserved more often in small populations and more commonly when they are tandemly linked.

Neofunctionalization and DDC are models that help to explain the fates of gene duplicates after a duplication event. Evidence suggests that WGD events are the major trigger for genetic diversity and speciation (Ohno 1970; Taylor and Raes 2004; Gerstein and Otto 2009). Recent evidence supports a third WGD (3R) occurring 320-400 million years ago in teleosts (Sato and Nishida 2010). Further studies in teleosts have provided strong evidence that reciprocal paralog loss following a WGD may have contributed to the burst of speciation in teleosts (Semon and Wolfe 2007). There have even been attempts to generalize rules of how eukaryotic genomes react to a WGD (Hufton and

Panopoulou 2009), but such hypothesizing attempts are difficult and species' responses range from increased genome restructuring to stasis. In other words, WGD reaction patterns are highly varied.

Speciation can also occur without WGD events. Enough genetic diversity, arising from evolution over millions of years, can be present in a species to give rise to distinct forms. It has been suggested that if variation between similar populations is present at a significant level (i.e. greater than 3% at the gene level) molecular biologists consider the two populations to be distinct species (Hebert et al. 2003). A recent study by Yazawa et al. (2008) compared salmon lice (*Lepeophtheirus salmonis*) from the Pacific and Atlantic Oceans. It is estimated that Atlantic salmon lice were probably introduced into the Pacific Ocean roughly 5 million years ago. Sequence comparisons of the two forms show an average nuclear gene difference of 3.2%, and 7.1% difference at the mitochondrial level. This evidence supports the existence of two distinct forms, arising over millions of years, from a single ancestral species.

As a major driver behind genetic variability and speciation, the effects of a WGD are of particular interest to molecular evolutionary studies. Although these WGD events cannot be observed directly, by comparing transcriptome or genome data from extant species, evolutionary trees can be constructed that indicate species' relationships to each other and at the same time hypothesize common ancestors and significant evolutionary events.

Implications of genome duplication

There has been a long history of discussion regarding genome duplications and their evolutionary implications. As far back as 1911, the study of maize by Kuwada led to the proposal of the species being a result of an ancient tetraploid (Taylor and Raes 2004).

Susumu Ohno proposed in 1970 in his book, *Evolution by Gene Duplication*, that major evolutionary transitions were driven by genome duplication events (e.g., not by point mutations). The significance of gene duplication and their associated mechanisms have been studied since that time. The ideas from studies on this topic strongly suggest that duplication is the overwhelming manner in which major evolutionary events, such as speciation or morphological changes, could arise.

It is hypothesized that vertebrates arose from an invertebrate ancestor and that this event was a result of one or more genome duplication events. The idea of two genome duplication events giving rise to all vertebrates was popularized by Ohno in 1968 (Figure 5 – Diamond 1, 2) (Ohno et al. 1968). Ohno's hypothesis has been refined in recent years, to become the 2R (also known as the 'one to four rule') hypothesis. The name of this hypothesis has been derived from the 2 rounds of proposed genome duplications, even though the actual number of duplications is still being disputed. Ohno also found evidence for a salmonid tetraploidization based on his observations of DNA content (Figure 5 – Diamond 4) (Ohno et al. 1968). It is now more accurately known that salmonids recently underwent a WGD in the last 25-100 million years (Allendorf and Thorgaard 1984), and are currently in the process of reverting back to a stable diploid state (Danzmann et al. 2008). Comparisons of the *Tetraodon* and human genomes show strong evidence of a genome duplication in the teleost fish lineage (Taylor et al. 2001; Jaillon et al. 2004) (Figure 5 – Diamond 3). These observations would suggest that genome duplications have the ability to give rise to bursts of speciation.

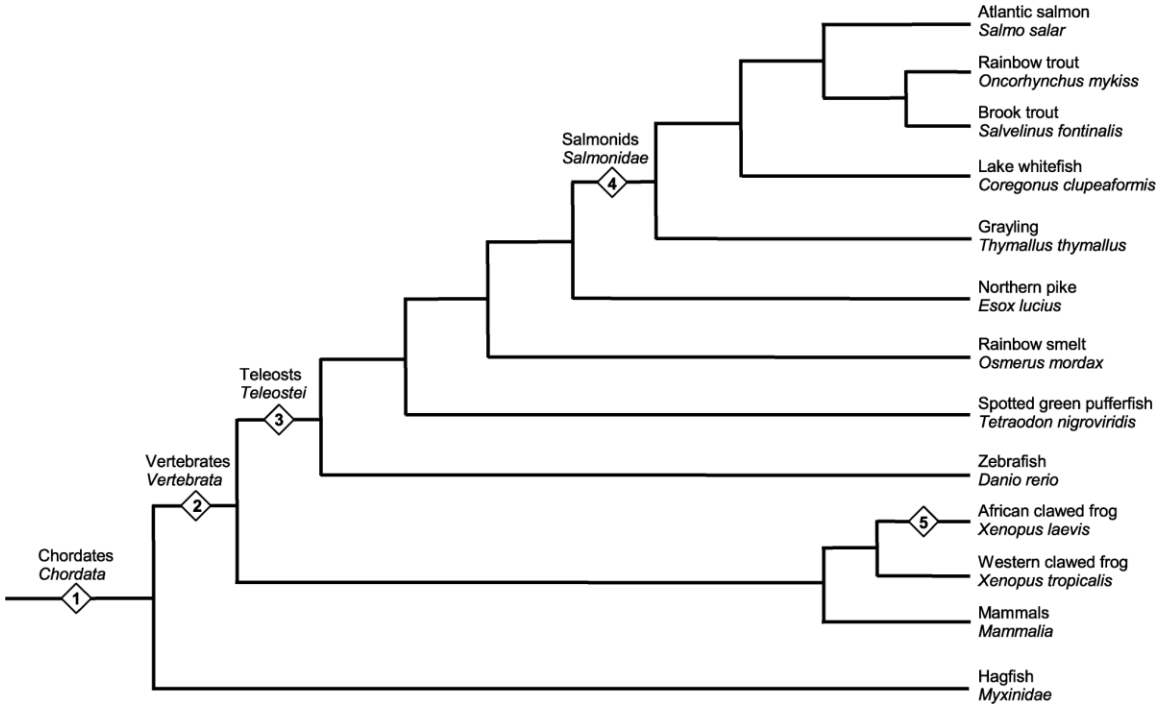

Figure 5. Molecular evolutionary tree with 5 proposed genome duplications. Diamonds 1 and 2 represent duplication events from Ohno's 2R hypothesis. Diamond 3 represents the ancient teleost WGD 320-400 MYA. Diamond 4 represents the salmonid WGD 25-100 MYA. Diamond 5 represents the recent *Xenopus laevis* duplication (Bisbee et al. 1977). Data modified from: (Brown 2008; Koop et al. 2008; Leong et al. 2010)

Atlantic salmon as a target for transcriptome sequencing

The consortium for genome research on all salmonids (cGRASP) provided a unique opportunity to study fish genetics. In 2006, the three-year cGRASP project began with funding from Genome Canada. The project's scientific mandates focused on Atlantic salmon (*Salmo salar*) and included three areas: coordinating the linkage map with the physical map, locating genes on the physical map, and studying gene expression at the transcriptional level. Information to fulfill these mandates needed to be obtained as part of the project. Teleosts, of which Atlantic salmon belongs to, represent an extremely diverse group of species. They make up roughly half of all vertebrates, the study of which would provide much needed insight into such a wide-range of species.

Atlantic salmon was chosen out of the sixty-six species in the Salmonidae family in particular because it is commonly used as an important sentinel species to monitor the health of aquatic environments. In Canada, conservation of many wild stocks are carefully monitored, as overfishing and declining stocks are prevalent issues. Used for aquaculture worldwide, Atlantic salmon is of particular economical importance to Canada, as well as to Norway and Chile. In addition, this species can act as a genomic model for other salmonid species such as trout or Arctic char (Thorgaard et al. 2002). Thus, through a series of funding opportunities, Atlantic salmon became the focus of the cGRASP joint international effort. In keeping with one of cGRASP's mandates, it was necessary to annotate and identify Atlantic salmon genes. Achieving the goals of the mandate presented a particular challenge, as Atlantic salmon contains many functional and non-function gene duplicates - remnants of gene duplication, and repeated non-coding sequence. Overall, there are challenges in discerning between alleles, paralogs,

gene duplicates, and even sequencing errors in functional genes. These challenges once overcome, will form the foundation of future salmonid research endeavours.

The areas of research that will be impacted with the work presented here are multi-faceted, with both short- and long-term uses with respect to their potential applications. In other words, while the results from this work can be used immediately, they also serve as the building blocks for future research. In the broadest sense, a complete workflow for the *in silico* identification of full-length genes in a duplicated genome will be useful not only among salmonids, but across almost any eukaryotic species. Across salmonids, conservation studies benefit from the identification of genetic markers in Atlantic salmon. These markers, once determined, can be used to correlate unique populations with geography or to track how different populations have been interacting over time. As an increasingly popular tool to aid biological studies, gene-based microarray design is greatly assisted by full-length gene information. From this information, short probes can be designed with a desired level of species or of gene region specificity. With these probes, microarray experiments become vital tools for efficient high-throughput gene expression analyses. Experiments can be performed to examine how an organism reacts to various environmental conditions. These conditions can involve the effects of pollution, drugs, nutrition, stress, parasites, and even migration patterns. The results from these microarray experiments are further optimized if there is reliable annotation from full-length genes. Whole genome sequencing (WGS) clearly benefits from EST data, helping to define coding regions as well as possible regulatory regions in the genome. A near-complete transcriptome serves as scaffolding to compare to long genomic segments; a transcript will match against exons, 5' untranslated region (UTR), and 3' UTR, thus

providing a method by which a gene's genomic structure can be verified. Signatures of evolution can be studied once genes are identified. Evidence that *Salmo salar* is in the process of reverting back to a stable diploid state after a recent whole genome duplication is of particular interest to molecular evolution studies. Comparing genes from species in which a WGD has occurred with a non-duplicated sister family enables evidence to be gathered about rates and patterns of evolution. The fate of gene duplicates in this species must also be studied so that the role of WGD in genetic variation can be better understood. From these examples, it is clear that a detailed characterization of *Salmo salar*'s transcriptome will serve as a significant foundation from which many other meaningful analyses will be possible.

The main goal of this thesis is to be able to identify the majority of full-length genes in Atlantic salmon. One method of accomplishing this goal involves the gathering and analysis of transcript data. Transcriptome sequencing requires the initial collection of raw sequence data. In the case of the cGRASP project, this step involved collecting raw EST data. However, this data alone is of incomplete biological value. The short EST reads, sometimes truncated, must be correctly assembled into contigs that can represent a single transcript. These contigs are the longest possible representations of a cell's transcriptome. Further in-depth analysis is required to determine what gene a contig represents. It must also be decided if the gene has been accurately sequenced or assembled from the EST data.

At the start of my MSc., the data for approximately 70,000 raw ESTs had been obtained. However, the existing computational tools for analysis and visualization were not suitable for dealing with such large datasets especially from duplicated genomes. My

work with Atlantic salmon focuses on the analysis of existing raw EST data, over 400,000 new EST sequences, plus additional full-length DNA sequence data. Such large disparate datasets require methods of automated storage and analysis. Due to the WGD, the possibility of misrepresenting genes through sequence assembly must be mitigated. In generating sequence data for accurate full-length gene identification, the optimal experimental protocols must be chosen in order to produce meaningful data to work with. In turn, it is imperative that this data be analyzed with the proper Bioinformatics algorithms. To best represent the final results from the data, it is necessary to implement an easy to use web-based system of visualization and collaboration for the full-length gene identification. To further extend the *Salmo salar* transcriptome work, my thesis touches on the sequencing and full-length gene identification of the closest non-WGD species, northern pike (*Esox lucius*). Using genes from *Salmo salar* and *Esox lucius,* a non-duplicated sister family, the topic of molecular evolutionary patterns was investigated.

Only by understanding the fundamentals of a species at a genetic level can one begin to fully understand its complex interactions with its ecosystem. Once basic transcriptome knowledge has been established, it will be possible to begin asking these more profound questions. The answers to these questions will no doubt one day aid policy makers in making informed decisions that will have implications for a species' continued well-being. In the case of a harvested natural resource such as fish, continued sustainability should be made a primary concern. Improper knowledge regarding the handling of such resources made today could have the potential to detrimentally affect its continued survival. Therefore, if the necessary genetic data and functional genomic tools exist, it is

the responsibility of these policy makers to gather accurate information on areas that will ultimately affect long-term species health.

The goal of the work presented here is to establish basic molecular knowledge about salmonids using Atlantic salmon as the model organism. In particular, the identity of its genes is analyzed by gathering and assembling data from EST and full-length cDNA sequences. These gene constructs should faithfully represent a grouping of partially overlapping transcribed sequences from one gene at a single locus. In individuals with a duplicated genome, there is a significant chance that contigs represent a combination of closely related genes. It is the hope of this work to illustrate that by using *in silico* analysis of raw EST data, full-length genes in a duplicated salmonid genome can be faithfully represented. Using a series of novel Bioinformatics approaches, I have attempted to produce an accurate identification of Atlantic salmon genes.

Chapter 2

Gene Identification and Analysis of Salmonid Expressed Sequence Tags

Koop, B. F., von Schalburg, K. R., Leong, J., Walker, N., Lieph, R., Cooper, G. A., Robb, A., Beetz-Sargent, M., Holt, R. A., Moore, R., Brahmbhatt, S., Rosner, J., Rexroad, C. E., McGowan, C. R. and Davidson, W. S. (2008), 'A salmonid EST genomic study: genes, duplications, phylogeny and microarrays' *BMC Genomics* **9**, 16.

Summary

Salmonids are of interest because of their relatively recent genome duplication, and their extensive use in wild fisheries and aquaculture. A comprehensive gene list and a comparison of genes in some of the different species provide valuable genomic information for one of the most widely studied groups of fish.

298,304 ESTs from Atlantic salmon (69% of the total), 11,664 chinook, 10,813 sockeye, 10,051 brook trout, 10,975 grayling, 8,630 lake whitefish, and 3,624 northern pike ESTs were obtained in this study and have been deposited into the public databases. Contigs were built and putative full-length Atlantic salmon clones have been identified. A database containing ESTs, assemblies, consensus sequences, open reading frames, gene predictions and putative annotation is available. The overall similarity between Atlantic salmon ESTs and those of rainbow trout, chinook, sockeye, brook trout, grayling, lake whitefish, northern pike and rainbow smelt is 93.4, 94.2, 94.6, 94.4, 92.5, 91.7, 89.6, and 86.2% respectively. An analysis of 78 transcript sets show *Salmo* as a sister group to *Oncorhynchus* and *Salvelinus* within Salmoninae, and Thymallinae as a sister group to Salmoninae and Coregoninae within Salmonidae. Extensive gene duplication is consistent with a genome duplication in the common ancestor of salmonids. Using all of the available EST data, a new expanded salmonid cDNA

microarray of 32,000 features was created. Cross-species hybridizations to this cDNA microarray indicate that this resource will be useful for studies of all 68 salmonid species.

An extensive collection and analysis of salmonid RNA putative transcripts indicate that Pacific salmon, Atlantic salmon and charr are 94–96% similar while the more distant whitefish, grayling, pike and smelt are 93, 92, 89 and 86% similar to salmon. The salmonid transcriptome reveals a complex history of gene duplication that is consistent with an ancestral salmonid genome duplication hypothesis. Genome resources, including a new 32 K microarray, provide valuable new tools to study salmonids.

Introduction

Extensive knowledge of trout and salmon is a result of their widespread use in scientific research, as an environmental sentinel species and as a food and sport fish. Perhaps more is known about the physiology, ecology, genetics, behavior and biology of salmonids than any other fish group (Thorgaard et al. 2002). This background provides a wealth of data from an economically important and phylogenetically distinct group of fish that can help guide, and benefit from, new genomic studies.

The Salmonidae family includes: whitefish and ciscos (subfamily Coregoninae); graylings (Thymallinae); trout, salmon and charr (Salmoninae) (Nelson 2006). Salmonids are classified into nine genera and sixty-six species. They are native of the cooler climates of the Northern Hemisphere, but have been widely introduced around the world. Salmonids belong to a basal teleost Protacanthopterygii suborder (smelt, pike and salmon) group, which has been separated from other well studied euteleost lineages such as Ostariophysi (zebrafish, catfish, flathead minnow, etc.), and Acanthopterygii (cod, cichlids, fugu, sticklebacks, rockfish) for 217–290 MY (Ishiguro et al. 2003; Hoegg and Myer 2005; Nelson 2006; Steinke et al. 2006).

The common ancestor of salmonids is purported to have experienced a whole genome duplication event between 25 and 100 MYA (Ohno 1970; Allendorf and Thorgaard 1984). Extant salmonids are considered pseudo-tetraploid as they are in the later stages of reverting to a stable diploid state. Evidence for the ancestral salmonid autotetraploid genome duplication includes: multivalent chromosome formation during male meiosis and evidence for tetrasomic segregation at some loci (Allendorf and Thorgaard 1984); one of the larger euteleost genome sizes (3–4.5 pg) with double that of sister groups

Esociformes (0.8–1.8 pg, pike) and Osmeriformes (0.7 pg, smelt) (Gregory 2002); homeologous chromosomal segments based on recent genetic maps and comparative studies using microsatellite markers, and duplicated gene family studies such as Hox, Major Histocompatibility complex (MH), growth hormone, and nineteen allozymes (Allendorf and Thorgaard 1984; McKay et al. 2004; Moghadam et al. 2005; Danzmann et al. 2006; Lukacs et al. 2007).

The genome duplication in salmonids is the most recent genome duplication in this lineage. There are now a number of studies and good evidence, primarily from sequenced zebrafish and pufferfish genome sequences, for tetraploidization/rediploidization early in the ray-finned fish lineage (350–400 MYA) (Vandepoele et al. 2004; Panopoulou and Poustka 2005; Volff 2005; Blomme et al. 2006). Several of these studies have suggested that the ancestral fish duplication, in addition to the two ancestral vertebrate genome duplications, are part of the reason why ray-finned fishes make up nearly half of all extant vertebrates species and exhibit tremendous biodiversity affecting their morphology, ecology, behaviour and evolution.

Vertebrate species diversity and body plan diversity have commonly been linked to genome duplications, although there is some debate on how well we can draw these conclusions based on the very old genome duplications commonly studied. Mechanistically, how a genome reorganizes itself to cope with duplicated chromosomes, gene dosage effects, and the role of gene duplications for evolution and adaptation are long-standing issues in biology that remain unresolved (Allendorf and Thorgaard 1984; Force et al. 1999; Vandepoele et al. 2004; Panopoulou and Poustka 2005; Volff 2005; Blomme et al. 2006). The number and diversity of salmonid species, and their relatively

recent genome duplication, make salmonids ideal for examining recent events that could have played such a pivotal role in generating gene diversity and species diversity found in modern vertebrates.

The genomics resources of salmonids are being rapidly expanded through a few large-scale genomics programs (Rexroad et al. 2003; Rise et al. 2004; Ng et al. 2005; Govoroun et al. 2006; Adzhubei et al. 2007; von Schalburg et al. 2008). Here we identify 354,061 new ESTs from Atlantic salmon and several other salmonid and related species in order to obtain a comprehensive view of the salmonid transcriptome, identify species relationships, identify gene duplications and introduce a new 32 K microarray tool for transcriptome analysis.

Methods and Materials

Tissues, RNA, Aquaculture and Sampling

Salmo salar (McConnell strain), *Oncorhynchus tshawytscha* and *Oncorhynchus nerka* tissues were obtained from the Department of Fisheries and Oceans (Robert Devlin, WestVan Lab., West Vancouver, British Columbia). *Salvelinus fontinalis* and *Coregonus clupeaformis* tissues were obtained from Louis Bernatchez (Laval University, Quebec). *S. salar* (Saint John River strain; brain, kidney and spleen) were obtained from Vanya Ewart (NRC Institute for Marine Biosciences, Nova Scotia). *Thymallus thymallus* brain, kidney and spleen tissues were obtained from Craig Primmer (University of Turku, Finland). *Esox lucius* were captured by gill net from Charlie Lake British Columbia. All fish were euthanized, followed by rapid dissection of tissues. Tissues were flash frozen in liquid nitrogen or dry ice and stored at -80°C until RNA extraction.

cDNA libraries

Total RNA or poly(A)+ RNA (FastTrack MAG kit; Invitrogen) was extracted from flash frozen tissues. *Salmo salar* and *Oncorhynchus tshawytshca* mixed tissue (spleen, head kidney, brain) libraries were directionally constructed in both pCMV Sport-6.1 (Research Genetics Inc.) and pAL-17.3 (Evrogen Co.). *S. salar* (normalized head kidney, thymus and thyroid), *Coregonus clupeaformis. Thymallus thymallus* and *Salvelinus fontinalis* libraries were constructed in pAL-17.3 (Evrogen). The *Oncorhynchus nerka* mixed tissue normalized library was also constructed in pCMV Sport-6.1 (ResGen). *S. salar* (mixed tissue St. John strain) and *Esox lucius* libraries were constructed in pDNR-Lib using Creator SMART cDNA library construction kits (Clontech). Insert sizes of

cDNA libraries were determined by visual comparison of clone restriction fragments with the DNA size markers *Hind*III (GibcoBRL) and 1 kb ladder (GibcoBRL).

Sequencing, Sequence Analysis, and Contig Assembly

Clone libraries were plated and robotically arrayed in 384-well plates. Glycerol stocks of overnight cultures were prepared in 384-well format (Rise et al. 2004). Plasmid DNAs were extracted and BigDye™ Terminator (ABI) cycle sequenced on ABI 3730 sequencers using conventional procedures and the following primers: 5'-T_{18}-3', M13 forward (5'-GTAAAACGACGGCCAGT-3'), M13 reverse (5'-AACAGCTATGACCAT-3' or 5'-CAGGAAACAGCTATGAC-3') and for the Evrogen libraries SP6WAN primer was used for the 3' end sequencing. Base-calling from chromatogram traces was performed using Phred (Ewing and Green 1998; Ewing et al. 1998). Vector, poly-A tails, and low quality regions were trimmed from EST sequences; sequences that had less than 100 good quality bases after trimming were discarded (Rise et al. 2004). Initial assembly of ESTs into contigs used PHRAP (Green), under stringent clustering parameters (minimum score: 100; repeat stringency: 0.99). Contig consensus sequences and singleton sequences were aligned with non-redundant GenBank nucleotide and amino acid sequence databases (SwissProt, PBL, CDD, and UniRef90) using BLASTN or BLASTX (Boguski et al. 1993; Altschul et al. 1997; Camon et al. 2004). Sequence databases, assemblies, consensus sequences, tools such as BLAST and RepeatMasker (Smit et al. 1996), and sequence and consensus annotations are freely available from the author and from the (GRASP website).

The number of *Salmo salar* contigs was assessed using the PHRAP assembly program because of its ability to assemble very large numbers of ESTs in a single run, and its

integration with PHRED base quality scores on primary reads and subsequent consensus sequences. The CAP3 assembler (Huang and Madan 1999) was also used and similar results were obtained for smaller datasets. For this study, contig assembly employed a two-stage process. The first stage assembly used parameters 100 minscore and 0.99 repeat stringency to build contigs and consensus sequences that appeared to separate alleles of many transcripts. The second stage used the consensus sequences (with quality scores) from the first stage and parameters 96% repeat frequency and 300 minscore to build contigs and consensus sequences that appeared to combine some of the contigs that contained some base calling discrepancies, as well as what appeared to be alleles or very recently duplicated genes. Various parameters were tested and final parameters were chosen to minimize the number of contigs, where the number of contigs changed the least with respect to small changes in parameter values, and where distinct contigs appeared to have some biological significance (i.e., 99/100 appeared to separate many alleles and 96/300 as a second stage appeared to join some alleles and provided values that separated a clear majority of orthologous salmonid gene comparisons). With both sets of parameters, we were able to discriminate between similar sequences from different salmonid species. Sequences in contigs containing more than one polyA site were removed from the assemblies as they may represent chimeric clones.

Assemblies provide rough estimates of transcripts. Several algorithms have been examined and all have strengths and weaknesses. Examples of other assemblies include DFCI gene indexes (Computational Biology and Functional Genomics Laboratory ; Quackenbush et al. 2001) that estimate 83,554 TCs+singletons from 244,984 rainbow trout ESTs and 63,138 contigs from a partial 236,009 EST dataset from Atlantic salmon

(these assemblies are periodically updated). INRA (Govoroun et al. 2006) using CAP3 estimates 56,392 transcripts (contigs + singlets) from 326,719 rainbow trout ESTs and 45,349 contigs from a partial Atlantic salmon EST database. UniGene (NCBI Unigene database), from NCBI does not provide true assemblies and may cluster duplicated genes into single bins, which is problematic in salmonids. UniGene estimates approximately 30,000 and 25,000 UniGene sets in Atlantic salmon and rainbow trout respectively. While differences exist, the general number of estimated transcripts is similar. Problem areas that have been identified in assemblies tend to be associated with long transcripts, so these contigs will have to be treated carefully and perhaps manually edited. Assemblies are freely available from the author and the (GRASP website). As a caveat, because of the purported recent duplication of the salmonid genome and potential for miss-assembly of duplicated transcripts, these contigs have to be treated with caution.

Percent identity measures between contig consensus sequences from the various species were obtained from BLASTN alignments where a minimum length of 200 bp was observed. As in other distance measures, this finds the most similar sequence fragments and is biased high, particularly for more distant comparisons. A partial estimate of the impact on more distantly related sequence comparisons is the increased number of contigs for which no cross-species alignments were found and the reduction in average length of alignments. These values are provided in Table 3. However, the percent identity measure provides an estimate of observed similarity that is useful for evaluating potential cross-species DNA hybridizations in microarray experiments (see below).

Gene phylogenetic analysis

Contig sequences from *Salmo salar* (Atlantic salmon), *Oncorhynchus mykiss* (rainbow trout), *Osmerus mordax* (rainbow smelt), *Coregonus clupeaformis* (lake whitefish), *Salvelinus fontinalis* (brook trout), and *Thymallus thymallus* (grayling) (Table 2) were BLASTed against each other (evalue < 1e-35, hits > 100bp) and the results used to generate clusters of contigs. Bins of similar sequences, or clusters, were generated containing all contigs irrespective of species origin that had alignments with greater than 75% of the length of the shorter sequence and had greater than 70% identity in the overlapping regions (alignments consisted of ends-free alignment with scores of 2/-2/-5/-1 for match/mismatch/open gaps/extend gaps (Gusfield 1999). After the contigs had been grouped into clusters, the individual clusters were then further selected to only contain contigs that had mutually overlapping regions and all contig members were trimmed to the largest common alignment (same alignment parameters as above). A good alignment was considered to be greater than or equal to 300 bp in length with greater than 60% identity in the overlapping region. At this point, clusters that did not contain at least one sequence from each of the six target species were discarded. This resulted in a dataset of 78 clusters or gene sets. All gaps (and their corresponding positions in other sequences of the cluster) were removed, and the data within each gene set were bootstrapped 500 times. The PHYLIP package was used because it offers many different analysis methods, is freely available and is commonly used (Felsenstein 2004). Distance matrices were computed for each bootstrapped dataset within each cluster using the F84 model of nucleotide substitution and Gamma-distributed rates of variation across sites with a coefficient of variation of 0.5 (Felsenstein 2004). Neighbour-joining trees

were then computed from each set of distance matrices and the set of resulting bootstrapped trees was used to derive a 70%-majority consensus tree (Felsenstein 2004). The consensus trees were rooted with *Osmerus mordax*, and simplified by iteratively collapsing all pairs of leaf nodes having the same species and showing > = 98% similarity in the aligned portion of their sequences. Independently, maximum likelihood trees were generated for all 78 data sets using the default options with the Phylip program dnaml (transition/transversion ratio of 2.0, empirical base frequencies, constant rate variation among sites). A general evolutionary model was used for the 78 data sets because each set potentially consisted of a mixture of unidentified coding and non-coding data. All of the 78 ML trees were consistent with their 70%-consensus bootstrapped Neighbour-joining counterparts. EST accession numbers used to make contig consensus sequences, alignments and the 70% consensus trees are available or online at the (GRASP website).

Microarray Clone selection

Starting from the existing GRASP 16 K cDNA microarray (GRASP website), additional clones were selected for representation on the following basis: a) the contig (Table 2) includes at least one clone that is on hand; b) the contig is of high quality with few conflicting positions, few singleton positions, no interior singleton positions (potential chimeric sites) and there are at least two clones in the contig (from at least 2 plates, and preferably from at least 2 libraries); c) if the contig is singleton then it must have a good BlastX hit (e-value < 1e-8) or other indication of orientation (eg. consistent poly(A) tail information); d) contig must have < = 94% identity to another sequence on the chip (the existing 16 K plus any new contig; not counting rainbow trout orthologs); and e) no tRNA, ribosomal, or mitochondrial sequences. We chose clone representatives

within each contig based on: a) the reliability of the cDNA library and sequence; b) high similarity to consensus of contig (allow 20 bp at ends for poor trimming); c) reliable sequence from the 3'-end of contig and correct (3' -> 5') orientation; and d) ownership of clone.

Microarray fabrication

The initial clones were robotically rearrayed from daughter glycerol stock 384-well plates into 96-well plates prefilled with 8% glycerol in 2XYT + ampicillin with a MicroGrid II-610 (Biorobotics, Cambridge, UK), incubated overnight at 37°C, and checked for uniform optical density. Plasmid inserts were PCR-amplified in a MJ Tetrad PTC-205 thermocycler (Bio-Rad, Hercules, CA, USA) by using 1.0 μL overnight culture, 0.3 μM M13/pUC forward primer (5'-CCCAGTCACGACGTTGTAAAACG-3'), 0.3 μM M13/pUC reverse primer (5'-AGCGGATAACAATTTCACACAGG-3'), 2 mM MgCl2, 10 mM Tris-HCl, 50 mM KCl, 200 μM dNTPs, 1U AmpliTaq (Roche Diagnostics, NJ, USA), and nuclease-free H2O (Qiagen, Valencia, CA, USA) to 100 μL. PCR conditions were as follows: 2 min at 95°C denaturation; 35 cycles of 30 sec at 95°C, 45 sec at 59°C, and 4 min at 72°C; and 7 min at 72°C. Hotstar taq (Qiagen) was used to amplify additional inserts (clone set 2) with an initial denaturation of 15 mins. Amplicon specificity and yield was analyzed by capillary electrophoresis using the HT DNA SE 30 LabChip on Caliper AMS 90 system (Zymark-Caliper Life Sciences, MA, USA). PCR products were robotically cleaned (Qiagen) and consolidated into 384-well plates, lyophilized by speed-vac, and resuspended in 20 μL 3× SSC plus 1.0 M betaine. All cDNAs (average printing concentration of 165 ng/ul [original inserts] and 100 ng/ul [new inserts]) were printed as single spots on Erie Aminosilane slides (Erie, Portsmouth, N.H.,

USA) with a Genetix QArraymax microarray printer (Genetix, New Milton, Hampshire, UK) or MicroGridII-610 printer (Biorobotics, Cambridge, UK). All clones and controls were distributed randomly on the array. Genetix aQu 65 um quill pins or Biorobotics 10 k quill pins in a 48-pin tool were used to deposit < 1.0 nL (0.1 ng cDNA) per spot onto the slide. The resulting microarrays have a 4 × 12 subgrid layout with 699 spots per subgrid, each spot having diameter and pitch of 90–130 and 160–190 μm, respectively. A 280-bp GFP (green fluorescent protein) cDNA was amplified from a GFP clone (BD Biosciences, Mountain View, CA, USA) by using the primers (5'-GAAACATTCTTGGACACAAATTGG-3') and (5'-GCAGCTGTTACAAACTCAAGAAGG-3'), and printed in subgrid corners to assist in placing on the grid. The slides were crosslinked in a UV Stratalinker 2400 (Stratagene, La Jolla, CA, USA) at 300 mJ. One slide every 20 to 30 slides was hybridized with labeled random 9-mer oligonucleotide (SpotQC, Integrated DNA Technologies, Coraville, IA, USA) and scanned using GenePix 4200AL scanner (Molecular Devices, Sunnyvale, CA, USA). Presence/absence, shape, signal intensity vs. background, diameter and DNA binding site capability were measured for each spot on the slide using files generated by Imagene software (BioDiscovery Inc., El Secundo, CA, USA). Position and description of flagged spots (spots absent or thought to be unusable during post hybridization analysis), sub-grid defects and other noticed irregularities are recorded. Two PCR fragments from each plate were randomly selected and sequenced to ensure correct matches to the original clone sequence in the EST database. For controls, Stratagene SpotReport Alien cDNA Array Validation system PCR products (Cat # 252550) composed of 10 unique PCR products are spotted five times on the array.

Corresponding mRNA for these PCR products can be purchased from Stratagene. The alien mRNA spikes can be used to determine mRNA quality, cDNA synthesis efficiency, positive and negative hybridization control, normalization for dye differences and determination of hybridization consistency.

Microarray hybridizations

The microarray experiments were designed to comply with MIAME guidelines. To minimize technical variability, all targets were synthesized in one round and hybridization experiments were conducted on slides from a single batch. Each hybridization experiment included dye-flips to compensate for cyanine fluor effects. Total RNA samples were quantified and quality-checked by spectrophotometer and agarose gel, respectively. All hybridization experiments were performed using the SuperScript III Indirect cDNA Labeling System kit and following manufacturers instructions (Invitrogen). Briefly, total RNA was reverse transcribed using an anchored oligo $d(T)_{20}$ primer in cDNA synthesis reactions that incorporated aminoallyl- and aminohexyl-modified nucleotides. The modified cDNAs were then labeled with fluorescent Cy5 or Cy3 dye in reactions with the amino-functional groups in coupling buffer.

All microarrays were prepared for hybridization by washing 2 × 5 min in 0.1% SDS, washing 5 × 1 min in MilliQ H_2O, and drying by centrifugation (520 g for 5 min in 50 ml conical tube). All slides were prehybridized in 5 × SSC, 0.1% SDS, 3% BSA for 1.5 h at 49°C. Arrays were briefly washed 3 × 20 sec in MilliQ H_2O, then dried by centrifugation. A total of 200 ng of labeled cDNA with each fluor was applied to prewarmed microarrays in a formamide-based buffer (25% formamide, 4× SSC, 0.5%

SDS, 2× Denhardt's solution) 16 h at 49°C. The arrays were washed 1 × 10 min at 49°C (2× SSC, 0.1% SDS), and then 2 × 5 min in 2× SSC, 0.1% SDS, 2 × 5 min in 1× SSC and 4 × 5 min in 0.1× SSC at room temperature, then dried by centrifugation.

Microarray analyses

Fluorescent images of hybridized arrays were acquired immediately at 10 um resolution using ScanArray Express scanner (PerkinElmer). The Cy3 and Cy5 cyanine fluors were excited at 543 nm and 633 nm, respectively, at the same laser power (90%), with adjusted photomultiplier tube settings between slides to balance the Cy5 and Cy3 channels. Fluorescent intensity data was extracted from TIFF images using Imagene 5.6.1 software (Biodiscovery). Quality statistics were compiled in Excel from raw Imagene fluorescence intensity report files. The hybridization performance of labeled targets to salmonid features was assessed as a percentage of features bound from the numbers of AS and RT features passing a hybridization signal threshold. Signal threshold was defined by 2 standard deviations above the signal mean for the 3× SSC/betaine buffer spots. Outliers of buffer spots were removed based on the Median Absolute Deviation method (Hample 1974) whereby elements with a test statistic value greater than 5 were removed. No transformations or normalizations were performed on these data. Only features deemed present by Imagene 5.6.1 (excluding marginal and absent values) were used for analyses.

Results and Discussion

cDNA libraries

New, directionally cloned, mixed tissue (brain, kidney and spleen), normalized cDNA libraries were constructed for Atlantic salmon (*Salmo salar*; European McConnell, and Canadian, Saint John River strains), chinook salmon (*Oncorhynchus tschawytscha*), sockeye salmon (*Oncorhynchus nerka*), brook trout (*Salvelinus fontinalis*), lake whitefish (*Coregonus clupeaformis*), grayling (*Thymallus thymallus*), and northern pike (*Esox lucius*). Separate normalized libraries were constructed from *Salmo salar* thymus, thyroid, and head kidney tissues. In addition, one full-length, mixed tissue, large insert (> 2 kb), non-normalized library was constructed to identify longer gene transcripts. cDNA clones were isolated, purified and sequenced from the 5' and 3' ends. Clone numbers and insert sizes for the different libraries and species that were done as part of this study are listed in Table 1.

Table 1. Salmonid cDNA libraries, sequencing and assembly summary statistics for data provided in this study.

Species/Tissue/(library)	# clones [a]	Insert size [b]	# seq [c]	# contigs [d]	#. of singlets [e]	Max. contig [f]	Ave. contig [g]	% new (sp.) [h]
Salmo salar								
Thymus (evd)	31488	1.5	59264	23768	8685	66	2.3	16
Thyroid (eve)	30720	1.9	58700	28045	12378	37	2.1	15
Head kidney (evf)	31104	1.5	59541	28316	10832	30	2.1	16
Pyloric Caecum (pla, plb, plc, plna, plnb, pha, phc)	9584	0.9	13543	5691	2766	35	2.3	17
Brain, kidney, spleen (rgb2)	60288	1.6	97171	42562	26504	58	2.1	15
Brain, kidney, spleen (sjb)	5835	1.8	10085	6656	3541	8	1.5	18
Oncorhynchus tschawytscha								
Brain, kidney, spleen (rgd)	3840	2.1	5935	3941	2970	31	1.5	82
Brain, kidney, spleen (evc)	3744	1.8	5729	3841	2487	10	1.5	80
Oncorhynchus nerka								
Brain, kidney, spleen (rge)	7296	2.0	10813	6123	3924	173	1.8	98
Salvelinus fontinalis								
Brain, kidney, spleen (evi)	5376	1.4	10051	5424	1247	9	1.9	100
Coregonus clupeaformis								
Eye, kidney, spleen (evb)	4800	1.6	8630	5537	3359	12	1.6	93
Thymallus thymallus								
Brain, kidney, spleen (evl)	5760	1.5	10975	5926	1309	6	1.9	100
Esox lucius								
Brain, kidney, spleen (bkhp)	2304	0.9	3624	2420	1346	6	1.5	100

[a] number of clones from which at least one sequence (5' or 3') was obtained
[b] average EST fragment size cloned (kb), estimated from > 30 clone digests.
[c] number of 5' and 3' EST sequences obtained
[d] number of EST contigs (1st stage assembly) that includes singlets
[e] number of contigs containing a single sequence
[f] the size of the contig containing the largest number of sequences
[g] the average size of all contigs (includes singletons)
[h] percent of the putative transcripts that are unique to the species.

Transcript analysis: sequence and assembly

To obtain a comprehensive list of genes in salmonids, we used a strategy of deep 5' and 3' EST sequencing from a few high quality libraries. This approach complements previous studies, which examined more limited EST surveys of cDNA libraries from a large number of different tissues and developmental stages (Rexroad et al. 2003; Rise et al. 2004; Ng et al. 2005; Govoroun et al. 2006; Adzhubei et al. 2007). For Atlantic salmon, over 30,000 clones were sequenced from each of the thymus, thyroid, and head kidney tissue libraries. From previously described normalized libraries, (Rise et al. 2004) 9,584 additional clones were sequenced from the Atlantic salmon pyloric caecum tissue library and 60,288 additional clones were sequenced from a mixed tissue library (rgb2; Table 1). The total number of clones examined from the rgb2 library was 84,176 which yielded 127,660 sequence reads or 30% of the total Atlantic salmon EST database. Even with this deep sequencing, nearly 13% of the last 637 reads were novel (< 99% over 100 bp) and the maximum redundancy for a single transcript from the rgb2 library was 58 (Table 1).

The results of the assembly of 298,304 Atlantic salmon ESTs obtained in this study along with 138,325 ESTs from previous studies [(Rise et al. 2004; Adzhubei et al. 2007), GenBank] are shown in Table 2. Due to the complexities of the salmonid genome duplication and because it provides a stable, conservative starting point for all subsequent analyses, our analysis began with a first stage assembly using stringent parameters (PHRAP: 0.99 repeat stringency and 100 minscore). A second stage assembly (96% repeat stringency and 300 minscore) was implemented to combine some of contigs which may be alleles, or possibly very recent gene duplications (distinguishing among alleles,

minor assembly errors, miss-calls and very recent gene duplications, particularly in lower quality sequence regions is very difficult in the absence of genomic sequence data). In Atlantic salmon, 81,398 potential transcripts (2 stage assembly) were identified, of which 29,844 (37%) were similar (BLASTX, 1e-10) to annotated sequences in CDD or SwissProt protein databases. For comparison, an assembly of 246,704 ESTs from rainbow trout [(Rexroad et al. 2003; Govoroun et al. 2006), GenBank] resulted in 51,199 transcripts, of which 19,266 (38%) had BLASTX hits. Assembled contigs are available (GRASP website).

Table 2. Summary of salmonid ESTs and contig assemblies.

	Atlantic salmon	Rainbow trout	Chinook salmon	Sockeye salmon	Brook trout	Lake whitefish	Grayling	Northern pike	Rainbow smelt
# EST sequences [a]	436629	246704	14535	12056	10051	10842	10975	3624	36785
Assembly Stage1 [b]									
# contigs (2+) [c]	70,845	42423	2890	2480	4178	4464	4616	1074	9044
# singletons [d]	47,139	26935	6295	4118	1247	2510	1314	1346	7019
# transcripts [e]	117,984	69358	9185	6598	5425	6974	5930	2420	16063
Assembly Stage2 [f]									
# transcripts [g]	81398	51199	8517	6200	4946	6446	5408	2380	12159
# hits [h]	29844	19266	3684	3561	1838	2314	1780	198	6139
% with hits [i]	37	38	43	57	37	36	33	8	50

[a] number of EST sequences for all of the species including those in GenBank
[b] Assembly stage 1 refers to PHRAP assembly using parameters 0.99 repeat_frequency and 100 minscore
[c] number of contigs with 2 or more sequences
[d] number of contigs with 1 sequence
[e] total number of transcripts including singletons
[f] Assembly stage 2 refers to PHRAP assembly using parameters 0.96 repeat_frequency and 300 minscore
[g] the number of transcripts that result from a re-assembly of all stage 1 transcripts using PHRAP parameters 96 repeat_frequency and 300 minscore
[h] number of transcripts that have a BLASTX hit of < 1e-10 to SwissProt/CDD databases.
[i] percent of stage 2 assembled transcripts that have a BLASTX hit.

Transcript surveys of additional salmonid species included 4,800–7,500 clones sequenced from each of chinook salmon, sockeye salmon, brook trout, grayling and lake whitefish. 11,664 sequences were obtained from chinook salmon, 10,813 sequences from sockeye salmon, 10,051 sequences from brook trout, 10, 975 sequences from grayling and 8,630 sequences from lake whitefish. Sequence, assembly and summary statistics are shown for those data obtained in this study (Table 1) and when combined with data from public databases (Table 2). In addition, to provide non-genome-duplicated sister group comparisons, 2,304 clones were sequenced from northern pike (3,624 sequences) (Table 1 and 2). For many of these species, the ESTs provided in this study represent nearly all or most of the known transcripts. Recently published data from rainbow smelt (*Osmerus mordax*) (von Schalburg et al. 2008) was also included in Table 2.

To examine the relationships among the contig consensus sequences of Atlantic salmon we compared all contigs (including singletons) against each other by BLAST and plotted the number of top pair-wise alignments (E-value < 1e-50; length > 200 bp) with the identity score (Figure 6). 36,775 contigs showed greater than 80% identity over 200 bp to at least one other contig. Of these, 12,883 were 97–99.9% similar to at least one other contig. These contigs may represent alleles, recent duplicates or errors in sequence data. 23,892 contigs show between 80 and 96.9% identity with at least one other contig. The large number of duplicated transcripts observed in the Atlantic salmon genome is consistent with the hypothesis of an ancestral salmonid genome duplication, though it is surprising that so many of the duplicated contigs are so similar. This observation is being pursued further in a separate study. The analysis of contig similarity shows that the majority of the 81,398 contigs represent distinct transcripts. Note that since the assembly

process itself combines sequences with high levels of similarity (> 96% repeat stringency with minscore > 300; see Methods), very recent duplications may not all be identified in this process. Furthermore, since the species used in this study differ by greater than 5% (Table 3), this process would be expected to identify ancestral salmonid duplications occurring at or prior to the rainbow trout and Atlantic salmon speciation.

Table 3. Cross-species comparisons of contig transcripts.

	# contigs	# missing in AS[a]	# missing in RT[b]	# missing in both[c]	% sim to AS[d]	Ave len[e]	% sim to RT[f]	Ave len
Atlantic salmon (SJ)[g]	5781	479	1210	354	98.4	705	93.4	493
Atlantic salmon (all)[h]	81398	*na*	36351	*Na*	*na*	*na*	93.3	504
Rainbow trout	50256	13626	*na*	*Na*	93.8	495	*na*	*na*
Chinook salmon	8517	797	1224	426	94.2	510	95.5	510
Sockeye salmon	6200	577	770	298	94.6	571	95.7	569
Brook trout	5424	285	627	174	94.4	580	93.9	522
Lake whitefish	6446	804	1420	608	92.5	425	92.2	399
Grayling	5408	657	1136	506	91.7	435	91.3	400
Northern pike	2380	1894	2001	1846	89.6	241	89.4	251
Rainbow smelt	12159	7462	7812	6920	86.2	431	86.1	419

[a] number of contigs that are not found in the Atlantic salmon database
[b] number of contigs that are not found in the rainbow trout database
[c] number of contigs that are not found in either the Atlantic salmon or rainbow trout database
[d] percent identity compared to the top BLASTN hit to the Atlantic salmon database over 200bp and e-value < 1e-25. In the case of Atlantic salmon (SJ) the comparison is to the McConnell strain.
[e] average length of the BLASTN hit
[f] percent identity compared to the top BLASTN hit to the rainbow trout database over 200bp and e-value < 1e-25
[g] only Atlantic salmon ESTs from the Saint John River strain
[h] all Atlantic salmon ESTs other than those in note "g" above

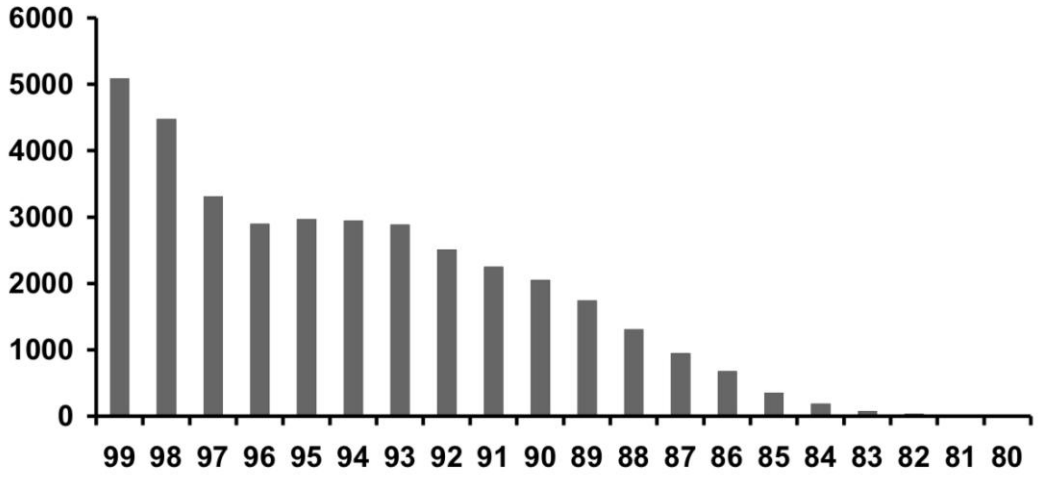

Figure 6. Number of aligned contigs (y-axis) out of 81,398 total contigs is plotted against percent similarity of alignments (x- axis).

Determining the number of genes in Atlantic salmon from the number of EST contigs is difficult for several reasons; 1) the partial representation of genes by EST sequences may result in several contigs associated with a single gene transcript, 2) allelic or recently duplicated genes may be represented by similar but unique transcripts (this latter case is particularly problematic in pseudotetraploid salmonids), 3) alternative splicing, alternative poly adenylation and termination sites from the same gene can result in different transcripts, and 4) transcription products can occur from intergenic regions. An estimation of the number of genes in salmonids will require additional information such as full-length cDNA sequences and gene mapping information.

Salmonid comparisons

Similarity among the different salmonid species was assessed using the top BLASTN hit against Atlantic salmon and rainbow trout EST contig databases. The similarity

values from chinook salmon, sockeye salmon, rainbow trout, Atlantic salmon (McConnell and Saint John River strains), brook trout, grayling, lake whitefish, northern pike and rainbow smelt are shown in Table 3. Assembled contigs (2-stage), rather than individual reads were used for all comparisons to reduce the impact of redundant transcripts. Chinook, sockeye, brook trout, grayling and lake whitefish average 95.5, 95.7, 93.9, 91.3 and 92.2% identity to rainbow trout, and 94.2, 94.6, 94.4, 91.7 and 92.5% identity to Atlantic salmon with over 87% of the contigs matching (E-value < 1e-25) at least one contig in the rainbow trout or Atlantic salmon databases. These comparisons provide only a very general indication of the similarity between transcriptomes of various salmonids, as assemblies contain both 5' (generally genic regions) and 3' (generally 3'-UTR regions) transcript reads. However, these DNA sequence similarity values correspond well to the limited number of values in the literature. Non-coding sequence similarity between rainbow trout and Atlantic salmon are 95% over 120 kb in MH class IA and B loci (Lukacs et al. 2007), and 93–97% over 4 kb in growth hormone (GH) genes (McKay et al. 2004). Similarity between salmon and whitefish is 90–93% in GH genes (McKay et al. 2004).

Northern pike and rainbow smelt average 89.4 and 86.1% identity to rainbow trout and 89.6 and 86.2% identity to Atlantic salmon, but only 25–39% of these contigs matched anything in the rainbow trout or Atlantic salmon database. These latter comparisons have many fewer significant similarities identified partly because of the much older divergence times (Hoegg and Myer 2005). However, the reason for the lower than expected number of matches between northern pike and rainbow trout or Atlantic salmon is not clear. While the more distantly related rainbow smelt contigs show similar numbers of

BLASTX hits to protein databases as salmonids, the northern pike contigs showed very few similarities to Atlantic salmon and rainbow trout contigs (25% compared to 39% for rainbow smelt and 87% for lake whitefish) and very few BLASTX hits to protein databases (8% compared to 50% for rainbow smelt and 36% for lake whitefish). One possible explanation may be due to longer 3'-UTRs in northern pike, but this remains to be confirmed.

Transcriptome representation

It is difficult to assess how comprehensive the extensive Atlantic salmon and rainbow trout EST databases are. However, 73% (37,573 of the 51,199) of all rainbow trout contigs are also found in Atlantic salmon. Moreover, only 28% of those transcripts unique to rainbow trout (13,626) have protein hits ($E < 1e^{-25}$) that support their legitimacy as genic regions, while other single ESTs may be from spurious transcription. 91% of lake whitefish transcripts have a significant similarity (BLASTN comparisons with e-values less than $1e^{-25}$) to the Atlantic salmon or rainbow trout databases. Comparative data from chinook salmon, sockeye salmon, brook trout, grayling, lake whitefish and rainbow smelt are provided in Table 3. Overall, these data provide support for extensive gene coverage in salmonid EST databases.

Full-length analysis

The rapid progress of EST sequencing has enabled an estimation of the number of full-length cDNA clones. Full-length cDNAs (FLcDNAs) are defined as having a "Start – Open Reading Frame (ORF) – Stop – 3' UTR – polyA signal" with the ORF corresponding to a full-length protein. Given multiple start and stop sites, alternative

splicing and partial homologies to known proteins, it is difficult to give precise numbers of completed FLcDNAs. However, TargetIdentifier (using BLAST comparisons to full-length genes in databases and Start signals; (Min et al. 2005)) identifies 17,399 possible FLcDNAs (averaging 1,361 bp in length) from the 81,398 possible transcripts in Atlantic salmon and 10,453 FLcDNAs from the 51,199 rainbow trout transcripts. Thus far, about half of the predicted FLcDNA meet all of the criteria above, and many of the FLcDNAs are already fully characterized on a single clone. These tend to be the shorter (< 1.5 kb) genes. The list of over 10,000 putative FLcDNA transcripts assembled from ESTs is available at the (GRASP website) and further identification of clones for complete sequence analysis is underway.

Salmonid EST, assembly, ORF and annotation database

All ESTs have been deposited in GenBank, however the EST assemblies themselves and the resulting consensus sequences are also very useful in identifying genes. These assemblies, together with the raw data are available (GRASP website). The assembly consensus sequences are available for download and for searching using BLAST tools. A contig visualization tool was developed to allow users to search for similar consensus sequences using BLAST searches, identifying consensus names and then visualizing the sequences, alignment, open-reading frames (ORFs), TargetIdentifier predictions, and BLASTX hits in a single view (Figure 7: Cluster tools). Until such time as the genomes are completed, this database provides the salmonid community with access to several levels of EST and gene analyses.

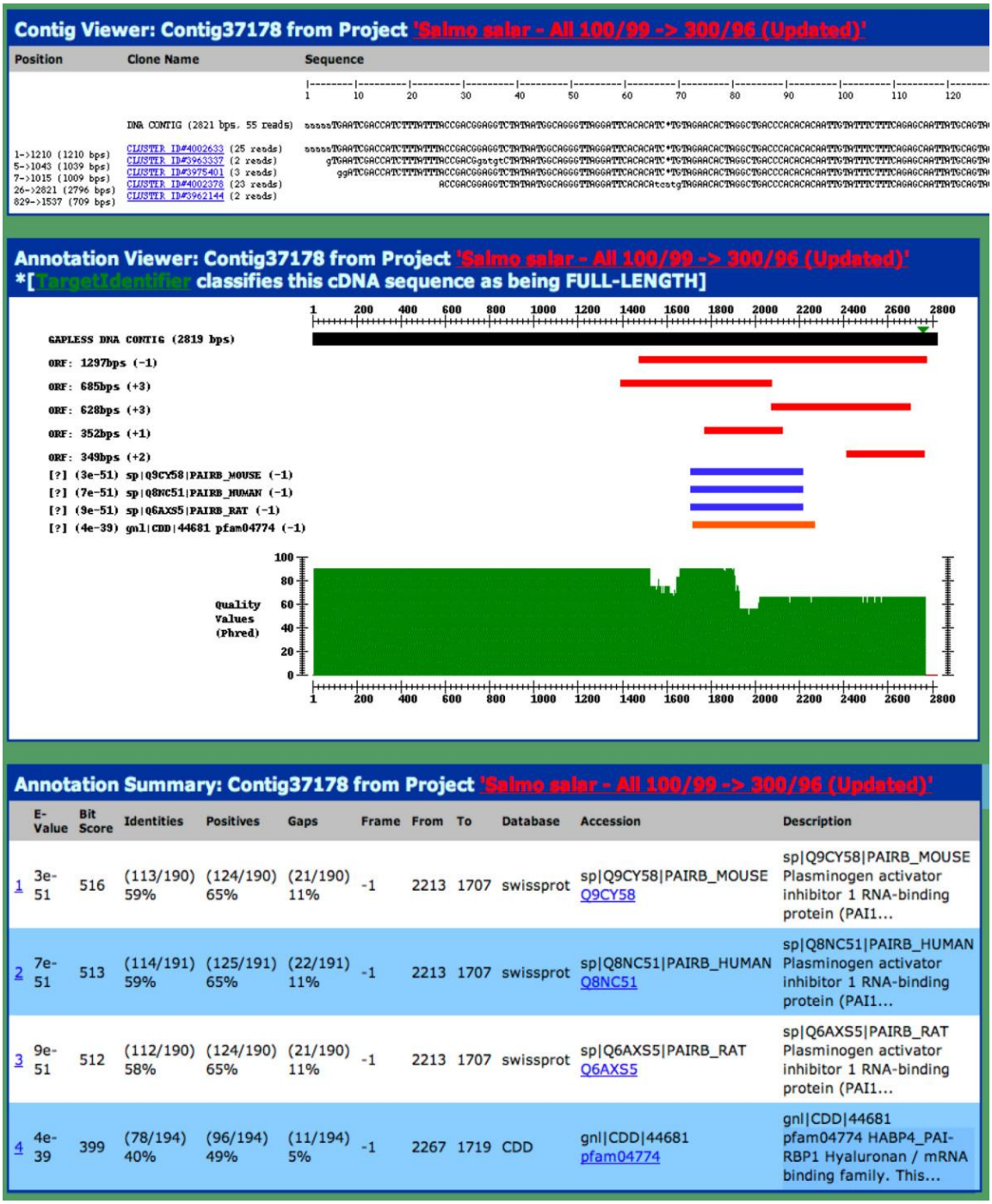

Figure 7. Screen shot of Atlantic salmon contig viewer. The top panel shows the alignment of 100/99 (first stage) clusters along with the number of individual EST reads in each. The second panel shows the 5 largest ORFs and reading frame, the BLASTX hits and reading frame, the Phred quality scores for each aligned position, and indicates whether TargetIdentifier has indicated that this clone is full-length and the predicted position of the START codon (green triangle). Selectable colored bars provide alignment links. The third panel gives specifics of the database hits and links to alignments and database entries.

Salmonid phylogeny and gene duplication

The relationships among major groups of salmonids have been largely unresolved, particularly with respect to the placement of *Salvelinus* (represented in this study by brook trout), *Oncorhynchus* (represented here by rainbow trout, chinook and coho salmon) and *Salmo* (Atlantic salmon) within Salmoninae, and the placement of Thymallinae (grayling), Coregoninae (whitefish) and Salmoninae (salmon) within Salmonidae (Stearley and Smith 1993; Oakley and Phillips 1999; Osinov and Lebedev 2000; Ramsden et al. 2003; Camon et al. 2004; Crespi and MJ 2004). From the EST contigs (Table 2), 78 separate gene sets have been identified, each of which contained at least one EST contig sequence from each of five major salmonid genera (*Oncorhynchus, Salmo, Salvelinus, Coregonus, and Thymallus*), in addition to representation by a non-salmonid (*Osmerus*). Contig sequences within each gene set were aligned, trimmed to a common length (minimum of 300 bp) and analyzed using phylogenetic methods. 73 of the 78 gene sets could be identified by BLASTX searches to SwissProt databases (Table 4). For each gene set, a 70% neighbour-joining (NJ) consensus tree based on 500 bootstrap replicates was generated and the consensus tree rooted with *Osmerus mordax* sequences (rainbow smelt). The single species tree shown in Figure 8 represents a compilation of the phylogenetic results from 78 gene sets. In the summary tree, each branch is noted by; i) the number of 70% consensus NJ trees supporting the branch ii) the number of 70% consensus trees providing no resolution to the branch point, and iii) the number of consensus trees that conflict with the shown result. In this summary, the placement of *Salmo* as a sister group to *Oncorhynchus* and *Salvelinus* is supported in 18 of the 27 gene consensus trees for which resolution was found. Eight alternative

consensus trees support grouping *Salmo* and *Salvelinus*, one consensus tree supports grouping *Salmo* and *Oncorhychus*, and the remaining 51 trees provide no resolution. Thus the overall result is in agreement with some of the more recent studies examining mitochondrial and nine nuclear genes (Crespi and MJ 2004), and suggests good support for grouping *Oncorhynchus* and *Salvelinus* apart from *Salmo* within the Salmoninae subfamily.

Table 4. Gene sets used in phylogenetic analysis.*

Tree	# of Contigs	Align Length	Salmon Group	Subfamily Group	Duplication Salmonidae	Gene Description		
						Accession	e-value	SwissProt Description
1	25	302	Om/Sf	-	no	Q9EPH8	0.00E+00	Polyadenylate-binding protein 1
2	11	287	-	-	yes	Q92572	1.00E-103	AP-3 complex subunit sigma-1
3	18	455	Om/Sf	-	yes	P41134	4.00E-31	DNA-binding protein inhibitor ID-1
4	16	283	-	-	yes	Q24117	2.00E-46	Dynein light chain 1, cytoplasmic
5	11	438	Om/Sf	C/T	yes	P38400	0.00E+00	Guanine nucleotide-binding protein G(i)
6	26	271	-	-	-	P09486	1.00E-138	SPARC precursor
7	12	301	-	-	no			Unknown
8	17	307	Om/Sf	-	yes	P62161	4.00E-80	Calmodulin
9	12	341	Ss/Sf	-	no	Q9Y5S9	4.00E-80	RNA-binding protein 8A
10	16	370	Om/Sf	-	no	P51410	4.00E-95	60S ribosomal protein L9
11	11	305	Om/Sf	-	yes	Q3MHN0	3.00E-95	Proteasome subunit beta type-6 precursor
12	8	411	-	-	-	O60493	1.00E-82	Sorting nexin-3
13	7	448	Ss/Sf	-	yes	O15247	8.00E-97	Chloride intracellular channel protein 2
14	13	500	Om/Sf	-	no	Q9NPI5	6.00E-73	Nicotinamide riboside kinase 2
15	14	379	Om/Sf	S/C	no	P13668	2.00E-47	Stathmin
16	7	638	Ss/Sf	-	no	Q6IQU6	1.00E-154	Ribosome production factor 1
17	11	713	-	-	yes	Q9D915	6.00E-23	Uncharacterized protein C8orf4 homolog
18	9	438	-	-	-	Q8VHZ7	1.00E-125	U3 small nucleolar ribonucleoprotein
19	10	314	-	-	yes	Q05826	1.00E-35	CCAAT/enhancer-binding protein beta
20	18	313	-	-	yes	P97371	1.00E-68	Proteasome activator complex subunit 1
21	10	505	-	C/T	yes	Q96GG9	1.00E-135	DCN1-like protein 1
22	10	620	-	-	yes	Q3T0B6	1.00E-91	Complement 1 Q subcomponent-binding
23	12	442	-	S/C	-	P05141	1.00E-149	ADP/ATP translocase 2
24	8	517	-	-	no	Q9UM00	2.00E-78	Transmembrane and coiled-coil domain
25	14	471	Om/Sf	-	yes	Q6PC69	1.00E-101	60S ribosomal protein L10a
26	13	409	-	-	no	Q5RE33	2.00E-67	Receptor expression-enhancing protein 5
27	14	308	Om/Sf	-	yes	P50397	0.00E+00	Rab GDP dissociation inhibitor beta

28	19	311	Ss/Sf	S/C	yes	P30044	8.00E-64	Peroxiredoxin-5, mitochondrial prec.
29	6	428	-	S/C	no	P15156	1.00E-115	Calcium-dependent serine proteinase
30	10	355	-	S/C	yes			Unknown
31	15	486	-	-	yes	Q9CXL1	8.00E-78	Transmembrane protein 50A
32	11	332	-	-	yes	Q62636	2.00E-90	Ras-related protein Rap-1b precursor
33	22	291	Om/Sf	-	yes	Q3T0Q6	3.00E-71	Cellular nucleic acid-binding protein
34	10	268	-	-	yes	O54734	0.00E+00	Dolichyl-diphosphooligosaccharide
35	8	367	Om/Sf	S/C	no	Q9W719	1.00E-117	Hypoxanthine-guanine phosphoribosyltran.
36	9	609	Om/Sf	-	no	O42123	2.00E-48	FK506-binding protein 1A
37	7	379	-	-	yes	Q9Y5K5	1.00E-161	Ubiquitin carboxyl-terminal hydrolase
38	17	599	Ss/Sf	C/T	yes	P50897	1.00E-120	Palmitoyl-protein thioesterase 1 prec.
39	12	408	-	S/C	no	Q75AA8	7.00E-44	Translation machinery-associated protein
40	15	389	Ss/Sf	-	yes	Q9UL46	4.00E-76	Proteasome activator complex subunit 2
41	8	413	-	-	no	Q96A49	1.00E-100	Synapse-associated protein 1
42	20	333	-	S/C	yes	Q9JK11	8.00E-66	Reticulon-4
43	17	336	-	S/T	no	P67810	5.00E-97	Signal peptidase complex catalytic sub.
44	11	483	-	S/C	yes	Q5XIH7	1.00E-112	Prohibitin-2
45	7	419	Om/Sf	-	yes	P28497	1.00E-136	F-actin-capping protein subunit alpha-2
46	14	465	-	-	no	Q8BLR9	7.00E-16	Hypoxia-inducible factor 1 alpha inhibitor
47	9	610	Om/Sf	-	yes	O75940	9.00E-96	Survival of motor neuron-related-splicing
48	19	604	-	C/T	yes	P60517	4.00E-61	Gamma-aminobutyric acid receptor-
49	17	413	-	C/T	yes	Q6NUC2	1.00E-161	COP9 signalosome complex sub. 6
50	13	350	-	S/C	yes	Q28104	1.00E-141	Coatomer subunit epsilon
51	16	289	-	-	yes	Q2VIU1	7.00E-55	DNA-binding protein inhibitor ID-2
52	12	484	Om/Sf	S/C	yes	P30101	0.00E+00	Protein disulfide-isomerase A3 prec.
53	19	356	-	-	yes	Q6DH65	4.00E-81	Density-regulated protein
54	9	439	Om/Sf	S/C	yes	P26990	4.00E-99	ADP-ribosylation factor 6
55	8	300		-	yes	Q13491	1.00E-124	Neuronal membrane glycoprotein M6-b
56	10	273	-	-	yes	O93277	0.00E+00	WD repeat-containing protein 1
57	11	304	-	S/T	yes	P08132	1.00E-115	Annexin A4
58	12	562	-	-	no	Q12962	3.00E-63	Transcription initiation factor TFIID sub.
59	7	553	-	-	yes	Q9NS69	2.00E-18	Mitochondrial import receptor subunit
60	8	561	Ss/Sf	S/C	no	Q6AYU1	1.00E-158	Mortality factor 4-like protein 1

61	11	378	-	-	no	P50169	1.00E-95	Retinol dehydrogenase 3
62	16	509	-	-	yes	P59998	2.00E-87	Actin-related protein 2/3 complex sub. 4
63	7	619	-	-	no			Unknown
64	12	310	-	-	no	P62316	4.00E-52	Small nuclear ribonucleoprotein Sm D2
65	10	398	-	-	yes	P16527	8.00E-23	Myristoylated alanine-rich C-kinase sub.
66	10	277	-	S/T	yes			Unknown
67	7	388	-	-	yes	O54968	3.00E-80	Nuclear factor erythroid 2-related factor 2
68	8	311	-	-	no	P22232	1.00E-124	rRNA 2'-O-methyltransferase fibrillarin
69	14	300	Om/Sf	-	yes	P40926	1.00E-157	Malate dehydrogenase, mito. prec.
70	9	589	-	-	yes	Q64422	1.00E-138	Glucosamine-6-phosphate isomerase
71	13	577	Ss/Sf	S/C	yes	Q58DU5	1.00E-127	Proteasome subunit alpha type-3
72	14	210	-	C/T	yes	Q15008	0.00E+00	26S proteasome non-ATPase reg. sub.
73	12	348		C/T	no			Unknown
74	18	409		-	yes	Q16799	7.00E-77	Reticulon-1
75	19	462		-	yes	Q9D1J3	5.00E-39	Nuclear protein Hcc-1
76	8	621		-	yes	P19387	1.00E-145	DNA-directed RNA polymerase II subunit
77	16	482		-	yes	Q02878	1.00E-106	60S ribosomal protein L6
78	11	552		C/T	yes	Q802F2	3.00E-95	Selenoprotein T1a precursor

* Listed is the gene set identifier (tree number) along with the number of contigs used in each data set, the length of the respective nucleotide alignment (no gaps), and tentative identification based on BLASTX hits to the SwissProt database (accession number, E-value and description). For each gene set, the tree support for the various arrangements is listed; for example, Om/Sf supports an Oncorhynchus mykiss / Salvelinus fontinalis grouping; or S/C supports a Salmoninae / Coregoninae grouping. In addition there is an indication whether a tree is consistent (yes/no) with an ancestral Salmonidae gene duplication. "-" indicates that the data provides no clear evidence for any particular tree. All EST accession numbers used to make contig consensus sequences, all alignments and the 70% consensus trees are available or online (GRASP website).

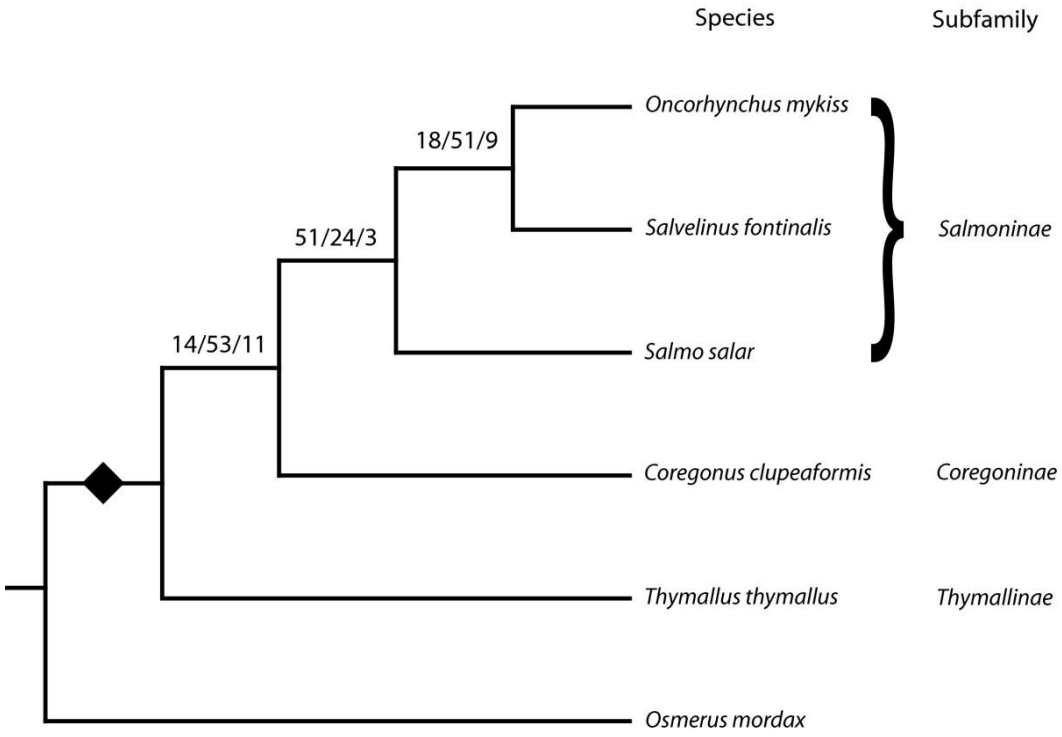

Figure 8. Summary of 78 gene set consensus (70%) trees depicting the relationships among the major groups of Salmonidae. Each branch shows the number of consensus trees supporting the branch, the number of trees providing no information and the number of trees contradicting the branch. The diamond at the base of the Salmonidae cladogram indicates the position where the majority of gene duplications were identified. The individual gene trees that pertain to each branch position are indicated in Table 4.

Consistent with traditional nomenclature, the Salmoninae group, which includes *Salvelinus*, *Oncorhynchus* and *Salmo* is also very well supported with 51 of the 54 resolved trees consistent with this grouping. The three discrepant trees supported a *Salmo/Coregonus* grouping.

The relationships among the three subfamilies within Salmonidae have not been extensively addressed at the molecular level. However, on the basis of a morphological analysis, Coregoninae (whitefish and ciscos) has been hypothesized as the earliest branch within the salmonids [(Stearley and Smith 1993), also see (Osinov and Lebedev 2000)].

In the present analysis, 14 of the 25 informative gene sets are more consistent with the basal position of Thymallinae (Figure 8). Of the discrepant trees, 8 sets support a Thymallinae/Coregoninae grouping and 3 support an ancestral position of Coregoninae. While these data are not definitive, there appears to be some support for an ancestral Thymallinae branching within the Salmonidae with Coregoninae as the sister group to Salmoninae. These data provide the first large-scale molecular view of salmonid subfamily relationships and provide an important perspective on future analyses of duplicated genes, as well as physiological and ecological traits (Ramsden et al. 2003) that have evolved subsequent to the ancestral salmonid genome duplication.

The salmonid whole genome duplication hypothesis makes it is difficult to separate an analysis of species relationships from gene phylogeny. One expectation arising from a relatively recent genome duplication is evidence for extensive nuclear gene duplicates. Subsequent to the genome duplication, the number of observed duplicated transcribed genes is expected to decrease as, over time, one of the duplicates becomes transcriptionally inactive. When multiple species are examined, some species may have both duplication products while other species may have only one representative. Evidence of an ancestral duplication is identified in gene trees that contain multiple species trees that may have missing representatives. Of the 78 gene sets examined in this study, 51 show clear evidence of multiple species trees within gene trees that are consistent with a gene duplication in the ancestor of Salmonidae, sometime after the separation of Osmeriformes and Salmoniformes fish. 23 gene sets (Table 4) provided no evidence for any ancestral gene duplication, and 4 sets could not be interpreted. The data from 78 gene sets representing 372 consensus sequences and 11,397 bp of aligned DNA

from five salmonid genera, indicate that a large number of salmonid genes show evidence of extensive gene duplication at a phylogenetic position that is consistent with the whole genome duplication in the ancestral Salmonidae hypothesis. Further studies of Esociformes fish will more precisely establish the timing of some of these gene duplications.

Salmonid 32 K microarray

To use the data generated by ESTs and assemblies for examining gene expression, a new 32,000 feature cDNA microarray was developed. This new array is based on the existing 16 K GRASP array (von Schalburg et al. 2005) plus 14,496 additional Atlantic salmon and 1,491 additional rainbow trout contigs that were identified as unique and were successfully amplified in this study. The 32 K cDNA microarray is composed mainly of 27,917 Atlantic salmon (AS) and 4,065 rainbow trout (RT) cDNA elements or features. 54% of the elements have fairly stringent ($1e^{-10}$) hits to annotated members in public protein databases. Hybridization performance of this array was evaluated using Atlantic salmon, rainbow trout, coho salmon, brook trout and lake whitefish RNA obtained from liver organs. The success of hybridization of labeled target to the salmonid elements was judged by the numbers of Atlantic salmon and rainbow trout elements passing background plus 2 SD threshold values (see Methods). No transformations or normalizations were performed on the data. Overall statistics are presented in Table 5. In summary, for RNA isolated from the liver of Atlantic salmon, rainbow trout, coho salmon, brook trout and lake whitefish, an average of 48% of the 32,018 elements showed significant detection levels of expression. Comparing these results to that from the previous 16 K GRASP arrays indicates that doubling the number

of elements from 16 K to 32 K resulted in the ability to assess expression patterns of approximately 61% additional transcripts. This represents a substantial increase in our ability to assess gene transcription patterns in salmonids. The hybridization performances of the different salmonid species (assessed from numbers of Atlantic salmon and rainbow trout elements passing threshold) conformed to expectations, given the close evolutionary relationships of the species tested (92–94% identity, Fig. 8) and, with the possible exception of brook trout, all members of the family Salmonidae tested showed similar levels of hybridization to the Atlantic salmon and rainbow trout elements on the 32 K microarray. As the Salmonidae family represents 68 closely related species, the 32 K cDNA array provides an excellent opportunity to evaluate gene expression patterns of a large group of culturally and economically important species.

Table 5. Cross species hybridization results for the salmonid 32 K cDNA microarray. *

Salmonid Species	% +'ve	%CV
Atlantic salmon (n=4)	48.6%	12.0%
Rainbow trout (n=4)	58.1%	9.8%
Coho (n=4)	52.3%	23.2%
Brook Trout (n=4)	35.0%	2.4%
Whitefish (n=4)	47.7%	7.8%

* Percent elements on cDNA array with median signal intensity greater than threshold (background signal+ 2SD). %CV is percent coefficient of variation and "n" is the number of biological replicates.

Chapter 3

Gene Identification and Full-Length Reference Genes in Salmonids

Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., Munro, S., Moore, R., Holt, R. A., Jones, S. J. M., Davidson, W. S. and Koop, B. F. (2010), '*Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome' *BMC Genomics* **11**, 17.

Summary

Salmonids are one of the most intensely studied fish, in part due to their economic and environmental importance, and in part due to a recent whole genome duplication in the common ancestor of salmonids. This duplication greatly impacts species diversification, functional specialization, and adaptation. Extensive new genomic resources have recently become available for Atlantic salmon (*Salmo salar*), but documentation of allelic versus duplicate reference genes remains a major uncertainty in the complete characterization of its genome and its evolution.

From existing EST resources and three new full-length cDNA libraries, 9,057 reference quality full-length gene insert clones were identified for Atlantic salmon. A further 1,365 reference full-length clones were annotated from 29,221 northern pike (*Esox lucius*) ESTs. Pairwise d_N/d_S comparisons within each of 408 sets of duplicated salmon genes using northern pike as a diploid out-group show asymmetric relaxation of selection on salmon duplicates.

9,057 full-length reference genes were characterized in *S. salar* and can be used to identify alleles and gene family members. Comparisons of duplicated genes show that while purifying selection is the predominant force acting on both duplicates, consistent with retention of functionality in both copies, some relaxation of pressure on gene

duplicates can be identified. In addition, there is evidence that evolution has acted asymmetrically on paralogs, allowing one of the pair to diverge at a faster rate.

Introduction

Salmonidae (including salmon, trout, charr, whitefish and grayling) are of economic and environmental importance, leading to a high level of interest in many different areas of biology. Of the sixty-six species in this family (Nelson 2006), Atlantic salmon (*Salmo salar*) has been used as a model for studies in several areas including osmoregulation, environmental toxicology, immunology, growth, physiology, and genomics (Handeland et al. 1998; Hutchings and Jones 1998; Boeuf and Le Bail 1999; Mommsen and Vijayan MM Moon 1999; Norris et al. 1999; Garant et al. 2000; Sutton et al. 2000; Bell et al. 2001; King et al. 2001; Landry et al. 2001; Jacobs et al. 2002; Bernatchez and Landry 2003; Grimholt et al. 2003; Moore et al. 2003; Ng et al. 2005; Zheng et al. 2005; Derome et al. 2006; Jorgensen et al. 2006; Krogdahl et al. 2006; Lukacs et al. 2007; Harstad et al. 2008; von Schalburg et al. 2008; Yazawa et al. 2008; Andreassen et al. 2009). Both *S. salar* and the closely related rainbow trout (*Oncorhynchus mykiss*) are commonly used as important sentinel species to monitor the health of aquatic environments (Mos et al. 2008). Conservation and enhancement of wild stocks of these fish continues to be the subject of very large internationally concerned groups (Klemetsen et al. 2003; McGinnity et al. 2003). Basic biological knowledge of *S. salar* serves as a foundation for improving fish health, conserving wild stocks, and increasing the commercial sustainability of aquaculture. Recent efforts in genomics have provided new tools to address fundamental questions regarding fish health, ecology, physiology, and genetics, as well as allowing investigation of post-tetraploidization genome remodelling (Rexroad et al. 2003;

Govoroun et al. 2006; Adzhubei et al. 2007; Danzmann et al. 2008; Koop et al. 2008; Phillips et al. 2009). Detailed efforts to annotate the entire complement of *S. salar* genes will greatly facilitate a better understanding of all aspects of salmonid biology.

The study of salmonid genomes is made more difficult and biologically interesting because of a whole genome duplication (WGD) that occurred through an autotetraploidization event in the common ancestor of salmonids between 25-100 million years ago (Allendorf and Thorgaard 1984). Extant salmonids are currently in a pseudotetraploid state and are in the process of reverting to a stable diploid state (Danzmann et al. 2008). Though many of the gene duplicates from the WGD have been lost through deletion events or by being converted into pseudogenes, many sets of paralogs remain. As a result, there are practical problems in distinguishing among alleles, recent segmental duplications, gene family members, and duplications arising from the WGD. Experimentally, *S. salar* genes have proven challenging to characterize because of the complexities resulting from assembling large numbers of partial mRNA sequences represented by ESTs obtained from these duplicated and other closely related sequences. In addition, interspersed repeat sequences (de Boer et al. 2007) can lead to the formation of incorrect assemblies of genomic sequences and transcripts (contigs). To resolve these potential errors, a gene containing coding sequence (CDS) flanked by 5' and 3' untranslated regions (UTR) coming from a single, completely characterized cDNA clone provides an important reference sequence representing a single allele of a single gene. The expansion of reference clone resources are particularly important not only in identifying other potential alleles and gene duplicates that are so pervasive in the

pseudotetraploid salmonids, but also in studying fundamental genetic rates and modes of evolutionary change.

Relatively few organisms and lineages have been used to examine the evolution of duplicated genes following a WGD. Morin et al. (2006) investigated the selective pressures acting on paralogs in *Xenopus laevis*, which resulted from allotetraploidization, and found wide-spread purifying selection but with some relaxation of pressure relative to orthologs in the diploid *Xenopus tropicalis*. Maere et al. (2005) studied substitution rates and found certain functional categories of genes that were selectively lost after genome duplication events in *Arabidopsis thaliana*. In a larger scale study, Conant and Wagner (2003) researched the genomes of a number of different organisms that have undergone WGDs, testing for asymmetric divergence of paralogs which they found in 20 - 30% of duplicates. Looking at asymmetrically evolving paralogs in yeast, Turunen et al. (2009) recently presented evidence for relaxation of selective pressures. Furthermore, in another examination of the WGD in *S. cerevisiae*, positive selection was detected in a substantial portion of paralogs (Fares et al. 2006). These studies examined the ratio of amino acid changing substitutions to silent substitutions (d_N/d_S) to measure evolutionary rates. The present study incorporates some of these approaches to identify evolutionary patterns in the genome of *S. salar*. Since there is not a large number of examples of post-tetraploidization evolution available for study, the WGD in salmonids becomes an important area for research.

Of the few organisms studied, some have been examined by a number of research groups. Since this is one of the first studies examining the post-tetraploidization evolutionary patterns in the salmonid genome, it is our hope that other groups in addition

to our own will expand on the work presented here, incorporating growing datasets and using a wide variety of phylogenetic and evolutionary methods.

Characterizing evolutionary changes in polyploid genomes requires comparison to a pre-WGD out-group species so that differences in substitution rates with respect to an ancestral genomic state can be determined. Ishiguro et al. (2003), Lopez et al. (2004), and Li et al. (2008) report that the Order Esociformes is the closest non-polyploid sister group to the Salmoniformes. Karyotypic data (Phillips and Rab 2001; Mank and Avise 2006) and C-values of ~3.0 - 3.3 pg in salmonids and ~0.9 - 1.4 pg in esocids (Gregory 2002) are consistent with the occurrence of the WGD after the divergence of esocids and salmonids. In particular, studying northern pike (*Esox lucius*) as a representative of the order would provide an opportunity to continue building upon existing efforts. As there were only 158 core nucleotide sequences, 83 protein sequences, and 3,612 EST sequences (Koop et al. 2008) available for northern pike prior to this study, it was necessary to expand sequence information of this species before a more thorough analysis of salmonid gene duplications could be done.

The objectives of this study were to: 1) obtain a large number of full-length reference cDNA clone sequences; 2) expand the transcriptomic resources (ESTs) of *E. lucius;* and 3) identify evolutionary patterns of duplicated genes in the autotetraploid *S. salar* species.

Material and Methods

Tissues, RNA, and Sampling

Adult *S. salar* tissues (brain, kidney, spleen) were obtained from Robert Devlin at the Department of Fisheries and Oceans (WestVan Lab, West Vancouver, British Columbia). Adult *E. lucius* tissues (head kidney, spleen, heart, gill) were obtained from Frank Koop at Charlie Lake (Fort St. John, British Columbia). Tissues were rapidly dissected, flash-frozen in liquid nitrogen or dry ice, and stored at -80°C until RNA extraction.

cDNA Libraries

Three full-length, non-normalized cDNA libraries were constructed using a full-length cDNA library protocol (Research Genetics Inc.). This protocol employed an enrichment of 5'-CAPed mRNA which prevents truncated mRNA from being reverse-transcribed, followed by transfer of intact double-stranded cDNAs directly into the library vector using Gateway® recombination cloning. An estimated 65-85% of the clones were full-length (Invitrogen Full-Length cDNA Library Construction).

Different mRNA size fractions were used in the construction of the three libraries. The libraries were created using transcripts between 0.6 to 1.1 kb (rgg), 1.1 to 2.0 kb (rgh) and > 2.2 kb (rgf). The cDNA libraries were directionally constructed (5' M13 Forward, 3' M13 Reverse) in pENTR222 vector (Research Genetics Inc.).

The *E. lucius* library (evq) was made from head kidney, spleen, heart and gill cDNAs that were normalized and directionally cloned (5' M13 Forward, 3' SP6) in pAL17.3 vector (Evrogen Co.). Sequences from a previously characterized *E. lucius* brain, kidney, and spleen library (Koop et al. 2008) were also utilized.

Sequencing, Sequence Analysis, and Contig Assembly

Clone libraries were plated and robotically arrayed in 384-well plates as detailed previously (Koop et al. 2008). Plasmid DNAs were extracted and BigDye Terminator (ABI) cycle sequenced on ABI 3730 sequencers using conventional procedures and the following primers: 5'-T18-3', M13 forward (5'-GTAAAACGACGGCCAGT-3'), M13 reverse (5'-AACAGCTATGACCAT-3' or 5'-CAGGAAACAGCTATGAC-3'), and SP6WAN (5'-ATTTAGGTGACACTATAG-3') for 3' end sequencing of Evrogen libraries. Base-calling was performed using PHRED (Ewing and Green 1998; Ewing et al. 1998) on chromatogram traces. Vector, polyA tails, and low quality regions were trimmed from EST sequences. Short (100 bp) low quality sequences were discarded. Assembly of *S. salar* ESTs into contigs employed two-stage processing using PHRAP (Figure 9 parts 1-2) (Koop et al. 2008). CAP3 (Huang and Madan 1999), using default parameters, was employed for a single assembly of *E. lucius* ESTs in place of the PHRAP two-stage approach, the purpose of which is to handle WGD transcriptomes.

FLcDNA contig identification

The analysis of full-length transcripts began with all EST contig sequences. Since each contig represents a potential transcript, it must be determined if a transcript is complete or incomplete. A complete or full-length transcript contains an entire CDS for a gene product, along with the flanking 5' and 3' UTR. Incomplete transcripts are mRNA that have not been fully reverse-transcribed during cDNA library creation, and therefore may not contain the complete CDS or the 5' UTR. Because of the selection for polyA tails during cDNA library creation, both incomplete and complete transcripts contain a polyA

tail. Inherent experimental errors in the reverse transcription step during cloning result in 5' incomplete cDNA inserts.

Using an e-value filter of $e \leq 10^{-5}$, the top ten SwissProt high-scoring segment pairs (HSPs) from BLASTX for each contig were analyzed in succession to identify the correct open reading frame (Figure 9 part 3). Full database protein matches must be contained within a full-length transcript sequence. HSPs often do not match a homologous protein in its entirety. This situation exists for the following reasons: i) a transcript is incomplete; ii) a transcript represents a pseudogene; iii) a transcript represents a novel gene product, but contains a domain common to an existing non-homologous protein. In cases where the match region between a transcript query and a subject protein sequence does not fully encompass the length of the subject protein, the two complete sequences are checked to determine whether the 5' end of the transcript extends beyond the 5' end of the known database reference protein sequence. In situations where the transcript is not long enough to accommodate the full database protein length, transcripts are disregarded from further FLcDNA consideration (Figure 9 part 4). In cases where the transcript is long enough to contain the known database reference protein, the transcript is kept for further analysis.

An ORF is a single continuous region on a processed transcript sequence that encodes a complete protein. These regions are defined by a start codon (ATG) and end with an in-frame (non-coding) stop codon (TAG, TAA, or TGA). When a potential start codon is identified, a corresponding in-frame stop codon is verified to complete an ORF. Stop codons found upstream of the start are useful but not essential in defining the proper coding region. Start codon positions are determined by examination of ATG motifs

present upstream, in-frame or within 30 bp downstream of the beginning of the aligned reference protein. Coding regions often contain multiple methionine codons, which may obscure prediction of a start codon. If a methionine codon is not found between the first upstream stop codon and the predicted start codon, it is assumed that the start codon is correct. If a methionine is found upstream of the predicted start codon and still is in-frame with the downstream stop codon, this new ATG motif position is assigned as the correct start codon. Once a start codon is identified, a corresponding in-frame stop codon is verified to form the completed ORF (Figure 9 part 5).

Reference FLcDNA identification

Complete transcripts whose coding regions can be fully represented by a single cDNA clone sequence are considered reference FLcDNAs. These FLcDNAs contain 5' and 3' UTRs flanking an ORF that matches or is consistent with a known protein identified by a BLASTX similarity search.

Subsequent to the initial clustering and annotation of 434,384 ESTs to establish the putative transcript set, three full-length cap-trapped libraries (rgg, rgh, rgf) were created and bi-directionally sequenced. Of these libraries, rgf ESTs were assembled, using PHRAP, to produce transcripts to be compared to the established set of 81,398 putative transcripts (Koop et al. 2008). The clone reads from the original libraries that were used to produce the putative transcript set were mapped back, via local alignment, to this putative set to determine which clones contained a reference FLcDNA insert. Library rgf was also mapped back to the putative transcript set. Reads from identical clones that map against the same putative transcript and contain sequence overlap are considered to be from a reference clone. If the forward and reverse reads from the same clone both

overlap an identical region of the transcript, that clone is classified as being complete. There are cases where clones have forward and reverse reads that do not overlap when mapped to the same transcript. In this scenario, a gap exists between the reads when mapped to the cluster, suggesting an area for which primers can be designed for further sequencing. These clones are known as incomplete clones, and formed a subset of 4,380 rgf clones that were later resequenced to completion. Libraries rgg and rgh were not included in any of these comparisons but were analyzed on an individual clone basis (discussed below).

The 81,398 putative transcripts were established using a two-stage EST clustering process (Koop et al. 2008). As a result, the second-stage assembly begins with sequences from the first-stage assembly. Prior to assembly, gaps from the sequence set need to be removed. As a result of a two-stage assembly, not only does one lose gaps that initially may have been introduced, but EST read names are also lost. The modification of gaps in assembled sequences affects the positions in the reads the assemblies are composed of. To recalculate read positions and reference FLcDNA clones, a local alignment of all reads from all libraries (except rgg, rgh) was performed against the putative second-stage transcript set of 81,398 sequences. Reads from identical clones that map against the same transcript set corresponding to FLcDNA contigs, regardless of sequence overlap, are determined (Figure 9 part 6).

All 6,081 complete (overlapping reads) clones (Figure 9 part 7) that flanked the entire predicted ORF region, in the set of 10,026 FLcDNAs, are selected and form the reference FLcDNA clone set. In this set, more than one complete reference clone may map to a single transcript. Therefore, to produce a non-redundant set of complete FLcDNA

reference clones, only the longest complete reference clone that maps to a specific transcript is selected. In the case where clones are of equal length, the clones are simply chosen according to alphabetical order, resulting in 5,853 non-redundant reference clones that are unique to a single transcript (Figure 9 part 8).

Reference FLcDNA identification using individual clone assembly

In addition to analyzing reference FLcDNA clones via transcript mapping, two full-length libraries (rgg, rgh) and a single fully sequenced full-length library subset (incomplete clones from rgf) were examined. Each of these three *S. salar* libraries was analyzed independently.

Clones were assembled individually so that reads that were already known to be from the same clone could be explicitly allowed to join, while erroneous additions of other sequences could be minimized. Using this method, libraries rgg, rgh, and a portion of rgf clones that were selected to be resequenced were analyzed independently from each other. For all sequence reads from rgg and rgh libraries, individual clone PHRAP assemblies (Green) (minscore 8, repeat stringency 99%) were performed (Figure 10 part 1).

The subset of 4,380 selected rgf library clones were fully resequenced (minimum PHRED 20 for entire sequence) (Ewing and Green 1998; Ewing et al. 1998). Those clones that contained a gap or the end sequences were of poor quality were rearrayed to a 384-well plate for further finishing via primer-walking. All sequences from this fully-sequenced group could therefore be directly selected for further full-length analysis.

Redundancy was minimized by performing an all versus all pairwise BLASTN comparison per library. Transcripts that showed greater or equal to 98% similarity over

200 bp were considered redundant. For sets of redundant transcripts, the longest sequence was taken as the non-redundant representative (Figure 10 part 5).

Reference FLcDNA assessment

To properly assess reference FLcDNAs, sequences were checked for polyA tails. A polyA tail is defined as a 3' region of 15 or more consecutive "A" resides. If such a polyA tail was detected, those sequences were deleted as well as all subsequent downstream sequence.

For *S. salar* and *E. lucius*, reference FLcDNAs that could be confirmed by a contig sequence were identified. Using BLASTN to determine matches, each reference FLcDNA set was compared to its contig assembly. Reference FLcDNAs that showed 100% similarity over ≥ 95% of its sequence were considered to be identical. Those that did not possess confirmed identity were categorized as unique reference FLcDNAs.

Selection of homologous genes

The 10,026 full-length *S. salar* cDNA contigs were used to identify homologous sequences and construct sets containing two paralogs from *S. salar* and one ortholog from *E. lucius* for determination of synonymous and non-synonymous substitution rates. It was necessary to start with known full-length contigs in order to be certain of the translation frame and ORF in the *E. lucius* and *S. salar* ESTs. Full-length sequences with the same accession number as another were removed from the query set resulting in a set of 5,219 unique contigs. This was because sequences with the same annotation would be likely to return the same cluster of ESTs when used to identify homologous sequences. The full-length sequences were translated to protein using ORF information. A

TBLASTN was performed using these amino acid sequences as queries against a translated nucleotide database consisting of all of the *S. salar* and *E. lucius* EST contig assemblies, 93,060 in total. An e-value of 10^{-10} or less was required for a match and 100 matches for each query were considered. The contigs corresponding to the BLAST matches were gathered into clusters, one cluster for each query sequence. As a preliminary screening function, the BLAST alignment was checked for percent coverage of the length of the amino acid query sequence. If the alignment covered 50% or greater, it was put into the cluster; otherwise, the alignment was discarded. BLAST information (hit region, frame of translation, and percent positive and identical matches) for each hit was retained. Each group of contigs was then translated using the frame information from the TBLASTN results and the resulting amino acid sequences were put into another cluster. Thus two corresponding sets of clusters were created, one protein and one nucleotide.

Determination of alignment regions

The DNA sequences in each individual cluster were trimmed to a common region of alignment with respect to the query protein sequence. The sequence that had the longest local alignment was compared with the sequence with the next longest alignment, and the common aligned region was retained, potentially trimming one or both ends of either sequence. This was repeated with sequences having shorter and shorter alignments until a common region was found for that cluster. The minimum length of the alignment was 300 bp; if a sequence's alignment would cause the common region to drop below 300 bp, that contig was removed entirely. In addition, the original TBLASTN alignment was required to have at least 75% positive amino acid matches. This same process was done

on the protein sequences to get the same alignment regions using 100 residues as the minimum length.

Sequence alignment

The trimmed protein sequences were aligned using ClustalW with default parameters (Thompson et al. 1994). Using the ClustalW alignments and the nucleotide clusters, RevTrans was used to create codon-aware DNA alignments (Wernersson and Pedersen 2003). The alignments were further screened for the presence of alleles and very similar sequences as well as odd sequences that did not closely match the cluster. This filtering was done by aligning each sequence in the cluster with every other sequence. If an alignment showed greater than 98% identity or less than 60% identity or the alignment was shorter than 90% of the length of the longer sequence, the sequence was dropped from the cluster.

d_N/d_S estimation

Only the final alignments containing one sequence from *E. lucius* and two sequences from *S. salar* were used in the analysis, 408 in total. The 408 clusters with the required three sequences were then converted from FASTA format to a sequential alignment form that the PAML package could use as input. The YN00 program in the PAML package was used with default parameters on each gene trio to determine d_N and d_S rates (Yang and Nielsen 2000). In addition, ω (d_N/d_S) values for the individual branches of the tree were estimated based on the formulae

$$d_S(AO) = (d_S(AC) + d_S(AB) - d_S(BC)) / 2 \qquad (1)$$

$$d_N(AO) = (d_N(AC) + d_N(AB) - d_N(BC)) / 2 \qquad (2)$$

$$\omega = d_N / d_S \quad (3)$$

where A and B are the extant paralogs, C is the extant ortholog, and O is the point of gene duplication (Miyata and Yasunaga 1980).

Gene Ontology analysis

Gene Ontology terms were found for the sequences that had the highest fold-change in ω between the post-duplication branches (> 3x; n = 67) as well as the lowest fold-changes (<1.75x; n = 61). BLASTX searches (NCBI Blast) were performed on sequences against the SwissProt database (Bairoch and Apweiler 1998). Gene Ontology terms were taken from Entrez Gene (NCBI Entrez Gene database) for the top hit using $e \leq 10^{-10}$.

Results

Full-length cDNA library construction and analysis

The majority of existing EST data from *S. salar* came from highly normalized cDNA libraries that were full-length (FL) biased (Rise et al. 2004; von Schalburg et al. 2005; Adzhubei et al. 2007; Koop et al. 2008). To specifically identify more full-length transcripts, a protocol for enrichment of 5'-CAPed mRNA was employed which prevents truncated mRNA from being reverse-transcribed, followed by transfer of intact double-stranded cDNAs directly into the library vector using Invitrogen Gateway® recombination cloning (Invitrogen Full-Length cDNA Library Construction). Starting from *S. salar* brain, head kidney, and spleen tissues, three non-normalized, size selected, full-length libraries were constructed. mRNAs were size-selected for 600 to 1,100 bp (rgg), 1,100 to 2,000 bp (rgh) and >2,200 bp (rgf). 7,680, 7,680, and 16,128 clones from rgg, rgh, and rgf, respectively, were bi-directionally sequenced. For the short insert library (rgg), 11,917 sequences were obtained and assembled into 1,833 transcripts (903 singletons and 930 contigs). This library had the fewest novel transcripts and had the highest redundancy in terms of identified sequences. The majority of transcripts in this library were identified as hemoglobins, ribosomal protein genes or other genes previously seen in our existing EST dataset (Koop et al. 2008). For the mid-sized insert library (rgh), 12,250 sequences were obtained and assembled into 5,305 transcripts (3,088 singletons and 2,217 contigs). While the sequence diversity of this library was higher, nearly all complete transcripts had been previously identified (Koop et al. 2008). For the large-sized insert library (rgf), 30,415 sequences were obtained and assembled into 15,125

transcripts (11,190 singletons and 3,935 contigs). This library contained the highest number of novel transcripts.

Identification of *S. salar* Full-Length cDNA contigs from existing EST assemblies

Starting with a 434,384 *S. salar* EST assembly (Koop et al. 2008) (Figure 9 part 1), 81,398 contigs (Figure 9 part 2) were compared to the SwissProt protein database (Bairoch and Apweiler 1998) (Figure 9 part 3) and 34,451 unique transcripts were identified. 14,021 of these were potential FLcDNA contigs as determined by similarity comparisons to known proteins (Figure 9 part 4). These assembled sequences represent potential full-length transcripts with significant similarity to SwissProt protein sequences. 10,026 (mean = 1,295 bp; range = 195 - 4,696 bp) of these contigs contained complete ORFs and 5,853 of these could be represented by a single completely characterized, non-redundant clone. These clone sequences are consistent with contig consensus sequences representing two or more different clones and were provisionally designated as reference FLcDNAs.

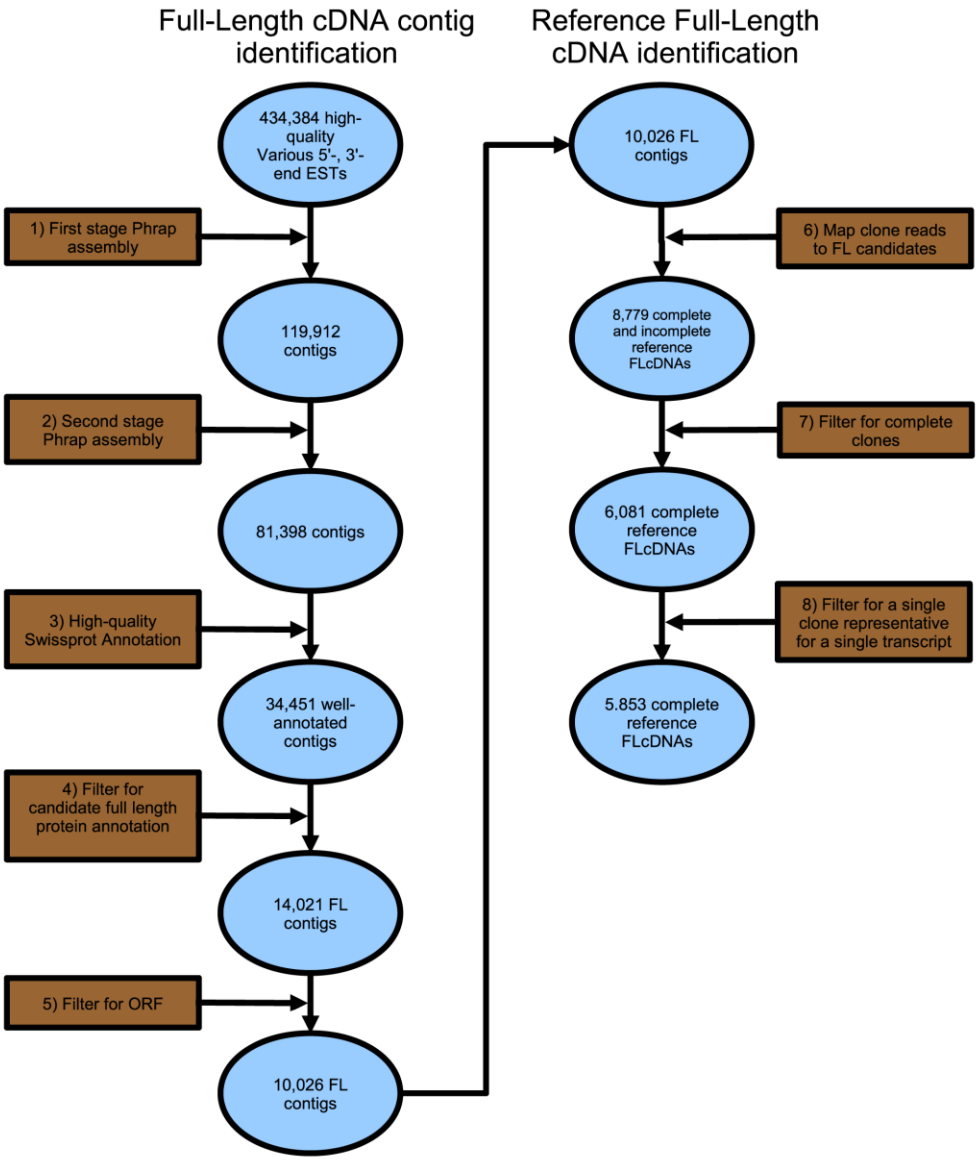

Figure 9. Schematic of *S. salar* FLcDNA contig identification and reference FLcDNA identification. Two-stage assembly of 434,384 high-quality 5'- and 3'-end ESTs identified 81,398 contigs (1-2) for FL contig identification. A BLASTX was carried out resulting in 34,451 well-annotated contigs (3), which were further reduced to 14,021 FL annotations by increasing the stringency of the local alignment length (4). In-frame annotation-flanking start and stop codons were found from the reduced set, resulting in a set of 10,026 FL contigs (5). The FL contigs represent the complete set of FL unique putative transcripts. A set of all reads and subsequently sequenced library rgf reads was mapped to the FL contigs (6). Those clones whose 5'- and 3'-end reads map to the same contig were analyzed to determine sequence overlap (complete) or non-overlap (incomplete) (7). Only complete clones are considered, and a single representative of a clone is taken for each transcript resulting in 5,953 complete reference FLcDNAs (8).

E. lucius ESTs

A full-length biased, normalized cDNA library from *E. lucius* head kidney, spleen, heart and gill tissues was constructed and 15,360 clones were bi-directionally sequenced. 29,221 sequences averaging 731 bp were obtained and, with the previously available 3,612 EST sequences (Koop et al. 2008), assembled into 11,662 contigs (2,791 singletons and 8,871 clusters; mean cluster = 2.2 reads, 1,384 bp; max cluster = 106 reads). BLASTX analysis (Altschul et al. 1997) revealed a total of 3,816 unique transcripts with strong SwissProt protein similarity (e-value $\leq 10^{-5}$). Using the same method outlined in Figure 9 part 4 for *S. salar*, 1,830 were identified as potential full-length transcripts. After ORF analysis (Figure 9 part 5), 1,543 FLcDNA contigs contained sequences corresponding to full-length proteins (mean = 1,044 bp; range = 312 - 2,984 bp) and 1,365 non-redundant reference clones were identified.

Reference Full-Length cDNA identification using individual clone assembly

Paired 5' and 3' sequence reads from short and mid-sized insert FLcDNA libraries from *S. salar* (rgg: 11,917 reads and rgh: 12,250 reads) were assembled individually to yield 6,941 rgg and 8,470 rgh cDNA clone sequences. These sequences were selected for further full-length, ORF and non-redundancy analysis (Figure 10). The short-insert library (rgg) yielded 274 new, full-length protein reference clone sequences. The midrange insert library (rgh) yielded 357 new FLcDNA reference clone sequences. The low yields of novel reference clones likely reflect similar clone insert sizes obtained from previous cDNA library characterizations.

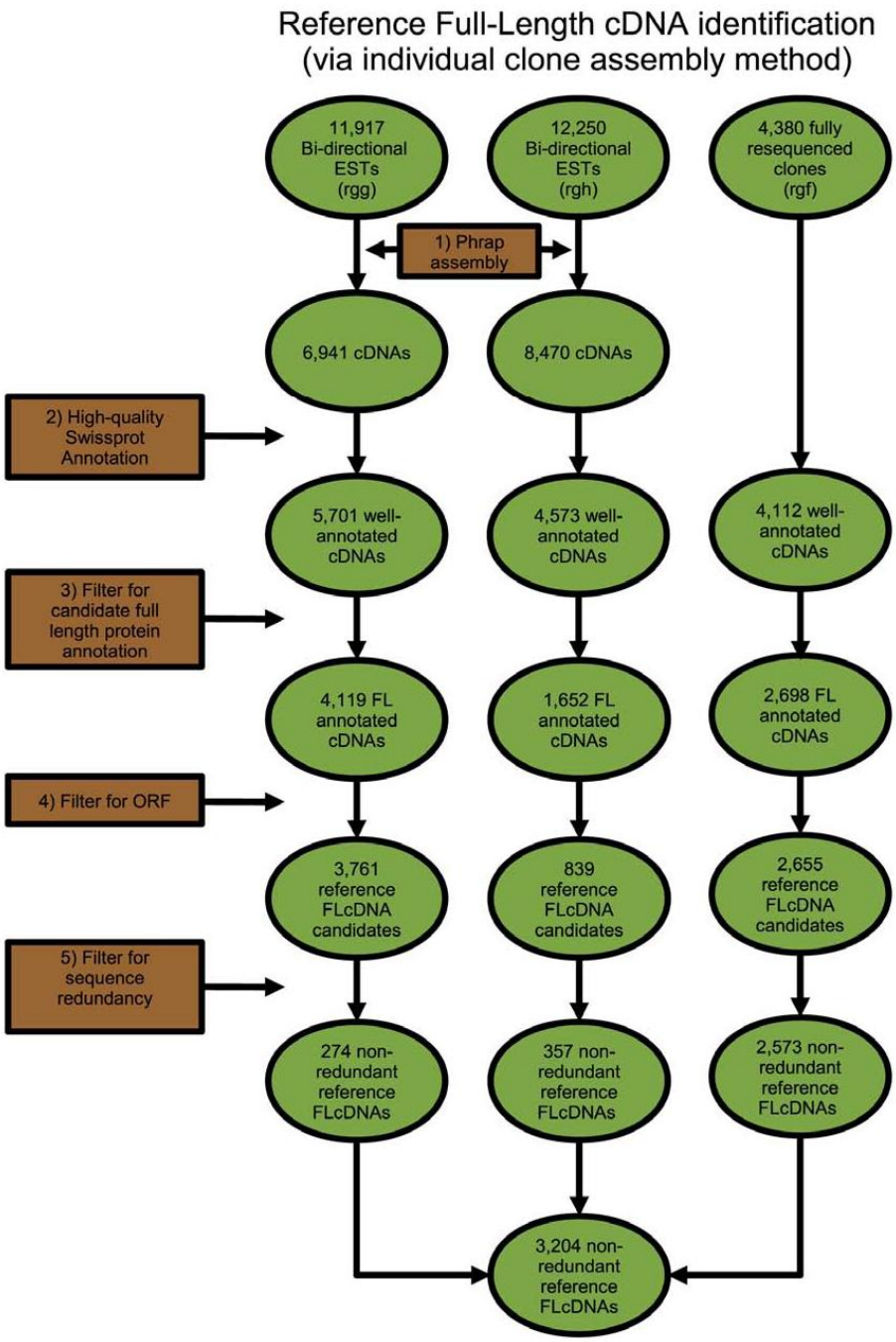

Figure 10. Schematic of *S. salar* reference FLcDNA identification through individual clone assemblies. Three full-length 5'-CAP enriched libraries were created. A 4,380 clone subset of library rgf was resequenced to completion. Libraries rgg and rgh were bi-directionally sequenced and individually assembled using PHRAP (1). A BLASTX was carried out resulting in a total of 14,384 well-annotated cDNAs (2), which were further reduced to 8,469 FL annotations by increasing the stringency of the local alignment length (3). In-frame annotation-flanking start and stop codons were found from the reduced set, resulting in a set of 7,255 reference FLcDNA candidates (4). Intra-library sequence redundancy was minimized using an all versus all pairwise BLASTN comparison (5), resulting in a total set of 3,204 non-redundant reference FLcDNAs.

The assembled 15,125 transcripts from the large-insert full-length cDNA library (rgf) from *S. salar* were initially examined for non-redundancy with existing full-length reference genes and potential for representing a full-length gene (5' and 3' non-overlapping clone sequences were consistent with the 5' and 3' ends of a known complete protein). Based on partial reference FLcDNA characterization, 4,380 clones were chosen for complete characterization using primer walking methods. Once these clones were completely sequenced, 4,112 were shown to have significant SwissProt similarity. Of those clones, 2,573 represented novel non-redundant *S. salar* transcripts with complete ORFs that corresponded to known proteins. Clones whose inserts contained full 5' annotation (Figure 10 part 3) and a proper ORF (Figure 10 part 4) were designated as reference FLcDNA clones. In total, 3,204 non-redundant reference FLcDNAs (Figure 10 part 5) were characterized from the three FLcDNA libraries of rgf, rgg, and rgh.

Reference Full-Length cDNA assessment

The 5' UTRs, ORFs and 3' UTRs for the 9,057 reference clones were characterized and the results are summarized in Figure 11. The mean reference FLcDNA length for 9,057 *S. salar* sequences (Figure 11a) is 1,450 +/- 794 bp (mean +/- SD), and ranges from 267 to 4,730 bp. Of these sequences the mean 5' UTR and 3' UTR is 142 +/- 171 bp and 608 +/- 509 bp, respectively. The mean reference ORF is 755 +/- 499 bp.

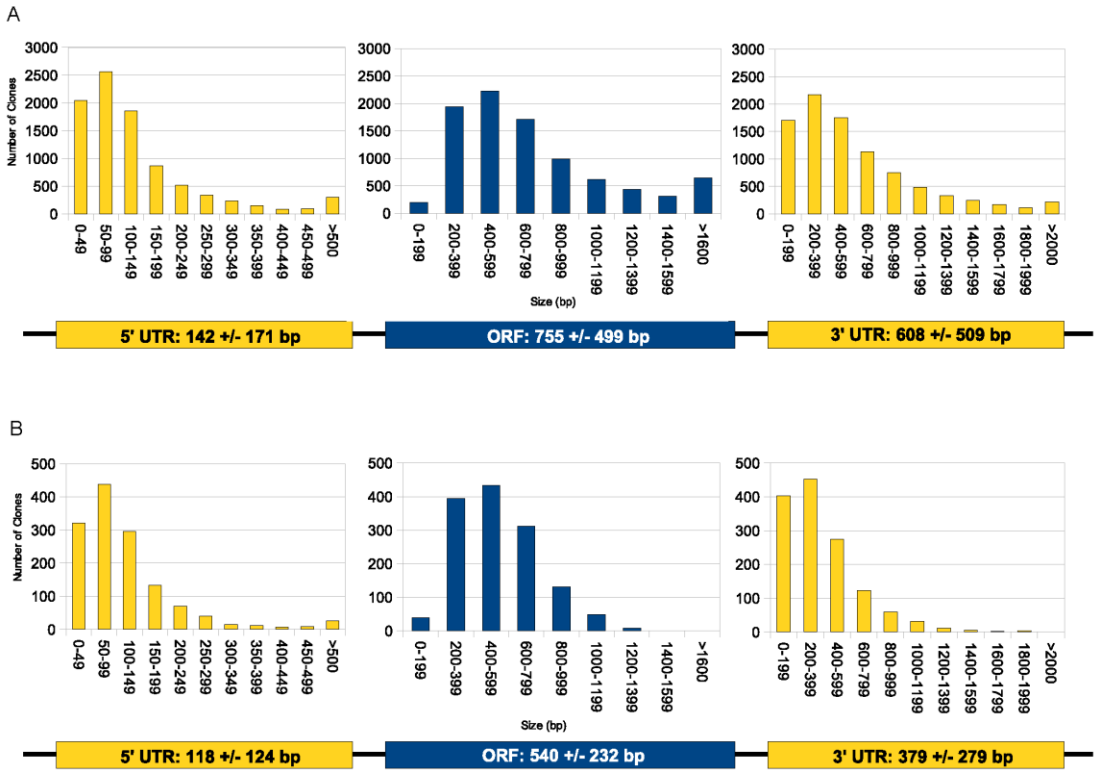

Figure 11. Distributions and means of ORF, 5' and 3' UTR sizes in reference FLcDNAs for (A) *S. salar* **(B)** *E. lucius*. Each reference FLcDNA, determined by in-house annotation methods, was examined for an ORF, 5' UTR, and 3' UTR. Means for each region were calculated (+/- standard deviation). An ORF is characterized by a start (ATG) and an in-frame stop codon (TGA, TAG, TAA). The 5' UTR is calculated as the entire area upstream of the start codon, while the 3' UTR is considered the entire area downstream of the stop codon. Any 3' polyA tails were masked and were not included in UTR length calculations.

Similar analysis of *E. lucius* reference FLcDNAs (Figure 11b) shows a mean length of 1,003 +/- 286 bp, ranging from 312 to 1,731 bp. Mean UTRs in the 5' and 3' regions are 118 +/- 124 bp and 379 +/- 279 bp, respectively. The mean reference ORF is 540 +/- 232 bp.

The UTR results are comparable with efforts from groups which have compiled mRNA UTR databases. In one such study of UTRs for a variety of species belonging to the 'fish' category, these species were shown to have an average 5' UTR of 107 bp, while 3' UTRs

averaged 397 bp (Pesole et al. 1996). These results indicate that the UTRs from this study are consistent with full-length sequences. While similar studies in this area are available for non-fish FLcDNAs, comparisons among more closely related organisms are lacking (Mignone et al. 2002).

A contig confirmation and uniqueness study was performed on each reference FLcDNA sequence set and the results are outlined in Table 6. A comparison was done on each corresponding contig set in an attempt to demonstrate which reference FLcDNAs could be confirmed by an existing contig. Reference FLcDNAs from *S. salar* full-length libraries were not included in the contig assembly and therefore possibly contain novel cDNA inserts. There were 6,115 reference FLcDNAs that could be confirmed by a contig sequence for *S. salar*. All *E. lucius* clones were included in its contig assembly. As a result, the entire 1,365 reference FLcDNA set could therefore be confirmed by a contig sequence.

Table 6. Summary of confirmed and unique reference FLcDNAs in contig sets for *S. salar* and *E. lucius*.

	Contigs	Reference FLcDNAs	Confirmed in Contigs	Unique
S. salar	81398	9057	6115	2942
E. lucius	11662	1365	1365	0
Total	93060	10422	7480	2942

Reference FLcDNAs were confirmed, using BLASTN, against their corresponding contig set. The remainder of the reference FLcDNAs are represented by clones that are similar to and consistent with SwissProt entries and are unique to the full-length libraries characterized in this study.

S. salar and *E. lucius* alignments

FLcDNA transcripts from *S. salar* were used to identify protein coding regions for an analysis of silent and amino acid changing substitution rates in duplicated genes. The

coding sequences were translated and used as queries in a TBLASTN comparison to the nucleotide database consisting of all *S. salar* and *E. lucius* EST assemblies. Contig sequences corresponding to the TBLASTN hits were organized into clusters, then translated, resulting in a cluster of nucleotide sequences and a corresponding cluster of protein translations for each full-length gene. A common region of alignment with respect to the translated ORF was found for the DNA sequences and the corresponding proteins based on length and quality of alignment criteria. A final screening process was performed to prevent allelic or distant homolog comparisons. 408 clusters contained the necessary one sequence from *E. lucius* and two sequences from *S. salar*.

Non-synonymous (d_N), synonymous (d_S) and ω (d_N/d_S) values were calculated for the 408 individual gene trees to investigate patterns of evolution. A value for ω that is < 1 over the alignment is indicative of purifying selection (the rate of amino acid changing substitutions is less than the rate of incorporation of synonymous mutations). A value for ω that is > 1 is indicative of diversifying or positive selection (Zhang et al. 2002).

Pairwise comparisons to determine d_S and ω

For each gene cluster, the corrected number of synonymous substitutions per synonymous site (d_S) was determined by comparing the *E. lucius* gene to each of the duplicate *S. salar* genes (gray and black lines) and the *S. salar* duplicate genes to each other (green line; Figure 12a). For each branch, the frequency of d_S values is plotted in Figure 12a. *S. salar* gene duplicate d_S values (green lines) have a median value of 0.192 and the *E. lucius* to *S. salar* d_S values (gray and black lines) have a median value of 0.434. This difference confirms that the salmonid genome duplication occurred more recently than the separation of *E. lucius* and *S. salar* lineages.

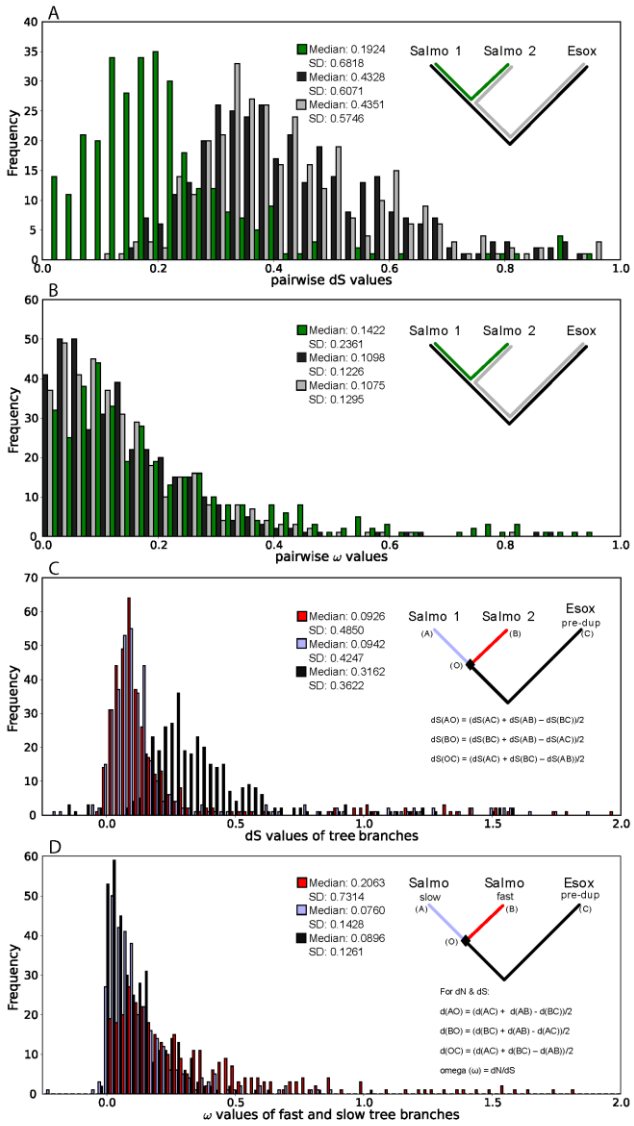

Figure 12. Frequencies of d_S and ω values for comparisons within *S. salar* and *E. lucius* gene trios. (A) Distributions of d_S values from pairwise comparisons within gene trios: between *S. salar* paralogs (green) and between each of the two *S. salar* paralogs and its corresponding *E. lucius* ortholog (gray and black). (B) Distributions of d_N/d_S ratios (ω) from pairwise comparisons within gene trios: between *S. salar* paralogs (green) and between each of the two *S. salar* paralogs and its corresponding *E. lucius* ortholog (gray and black). (C) Distributions of d_S values separated into individual tree branches based on gene trios. Values from pairwise comparisons were used to calculate silent substitution rates for periods before and after the salmonid tetraploidization event. The light blue curve represents frequencies of d_S values from the duplication event to one *S. salar* paralog, the red curve from the duplication event to the other paralog, and the black curve prior to the genome duplication to the *E. lucius* ortholog. (D) Distributions of d_N/d_S ratios separated into branches where one *S. salar* paralog, that which has the lower ω value, is considered to be a *slow* branch (light blue curve) and the other paralog (red curve) is considered to be more quickly diverging (*fast* branch for the purposes of labelling). The black curve displays frequencies of ω values between the *E. lucius* ortholog and the genome duplication.

The ratios of non-synonymous to synonymous substitution rates, or ω values, were in a similar manner calculated for each gene set and the frequency of ω values presented in Figure 12b. The median ω for *E. lucius* to each of the duplicate *S. salar* genes (gray and black lines) is 0.109 and the median ω for the duplicate *S. salar* genes (green line) is 0.142. The low ω values for all three sets of pairwise comparisons indicates an average of 7-9 synonymous substitutions for every non-synonymous substitution. This ratio confirms that purifying selection is the predominant evolutionary force in these genes. In most cases, both copies of these genes appear to have had their original functions retained based on the relatively low ratio of substitutions and high similarity between sequences. It is worth noting however, that poorly aligning and therefore potentially more divergent regions are trimmed from the overall alignments. Therefore, these estimates may be on the conservative side.

The paralogous comparisons (green line) produced ω values that were generally larger than the orthologous comparisons (gray and black lines). Upon using a Kruskal-Wallis test to compare distributions, both sets of orthologous comparisons were found to be significantly different from the paralogous comparisons (p-value = 1.671×10^{-5} and 2.359×10^{-5}) while orthologous sets were not significantly different with respect to each other (p-value = 0.9188). Therefore, while there is a large variance in the level of selection among the different genes, this result supports a small but significant relaxation in the level of selection pressure following gene duplication. This result is consistent with a comparison of 445 gene duplicates in the polyploid *Xenopus laevis* (Morin et al. 2006).

d_S and ω for tree segments

To more closely examine the effects of evolutionary pressures before and after duplication of the salmonid genome (represented by a diamond in the accompanying tree diagram), the substitution rates and ratios were separated into three tree segments (shown in Figure 12c, 12d). This subdivision was accomplished by using all three pairwise comparisons to calculate d_S values and ω values from the occurrence of the genome duplication (point O) to each extant gene (points A, B, and C).

This calculation provides an approximation for the number of substitutions before and after the duplication event. The post-duplication branches (in red and light blue) yielded median d_S values of 0.0926 and 0.0942 and the pre-duplication branch yielded a median d_s value of 0.3162. *E. lucius* sequences clearly diverge from the *S. salar* paralogs to a much greater extent than the *S. salar* paralogs diverge from each other. This again is consistent with the WGD occurring in salmonids but not in ecosids. The few values that are less than zero are presumably a result of the variation in the divergence of the three sequences with respect to one another and chance convergent substitutions.

To pursue the observation in Figure 12b showing a relaxation of selection following gene duplication (green line) and operating under the assumption that one duplicate could be retained and preserve its original function, freeing the other to diverge (Ohno 1970), the two post-duplicate branches were separated into two groups; one member of each pair represented the *slow* branch (i.e. the branch with the lower ω, indicating more purifying selection) and the other member represented the *fast* branch (i.e. with the higher ω, indicating possibly relaxed selection). While this method of sorting duplicate genes is somewhat arbitrary, the results can be reviewed with respect to an ancestral branch

leading to *E. lucius*. Selection values (ω) for the fast and slow duplicates along with the ancestral branch were calculated for each gene set and the frequency of ω values plotted in Figure 12d. In the slow *S. salar* branch (light blue) the median ω was 0.0760 and in the ancestral branch (black) the median ω was 0.0896. While the slow branch and pre-duplication branch do differ significantly from each other in terms of the means of the ranks of the data (Kruskal-Wallis test, P = 0.00623), the slow branch has a very similar median ω to the pre-duplication condition. The fast branch, on the other hand, has a much higher average ω (red curve, median = 0.2063) than both of the other branches. These results are consistent with the view that after the WGD there was little change in the evolutionary rate for one member of the pair; though in many cases, the rate of incorporation of non-synonymous changes increased dramatically for the other member. Despite a high level of variation, these data suggest that there is some asymmetry in evolutionary pressures on paralogs.

Gene Ontology analysis

In order to determine if there were ontological categories that were enriched for gene pairs that were subjected to more asymmetrical rates of evolution than others, Gene Ontology terms (The Gene Ontology Consortium 2000) were found for two groups of gene sets: those that had high fold differences (> 3x; n = 67) in the ω values found for the two *S. salar* branches and those that had low or no fold differences (<1.75x; n = 61). The results are shown in Table 7. For the most part, the two groups are populated by categories with similar proportions. A few notable exceptions include a larger proportion of genes involved in nucleic acid metabolic processes (GO:0006139) in the high fold-change group relative to the low. Likewise, the low fold-change group has a higher

proportion of 'other' metabolic processes (various IDs) such as lipid and carbohydrate metabolism. However, the relatively low number of genes in most categories limits the possibilities in this study of correlating ontological terms with specific patterns of evolution.

Table 7. Proportions of genes in GO categories

Categories	**ID**	**High fold-changes**		**Low fold-changes**	
localization (transport, cell motion)	GO:0051179	11	16.4%	10	16.4%
nucleic acid metabolism	GO:0006139	10	14.9%	3	4.9%
protein metabolism	GO:0019538	6	9.0%	6	9.8%
other metabolic process	Multiple IDs	8	11.9%	16	26.2%
development	GO:0032502	6	9.0%	5	8.2%
translation	GO:0006142	6	9.0%	4	6.6%
transcription	GO:0006350	5	7.5%	3	4.9%
apoptosis	GO:0006915	3	4.5%	5	8.2%
response to stimulus	GO:0050896	4	6.0%	2	3.3%
cell proliferation	GO:0008283	3	4.5%	2	3.3%
cell cycle	GO:0007049	3	4.5%	2	3.3%
signal transduction	GO:0007165	2	3.0%	3	4.9%
Total		67	100.0%	61	100.0%

Gene Ontology terms are given for gene trios in which there is a high fold-change in ω between paralogs (asymmetrically evolving duplicates) or a low fold-change in ω (symmetrically evolving duplicates). The proportions of genes contained within Gene Ontology categories are compared between the two groups. Nucleic acid metabolism (GO:0006139) is more highly represented in asymmetrically evolving gene duplicates.

Discussion

The objectives of this study were to: 1) characterize a large unambiguous set of reference gene sequences to compare with alleles and duplicates in *S. salar*, genes in other salmonid species, and genes in more distantly related fish species; 2) expand genomic resources for a representative member of the closest non-tetraploidized fish group (Esociformes: *E. lucius)* to provide a reference for the study of WGD in salmonids; and 3) identify patterns of change in the evolution of duplicated genes in the autotetraploid *S. salar*.

Genome duplications have a profound impact on the physiology, reproductive biology, ecology and evolution of a species. Salmonids (11 genera and 66 species) (Nelson 2006) are one of the most economically important and most studied groups of fish. A purported WGD in their common ancestor between 25-100 million years ago (Allendorf and Thorgaard 1984), after the separation from esocids, plays a prominent role in understanding the biology of this group. The different salmonid species are currently in the process of reverting to a stable diploid state through deletions and rearrangements (Ng et al. 2005; Danzmann et al. 2008; Koop et al. 2008; Phillips et al. 2009). In the absence of a completed genome, there is a significant problem in distinguishing among numerous duplicates, alleles and other very similar sequences. The integrity of genes resulting from assembling large numbers of partial mRNA sequences (ESTs) remains open to question. To resolve this problem, a collection of reference genes containing CDS and flanking UTRs coming from single, completely characterized cDNA clones provides an essential resource in gene identification, future genome annotation, and the study of evolutionary patterns.

An analysis of existing EST data from *S. salar* led to estimates of 10,026 FLcDNA contig consensus sequences. However, these contigs may represent an amalgamation of many unique transcript products with high similarity, rather than a single unique allele. 9,057 *S. salar* reference FLcDNA clones were determined in this study to resolve this issue. These FLcDNA sequences represent a significant community resource, adding to the current knowledge base on salmonid biology. These sequences can also serve as scaffolding with which to aid in genomic sequencing and creation of physical maps in other salmonids. The increasing popularity of microarrays in gene expression experiments has allowed for more precise control of probe design and information on full-length sequences enables probes to be optimally designed with higher specificity. An increase in probe-binding specificity reduces unwanted cross-species interactions. Salmonid research benefits from a fuller characterization of *S. salar* genes.

To expand evolutionary studies of salmonids, 29,221 *E. lucius* ESTs were obtained and when combined with the existing 3,612 EST sequences (32,833 total), 11,662 contigs and 1,365 FLcDNA reference sequences were identified. This resource not only provides an important initial genetic foundation for the study of pike throughout North America, Europe and Asia, but also provides essential information on a diploid reference species for the study of WGD in salmonids.

In this study, the FLcDNA sequences from *S. salar* along with homologous data from *E. lucius* were used to analyze evolutionary trends in some of the genes in the pseudotetraploid genome of *S. salar*. While the salmon genome may still be in the process of returning to a stable diploid state, it is evident that many gene duplicates have been retained. The peak in Figure 12c indicates a collection of genes that arose from a

duplication event after the separation of esocids from a salmonid ancestor. These genes are likely to be still active because the data for this study are based on mRNA, EST-derived sequences.

It is interesting to note that both Morin et al. (2006) and the present study started with approximately 10,000 full-length transcripts. By selecting a subset of sequence clusters that fulfil alignment and homology criteria, both studies ended up with only 400-450 gene sets. This ~4% yield is due in part to the strict criteria for usable sequences as well as the more limited *E. lucius* dataset from which to draw sequences. Further investigation should be undertaken to determine if both *X. laevis* and *S. salar* have retained similar proportions of gene duplicates, which would be of great interest in understanding responses to tetraploidization events. The numerous other species in the salmonid family provide an opportunity to facilitate finer analysis of the genome duplication, once additional data have been gathered for them. Moreover, the additional gene sets that contained more than two *S. salar* sequences in addition to the *E. lucius* ortholog (approximately 300 sets) could be studied to gain an understanding of some of the smaller scale duplication events or potentially more ancestral WGDs.

Over the last century, many individuals and groups have developed ideas about gene duplication in evolution and its importance in expanding on existing biological functions (Taylor and Raes 2004). A central model that has been strongly supported by Ohno (1970) states that a duplicate gene can accumulate mutations and become non-functional (non-functionalization) or diverge to a novel function (neo-functionalization) while the other duplicate keeps its original function. Other models have been proposed including sub-functionalization (Hughes 1994; Force et al. 1999), where both duplicates

accumulate mutations resulting in complementary expression, leaving each copy with its own sub-function. It is of interest to look for signatures of different types of selection in order to better understand models that may be directing the fates of one or both paralogs.

Based on the observation that the genes under investigation are in fact being transcribed to some degree in *S. salar*, it would be expected that purifying selection would be acting on both duplicates. The vast majority of ω values that are presented (Figure 12) are much less than one. It is apparent that negative selection is the predominant force in this evolutionary process as was also found in the similar analysis done by Morin et al. (2006). However, relative to the state of the genes before the duplication, there is significant relaxation of selective pressure on at least some of the paralogs, suggesting reduced constraints. This relaxation is consistent with the idea that having redundancy in the genome will result in increased freedom for divergence (Ohno 1970; Taylor and Raes 2004). These trends could facilitate neo-functionalization or modification of existing functionality taking place in some of the paralogs.

Data in this study provide evidence that selection constraints are not acting on both gene duplicates to the same extent. In a number of the 408 gene sets examined, one paralog may be relaxed, while it appears that the other is maintained to roughly the same degree as the pre-duplication single-copy gene. This asymmetrical pattern of evolution has recently been observed in specific Hox clusters in *S. salar* (Mungpakdee et al. 2008) as well as in earlier genome-wide studies in other organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans* (Conant and Wagner 2003).

The question that results from this observation centers around the fate of these duplicate genes that are operating under relaxed selection. The paralogs that were studied

are still being transcribed and are presumably functional (with the possible exception of some rarely transcribed pseudogenes) and have not been subject to non-functionalization. Duplicates that were deleted since the WGD would not be observed and neither would the presumably large number of duplicates that have become pseudogenes. Conclusive evidence for a general trend of positive selection was not found for the set of genes, since nearly all ω values were much less than one, though there were a few higher post-duplication values that suggested some duplicates may have been influenced by directional selection. The few genes that did have an ω value greater than one showed no enrichment for an ontological category (data not shown). Turunen et al. (2009) looked at asymmetrically evolving gene duplicates in yeast and found evidence for relaxation of selective pressure, sub-functionalization, and even neo-functionalization, though the average ω was significantly less than one. Therefore, it is not surprising that a strong signature of diversifying selection was not detected. Positive selection that may have occurred over a small region or short period of time could be masked by a larger overall pattern of negative selection. For example, once a neo-functionalization event has occurred, purifying selection would act to maintain that new function in the long term. Indeed, Hughes et al. (2000) reported 30-50 million years of divergence to be the upper limit of detection of positive selection in eukaryotes using d_N/d_S analysis. Looking at a variety of salmonid species in a comparative fashion could enable a higher resolution study of changes in evolutionary pressures and may provide more clues as to the events that took place in the duplicated genes soon after the tetraploidization. In addition, other groups (Hellsten et al. 2007; Chain et al. 2008) have studied polyploidization in *Xenopus* species using some alternative methods that may be applicable to *S. salar* in future

efforts. One example was using transversion rates at four-fold synonymous codon positions (4 DTv) to measure evolutionary divergence, though saturation of mutations at synonymous sites was not a problem for the present study.

Functional gene groups defined by Gene Ontology terms were found for *S. salar* gene duplicates that displayed either substantial or very small to no differences in selection constraints (i.e., evolving asymmetrically or symmetrically, respectively). The proportions of genes falling into the defined categories were generally quite similar (Table 7). However, one interesting result was the higher percentage of genes involved in nucleic acid metabolic processes (GO:0006139) (e.g., RNA processing and DNA metabolism) in the group of gene sets in which a large difference in selection constraints was identified. In this case, the conclusion that nucleic acid metabolism genes were more often present in the asymmetrical group than the symmetrical group would be consistent with earlier studies, which found that nucleic acid processing and nucleoside metabolism functional groups were selectively lost after whole genome duplications in *X. laevis* and *A. thaliana*. This suggests that nucleic acid processing and nucleoside metabolism functional groups of genes may have a greater chance of conferring dosage sensitivity (Maere et al. 2005; Morin et al. 2006).

Chapter 4
Conclusions

The work presented here describes gene characterization for Atlantic salmon (*Salmo salar*), derived from EST and full-length cDNA data, the most complete of any such study for this species. Although a seemingly straightforward endeavour, *in silico* transcriptome assembly and annotation of a recently duplicated species has many challenges. To further examine the duplicate genes in Atlantic salmon, genes were also characterized for northern pike (*Esox lucius*), a non-duplicated sister species. For Atlantic salmon, raw EST data was gathered from numerous individuals across multiple tissue types and lifestages, in the hope of providing an extensive representation of its transcriptome. To a lesser extent, EST data from northern pike was also gathered. Completion of this large-scale project required development of many novel computational methods and workflows.

For this study, analysis of EST and full-length cDNA data included multiple steps – sequencing, storage, curation, manipulation, assembly, annotation, gene-prediction, gene-comparison, and the sharing of results. Each of these steps has unique needs and required well thought-out solutions. All visualization software was developed in-house for the specific purpose of contributing results to the scientific and to the public community. Generalized software for the visualization of transcriptomes is available, however adequate solutions for displaying the intricacies of the recently duplicated genome of Atlantic salmon did not previously exist. Therefore, novel software solutions were created. Recently, more general genome browser systems have become available from Ensembl. The results generated in this thesis support implementation in such browsers,

as the prospect of a whole genome sequence using next-generation sequencing (NGS) methods for Atlantic salmon gets closer to reality.

Representing a milestone in genetic research, salmonid resources have benefitted a great deal from the efforts of this project. Atlantic salmon is currently the 21st most represented species for EST resources, with more than half of those sequences (298,304) derived from this project. In total, 434,384 Atlantic salmon ESTs sequenced in-house and from NCBI formed the basis of the work. The assembly of ESTs has identified 81,398 unique transcripts, with 17,399 being full-length gene assemblies (Koop et al. 2008), in Atlantic salmon's recently duplicated genome. In addition to sequencing support, a novel full-length cDNA prediction algorithm was created and integrated into the EST assembly pipeline. Subsequent identification of unique transcripts and full-length assemblies led to the characterization of 9,057 full-length reference genes from single clones in Atlantic salmon. These genes represent the largest full-length study for this species to date and provide an important resource for identifying reference alleles and gene family members. Moreover, reference full-length cDNA acts as a method to verify the accuracy of EST assemblies, as these sequences are derived unambiguously from a single clone. Using the same pipeline, 1,365 full-length reference genes in northern pike were also found. Comparison of the full-length genes between these two closely related sister taxa has provided insight into molecular evolution patterns. Using two paralogs from Atlantic salmon and the corresponding ortholog from northern pike, it was discovered that gene duplicates of Atlantic salmon showed asymmetrical evolutionary pressure and that purifying selection was the predominant force acting on those genes.

In parallel to the study of Atlantic salmon, resources for other salmonid species were created. EST sequencing and analysis for 48,653 rainbow trout (*Oncorhynchus mykiss*), 14,535 chinook salmon (*Oncorhynchus tshawytscha*), 12,056 sockeye salmon (*Oncorhynchus nerka*), 10,051 brook trout (*Salvelinus fontinalis*), 10,975 grayling (*Thymallus thymallus*), and 10,842 lake whitefish (*Coregonus clupeaformis*) ESTs was completed. 32,908 ESTs for northern pike (*Esox lucius*) were also sequenced and analyzed. As part of my thesis, a total of 438,324 in-house ESTs were sequenced and analyzed in-depth. The results of this work offer a major genetic resource for salmonid biological studies. Polymorphic marker design, and microarray development have already benefited from the results of the work presented here, and will continue to do so.

A novel web-based contig viewer was developed to efficiently and effectively visualize the EST assemblies and their consensus sequences. Contigs are categorized by size, and detailed information about alignments, assembly statistics, full-length gene predictions, open reading frame (ORF) predictions, and detailed protein annotations are shown (Figure 2, Koop et al. 2008). In addition, this viewer allows the data to be downloaded for independent evaluation. All relevant information has been made available to the public through this gateway. Access to the contig viewer has been well-received (Appendix A). Feedback from regular users of this gateway has been positive, and remains the most flexible method in which to share the results of the study.

By providing a solid foundation for salmonid genomics research, enhanced by an effective gateway to share results, the work presented provides for the next steps in annotation of the Atlantic salmon genome sequencing project.

Bibliography

Adzhubei, A., Vlasova, A., Hagen-Larsen, H., Ruden, T., Laerdahl, J. and Hoyheim, B. (2007), 'Annotated expression sequence tags (ESTs) from pre-smolt Atlantic salmon (*Salmo salar*) in a searchable data resource' *BMC Genomics* **8**, 209.

Allendorf, F. and Thorgaard, G. (1984), 'Tetraploidy and the evolution of salmonid fishes' *Evolutionary Genetics of Fishes*, 1 - 53.

Allendorf, F., Utter, F. and May, B. (1975), 'Gene duplication within the family *Salmonidae*. II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations' *Genetics and Evolution* **4**, 414-432.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs' *Nucleic Acids Res* **25**(17), 3389 - 3402.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M. and Postlethwait, J. H. (1998), 'Zebrafish hox clusters and vertebrate genome evolution' *Science* **282**(5394), 1711-1714.

Andreassen, R., Lunner, S. and Hoyheim, B. (2009), 'Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (*Salmo salar*)' *BMC Genomics* **10**, 502.

Bairoch, A. and Apweiler, R. (1998), 'The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998' *Nucleic Acids Res* **26**(1), 38 - 42.

Bell, J., McEvoy, J., Tocher, D., McGhee, F., Campbell, P. and Sargent, J. (2001), 'Replacement of fish oil with rapeseed oil in diets of Atlantic salmon (*Salmo salar*) affects tissue lipid compositions and hepatocyte fatty acid metabolism' *J Nutr* **131**(5), 1535 - 1543.

Bensasson, D., Feldman, M. W. and Petrov, D. A. (2003), 'Rates of DNA duplication and mitochondrial DNA insertion in the human genome' *J Mol Evol* **57**(3), 343-354.

Bernatchez, L. and Landry, C. (2003), 'MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years?' *J Evolutionary Biol* **16**(3), 363 - 377.

Bisbee, C. A., Baker, M. A., Wilson, A. C., Hadjiazimi, I. and Fischberg, M. (1977), 'ALBUMIN PHYLOGENY FOR CLAWED FROGS (XENOPUS)' *Science* **195**(4280), 785-787.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S. and Peer, Y. (2006), 'The gain and loss of genes during 600 million years of vertebrate evolution' *Genome Biol* **7**, R43.41 - 12.

Boeuf, G. and Le Bail, P. (1999), 'Does light have an influence on fish growth?' *Aquaculture* **177**(1-4), 129 - 152.

Boguski, M., Lowe, T. and Tolstoshev, C. (1993), 'dbEST - database for "expressed sequence tags"' *Nat Genet* **4**, 332 - 333.

Briscoe, A. D. (2001), 'Functional diversification of lepidopteran opsins following gene duplication' *Mol Biol Evol* **18**(12), 2270-2279.

Brown, G. D. (2008), 'An Analysis of Salmonid RNA Sequences and Implications for Salmonid Evolution' Journal Volume(Issue), Pages.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004), 'The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology' *Nucl Acids Res* **32**(suppl_1), D262 - 266.

Chain, F., Ilieva, D. and Evans, B. (2008), 'Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization' *BMC Evol Biol* **8**, 43.

Computational Biology and Functional Genomics Laboratory, Access Date: 01-10-2010, URL: http://compbio.dfci.harvard.edu/tgi/tgipage.html

Conant, G. and Wagner, A. (2003), 'Asymmetric sequence divergence of duplicate genes' *Genome Res* **13**, 2052 - 2058.

Consortium, T. G. O. (2000), 'Gene ontology: tool for the unification of biology' *Nat Genet* **25**(1), 25 - 29.

Crespi, B. and MJ, F. (2004), 'Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny' *Mol Phylogenet Evol* **31**, 658 - 679.

Danzmann, R., Cairney, M., Davidson, W., Ferguson, M., Gharbi, K., Guyomard, R., Holm, L.-E., Leder, E., Okamoto, N., Ozaki, A., Rexroad, C., Sakamoto, T., Taggart, J. and Woram, R. (2006), 'A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae)' *Genome* **48**, 1037 - 1051.

Danzmann, R., Davidson, E., Ferguson, M., Gharbi, K., Koop, B., Hoyheim, B., Lien, S., Lubieniecki, K., Moghadam, H., Park, J., Phillips, R. and Davidson, W. (2008),

'Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon)' *BMC Genomics* **9**, 557.

de Boer, J. G., Yazawa, R., Davidson, W. S. and Koop, B. F. (2007), 'Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids' *BMC Genomics* **8**.

Derome, N., Duchesne, P. and Bernatchez, L. (2006), 'Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis* Mitchill) ecotypes' *Mol Ecol* **15**(5), 1239 - 1249.

Doolittle, W. F. and Sapienza, C. (1980), 'SELFISH GENES, THE PHENOTYPE PARADIGM AND GENOME EVOLUTION' *Nature* **284**(5757), 601-603.

Ewing, B. and Green, P. (1998), 'Base-calling of automated sequencer traces using PHRED. II. Error probabilities' *Genome Res* **8**, 186 - 194.

Ewing, B., Hillier, L., Wendl, M. and Green, P. (1998), 'Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment' *Genome Res* **8**(3), 175 - 185.

Fares, M., Byrne, K. and Wolfe, K. (2006), 'Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species' *Mol Biol Evol* **23**(2), 245 - 253.

Felsenstein, J. (2004), 'PHYLIP (Phylogeny Inference Package) version 3.6' *Distributed by the author Department of Genome Sciences*.

Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. and Dujon, B. (2001), 'Evolution of gene order in the genomes of two related yeast species' *Genome Res* **11**(12), 2009-2019.

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

Garant, D., Dodson, J. and Bernatchez, L. (2000), 'Ecological determinants and temporal stability of the within-river population structure in Atlantic salmon (*Salmo salar* L.)' *Mol Ecol* **9**(5), 615 - 628.

Gerstein, A. C. and Otto, S. P. (2009), 'Ploidy and the Causes of Genomic Evolution' *Journal of Heredity* **100**(5), 571-581.

Gorbunova, V. and Levy, A. A. (1997), 'Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions' *Nucl Acids Res* **25**(22), 4650-4657.

Govoroun, M., Le Gac, F. and Guiguen, Y. (2006), 'Generation of a large scale repertoire of Expressed Sequence Tags (ESTs) from normalized rainbow trout cDNA libraries' *BMC Genomics* **7**, 196.

GRASP website, Access Date: 01-10-2010, URL: http://www.uvic.ca/grasp

Graur, D. and Li, W.-H. (2000), *Fundamentals of molecular evolution* Sunderland, MA, Sinauer Associates.

Green, P. 'Documentation for PHRAP'.

Gregory, T. (2002), 'Animal Genome Size Database'.

Grimholt, U., Larsen, S., Nordmo, R., Midtlyng, P., Kjoeglum, S., Storset, A., Saebo, S. and Stet, R. (2003), 'MHC polymorphism and disease resistance in Atlantic salmon (*Salmo salar*); facing pathogens with single expressed major histocompatibility class I and class II loci' *Immunogenetics* **55**(4), 210 - 219.

Gusfield, D. (1999), 'Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology'.

Hample, F. (1974), 'The influence cure and its role in robust estimation' *J Amer Stat Assoc* **69**, 383 - 393.

Handeland, S., Berge, A., Bjornsson, B. and Stefansson, S. (1998), 'Effects of temperature and salinity on osmoregulation and growth of Atlantic salmon (*Salmo salar* L.) smolts in seawater' *Aquaculture* **168**, 289 - 302.

Harstad, H., Lukacs, M., Bakke, H. and Grimholt, U. (2008), 'Multiple expressed MHC class II loci in salmonids; details of one non-classical region in Atlantic salmon (*Salmo salar*)' *BMC Genomics* **9**, 193.

Hebert, P. D. N., Cywinska, A., Ball, S. L. and DeWaard, J. R. (2003), 'Biological identifications through DNA barcodes' *Proceedings of the Royal Society of London Series B-Biological Sciences* **270**(1512), 313-321.

Hellsten, U., Khokha, M., Grammer, T., Harland, R., Richardson, P. and Rokhsar, D. (2007), 'Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*' *BMC Biol* **5**, 31.

Hilborn, R., Orensanz, J. M. and Parma, A. M. (2005), 'Institutions, incentives and the future of fisheries' *Philos Trans R Soc B-Biol Sci* **360**(1453), 47-57.

Hoegg, S. and Myer, A. (2005), 'Hox clusters as models for vertebrate genome evolution' *Trends Genet* **21**, 421 - 424.

Food and Agriculture Organization (2009), 'How to Feed the World in 2050', United Nations.

Huang, X. and Madan, A. (1999), 'CAP3: A DNA sequence assembly program' *Genome Res* **9**, 868 - 877.

Hudson, T. J., Engelstein, M., Lee, M. K., Ho, E. C., Rubenfield, M. J., Adams, C. P., Housman, D. E. and Dracopoli, N. C. (1992), 'ISOLATION AND CHROMOSOMAL ASSIGNMENT OF 100 HIGHLY INFORMATIVE HUMAN SIMPLE SEQUENCE REPEAT POLYMORPHISMS' *Genomics* **13**(3), 622-629.

Hufton, A. L. and Panopoulou, G. (2009), 'Polyploidy and genome restructuring: a variety of outcomes' *Current Opinion in Genetics & Development* **19**(6), 600-606.

Hughes, A. (1994), 'The evolution of functionally novel proteins after gene duplication' *P Roy Soc B-Biol Sci* **256**, 119 - 124.

Hughes, A., Green, J., Garbayo, J. and Roberts, R. (2000), 'Adaptive diversification within a large family of recently duplicated, placentally expressed genes' *P Natl Acad Sci USA* **97**(7), 3319 - 3323.

Hutchings, J. and Jones, M. (1998), 'Life history variation and growth rate thresholds for maturity in Atlantic salmon, *Salmo salar*' *Can J Fish Aquat Sci* **55**(Suppl 1), 22 - 47.

Invitrogen Full-Length cDNA Library Construction, Access Date: 01-10-2010, URL: http://www.invitrogen.com/site/us/en/home/Products-and-Services/Services/Molecular-Biology-Services/Library-Construction/Full-Length.html

Ishiguro, N., Miya, M. and Nishida, M. (2003), 'Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii"' *Mol Phylogenet Evol* **27**, 476 - 488.

Jacobs, M., Covaci, A. and Schepens, P. (2002), 'Investigation of selected persistent organic pollutants in farmed Atlantic salmon (*Salmo salar*), salmon aquaculture feed, and fish oil components of the feed' *Environ Sci Technol* **36**(13), 2797 - 2805.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D.,

Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, R., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J. P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J. N., Guigo, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J. and Crollius, H. R. (2004), 'Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype' *Nature* **431**(7011), 946-957.

Jorgensen, S., Lyng-Syvertsen, B., Lukacs, M., Grimholt, U. and Gjoen, T. (2006), 'Expression of MHC class I pathway genes in response to infectious salmon anaemia virus in Atlantic salmon (*Salmo salar* L.) cells' *Fish Shellfish Immun* **21**(5), 548 - 560.

Katju, V. and Lynch, M. (2003), 'The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome' *Genetics* **165**(4), 1793-1803.

Katju, V. and Lynch, M. (2006), 'On the formation of novel genes by duplication in the caenorhabditis elegans genome' *Mol Biol Evol* **23**(5), 1056-1067.

Kimura, M. and Ohta, T. (1969), 'AVERAGE NUMBER OF GENERATIONS UNTIL FIXATION OF A MUTANT GENE IN A FINITE POPULATION' *Genetics* **61**(3), 763-&.

King, T., Kalinowski, S., Schill, W., Spidle, A. and Lubinski, B. (2001), 'Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation' *Mol Ecol* **10**(4), 807 - 821.

Klemetsen, A., Amundsen, P., Dempson, J., Jonsson, B., Jonsson, N., Connell, M. and Mortensen, E. (2003), 'Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories' *Ecol Freshw Fish* **12**(1), 1 - 59.

Koop, B. F., von Schalburg, K. R., Leong, J., Walker, N., Lieph, R., Cooper, G. A., Robb, A., Beetz-Sargent, M., Holt, R. A., Moore, R., Brahmbhatt, S., Rosner, J., Rexroad, C. E., McGowan, C. R. and Davidson, W. S. (2008), 'A salmonid EST genomic study: genes, duplications, phylogeny and microarrays' *BMC Genomics* **9**, 16.

Krogdahl, A., Hemre, G. and Mommsen, T. (2006), 'Carbohydrates in fish nutrition: digestion and absorption in postlarval stages' *Aquacult Nutr* **11**(2), 103 - 122.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H. M., Yu, J., Wang, J., Huang, G. Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S. Z., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H. Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W. H., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J. R., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J. and Int Human Genome Sequencing, C. (2001), 'Initial sequencing and analysis of the human genome' *Nature* **409**(6822), 860-921.

Landry, C., Garant, D., Duchesne, P. and Bernatchez, L. (2001), "Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*)' *P Roy Soc B-Biol Sci* **268**(1473), 1279 - 1285.

Lees, F., Baillie, M., Gettinby, G. and Revie, C. W. (2008), 'The Efficacy of Emamectin Benzoate against Infestations of Lepeophtheirus salmonis on Farmed Atlantic Salmon (Salmo salar L) in Scotland, 2002-2006' *PLoS One* **3**(2).

Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., Munro, S., Moore, R., Holt, R. A., Jones, S. J. M., Davidson, W. S. and Koop, B. F. (2010), '*Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome' *BMC Genomics* **11**, 17.

Li, C., Lu, G. and Orti, G. (2008), 'Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci' *Syst Biol* **57**(4), 519 - 539.

Lilly, G. R. (2008). The Decline, Recovery, and Collapse of Atlantic Cod (Gadus morhua) off Labrador and Eastern Newfoundland. Resiliency of Gadid Stocks to Fishing and Climate Change. G. H. Kruse, K. Drinkwater, J. N. Ianelliet al. Fairbanks, Alaska Sea Grant Coll Program. **24:** 67-88.

Lin, Y. F. and Waldman, A. S. (2001), 'Capture of DNA sequences at double-strand breaks in mammalian chromosomes' *Genetics* **158**(4), 1665-1674.

Lin, Y. F. and Waldman, A. S. (2001), 'Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA' *Nucl Acids Res* **29**(19), 3975-3981.

Lopez, J., Chen, W. and Orti, G. (2004), 'Esociform phylogeny' *Copeia* **3**, 449 - 464.

Lukacs, M., Harstad, H., Grimholt, U., Beetz-Sargent, M., Cooper, G., Reid, L., Bakke, H., Phillips, R., Miller, K., Davidson, W. and Koop, B. (2007), 'Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (*Salmo salar*)' *BMC Genomics* **8**, 251.

Lynch, M. (2007), *The origins of genome architecture* Sunderland, MA, Sinauer Associates.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Peer, Y. (2005), 'Modeling gene and genome duplications in eukaryotes' *P Natl Acad Sci USA* **102**(15), 5454 - 5459.

Mank, J. and Avise, J. (2006), 'Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes' *Genetica* **127**, 321 - 327.

Maside, X., Assimacopoulos, S. and Charlesworth, B. (2000), 'Rates of movement of transposable elements on the second chromosome of Drosophila melanogaster' *Genet Res* **75**(3), 275-284.

Maside, X., Bartolome, C., Assimacopoulos, S. and Charlesworth, B. (2001), 'Rates of movement and distribution of transposable elements in Drosophila melanogaster: in situ hybridization vs Southern blotting data' *Genetics Research* **78**(2), 121-136.

McGinnity, P., Prodohl, P., Ferguson, K., Hynes, R., O'Maoileidigh, N., Baker, N., Cotter, D., O'Hea, B., Cooke, D., Rogan, G., Taggart, J. and Cross, T. (2003), 'Fitness reduction and potential extinction of wild populations of Atlantic salmon, *Salmo salar*, as a result of interactions with escaped farm salmon' *P Roy Soc B-Biol Sci* **270**(1532), 2443 - 2450.

McKay, S., Trautner, J., Smith, M., Koop, B. and Devlin, R. (2004), 'Evolution of duplicated growth hormone genes in autotetraploid salmonid fishes' *Genome* **47**, 714 - 723.

Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002), 'Untranslated regions of mRNAs' *Genome Biol* **3**(3), 1 - 10.

Min, X., Butler, G., Storms, R. and Tsang, A. (2005), 'TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences' *Nucl Acids Res* **33**, W669 - 672.

Miyata, T. and Yasunaga, T. (1980), 'Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application' *J Mol Evol* **16**, 23 - 36.

Moghadam, H., Ferguson, M. and Danzmann, R. (2005), 'Evidence for Hox Gene Duplication in rainbow trout (*Oncorhynchus mykiss*): A tetraploid model species' *J Mol Evol* **61**, 804 - 818.

Mommsen, T. and Vijayan MM Moon, T. (1999), 'Cortisol in teleosts: dynamics, mechanisms of action, and metabolic regulation' *Rev Fish Biol Fisher* **9**(3), 211 - 268.

Moore, A., Scott, A., Lower, N., Katsiadaki, I. and Greenwood, L. (2003), 'The effects of 4-nonylphenol and atrazine on Atlantic salmon (*Salmo salar* L) smolts' *Aquaculture* **222**, 253 - 263.

Morin, R., Chang, E., Petrescu, A., Liao, N., Griffith, M., Kirkpatrick, R., Butterfield, Y., Young, A., Stott, J., Barber, S., Babakaiff, R., Dickson, M., Matsuo, C., Wong, D., Yang, G., Smailus, D., Wetherby, K., Kwong, P., Grimwood, J., Brinkley, C., Brown-John, M., Reddix-Dugue, N., Mayo, M., Schmutz, J., Beland, J., Park, M.,

Gibson, S., Olson, T., Bouffard, G., Tsai, M., Featherstone, R., Chand, S., Siddiqui, A., Jang, W., Lee, E., Klein, S., Blakesley, R., Zeeberg, B., Narasimhan, S., Weinstein, J., Pennacchio, C., Myers, R., Green, E., Wagner, L., Gerhard, D., Marra, M., Jones, S. and Holt, R. (2006), 'Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling' *Genome Res* **16**, 796 - 803.

Mos, L., Cooper, G., Serben, K., Cameron, M. and Koop, B. (2008), 'Effects of diesel on survival, growth, and gene expression in rainbow trout (*Oncorhynchus mykiss*) fry' *Environ Sci Technol* **42**, 2656 - 2662.

Mungpakdee, S., Seo, H., Angotzi, A., Dong, X., Akalin, A. and Chourrout, D. (2008), 'Differential evolution of the 13 Atlantic salmon Hox clusters' *Mol Biol Evol* **25**(7), 1333 - 1343.

NCBI Blast, Access Date: 01-10-2010, URL: http://blast.ncbi.nlm.nih.gov

NCBI Entrez Gene database, Access Date: 01-10-2010, URL: www.ncbi.nlm.nih.gov/Entrez/

NCBI Unigene database, Access Date: 01-10-2010, URL: http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

Nelson, J. (2006), *Fishes of the world*, John Wiley & Sons, New York.

Ng, S., Artieri, C., Bosdet, I., Chiu, R., Danzmann, R., Davidson, W., Ferguson, M., Fjell, C., Hoyheim, B., Jones, S., de Jong, P., Koop, B., Krzywinski, M., Lubieniecki, K., Marra, M., Mitchell, L., Mathewson, C., Osoegawa, K., Parisotto, S., Phillips, R., Rise, M., von Schalburg, K., Schein, J., Shin, H., Siddiqui, A., Thorsen, J., Wye, N., Yang, G. and Zhu, B. (2005), 'A physical map of the genome of Atlantic salmon, *Salmo salar*' *Genomics* **86**, 396 - 404.

Norris, A., Bradley, D. and Cunningham, E. (1999), 'Microsatellite genetic variation between and within farmed and wild Atlantic salmon (*Salmo salar*) populations' *Aquaculture* **180**(3-4), 247 - 264.

Oakley, T. and Phillips, R. (1999), 'Phylogeny of Salmonine fishes based upon growth hormone introns: Atlantic (*Salmo*) and Pacific (*Oncorhynchus*) salmon are not sister taxa' *Mol Phylogenet Evol* **11**, 381 - 393.

Ohno, S. (1970), *Evolution by gene duplication*, New York: Springer-Verlag.

Ohno, S., Wolf, U. and Atkin, N. B. (1968), 'EVOLUTION FROM FISH TO MAMMALS BY GENE DUPLICATION' *Hereditas-Genetiskt Arkiv* **59**(1), 169-&.

Orgel, L. E. and Crick, F. H. C. (1980), 'SELFISH DNA - THE ULTIMATE PARASITE' *Nature* **284**(5757), 604-607.

Osinov, A. and Lebedev, V. (2000), 'Genetic divergence and phylogeny of the Salmoninae based on allozyme data' *J Fish Biology* **57**, 354 - 381.

Panopoulou, G. and Poustka, A. (2005), 'Timing and mechanism of ancient vertebrate genome duplications - the adventure of a hypothesis' *Trends Genet* **21**, 559 - 567.

Pesole, G., Grillo, G. and Liuni, S. (1996), 'Databases of mRNA untranslated regions for metazoa' *Computers Chem* **20**(1), 141 - 144.

Phillips, R., Keatley, K., Morasch, M., Ventura, A., Lubieniecki, K., Koop, B., Danzmann, R. and Davidson, W. (2009), 'Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*)' *BMC Genet* **10**, 46.

Phillips, R. and Rab, P. (2001), 'Chromosome evolution in the Salmonidae (Pisces): an update' *Biol Res* **76**, 1 - 25.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001), 'The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species' *Nucl Acids Res* **29**, 159 - 164.

Ramsden, S., Brinkmann, H., Hawryshyn, C. and Taylor, J. (2003), 'Mitogenomics and the sister of Salmonidae' *Trends Ecol & Evol* **18**, 607 - 610.

Rexroad, C., Lee, Y., Keele, J., Karamycheva, S., Brown, G., Koop, B., Gahr, S., Palti, Y. and Quackenbush, J. (2003), 'Sequence analysis of a rainbow trout cDNA library and creation of a gene index' *Cytogenet Genome Res* **102**(1-4), 347 - 354.

Ricchetti, M., Fairhead, C. and Dujon, B. (1999), 'Mitochondrial DNA repairs double-strand breaks in yeast chromosomes' *Nature* **402**(6757), 96-100.

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

Saksida, S. M., Morrison, D. and Revie, C. W. (2010), 'The efficacy of emamectin benzoate against infestations of sea lice, Lepeophtheirus salmonis, on farmed Atlantic salmon, Salmo salar L., in British Columbia' *J Fish Dis* **33**(11), 913-917.

Sato, Y. and Nishida, M. (2010), 'Teleost fish with specific genome duplication as unique models of vertebrate evolution' *Environmental Biology of Fishes* **88**(2), 169-188.

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A. and Webster, M. S. (2010), 'Population diversity and the portfolio effect in an exploited species' *Nature* **465**(7298), 609-U102.

Semon, M. and Wolfe, K. H. (2007), 'Rearrangement rate following the whole-genome duplication in teleosts' *Mol Biol Evol* **24**(3), 860-867.

Smit, A., Hubley, R. and Green, P. (1996), 'RepeatMasker Open - 3.0'.

Spaethe, J. and Briscoe, A. D. (2004), 'Early duplication and functional diversification of the opsin gene family in insects' *Mol Biol Evol* **21**(8), 1583-1594.

Food and Agriculture Organization (2008), 'The State of World Fisheries and Aquaculture 2008', United Nations.

Stearley, R. and Smith, G. (1993), 'Phylogeny of the Pacific trouts and salmon (*Oncorhynchus*) and genera of the family Salmonidae' *Trans Am Fish Soc* **122**, 1 - 36.

Steinke, D., Salzburger, W. and Meyer, A. (2006), 'Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs' *J Mol Evol* **62**, 772 - 784.

Sutton, S., Bult, T. and Haedrich, R. (2000), 'Relationships among fat weight, body weight, water weight, and condition factors in wild Atlantic salmon parr' *T Am Fish Soc* **129**(2), 527 - 538.

Taylor, J. S. and Raes, J. (2004), 'Duplication and divergence: The evolution of new genes and old ideas' *Annu Rev Genet* **38**, 615-643.

Taylor, J. S., Van de Peer, Y., Braasch, I. and Meyer, A. (2001), 'Comparative genomics provides evidence for an ancient genome duplication event in fish' *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **356**(1414), 1661-1679.

Thomas, E. E., Srebro, N., Sebat, J., Navin, N., Healy, J., Mishra, B. and Wigler, M. (2004), 'Distribution of short paired duplications in mammalian genomes' *P Natl Acad Sci USA* **101**(28), 10349-10354.

Thomas, J. H. (2006), 'Concerted evolution of two novel protein families in Caenorhabditis species' *Genetics* **172**(4), 2269-2281.

Thompson, J., Higgins, D. and Gibson, T. (1994), 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice' *Nucleic Acids Res* **22**(22), 4673 - 4680.

Thorgaard, G. H., Bailey, G. S., Williams, D., Buhler, D. R., Kaattari, S. L., Ristow, S. S., Hansen, J. D., Winton, J. R., Bartholomew, J. L., Nagler, J. J., Walsh, P. J., Vijayan, M. M., Devlin, R. H., Hardy, R. W., Overturf, K. E., Young, W. P., Robison, B. D., Rexroad, C. and Palti, Y. (2002), 'Status and opportunities for genomics research with rainbow trout' *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **133**(4), 609-646.

Turunen, O., Seelke, R. and Macosko, J. (2009), 'In silico evidence for functional specialization after genome duplication in yeast' *FEMS Yeast Res* **9**, 16 - 31.

Vandepoele, K., De Vos, W., Taylor, J., Meyer, A. and Peer, Y. (2004), 'Major events in the genome evolution of vertebrates: Paranome age and size differs considerably between ray-finned fishes and land vertebrates' *Proc Natl Acad Sci USA* **101**, 1638 - 1643.

Volff, J. (2005), 'Genome evolution and biodiversity in teleost fish' *Heredity* **94**, 280 - 294.

von Schalburg, K., Rise, M., Cooper, G., Brown, G., Gibbs, A., Nelson, C., Davidson, W. and Koop, B. (2005), 'Fish and chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray' *BMC Genomics* **6**, 126 - 133.

von Schalburg, K., Yazawa, R., de Boer, J., Lubieniecki, K., Goh, B., Straub, C., Beetz-Sargent, M., Robb, A., Davidson, W., Devlin, R. and Koop, B. (2008), 'Isolation, characterization and comparison of Atlantic and Chinook salmon growth hormone 1 and 2' *BMC Genomics* **9**, 522.

von Schalburg, K. R., Leong, J., Cooper, G. A., Robb, A., Beetz-Sargent, M. R., Lieph, R., Holt, R. A., Moore, R., Ewart, K. V., Driedzic, W. R., ten Hallers, B. F. H., Zhu, B., de Jong, P. J., Davidson, W. S. and Koop, B. F. (2008), 'Rainbow smelt (*Osmerus mordax*) genomic library and EST resources' *Mar Biotechnol* **10**(5), 487-491.

Wernersson, R. and Pedersen, A. (2003), 'RevTrans: multiple alignment of coding DNA from aligned amino acid sequences' *Nucleic Acids Res* **31**(13), 3537 - 3539.

Department of Economic and Social Affairs, Populaton Division (2009), 'World Population Prospects', United Nations.

Yang, Z. and Nielsen, R. (2000), 'Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models' *Mol Biol Evol* **17**(1), 32 - 43.

Yazawa, R., Cooper, G., Beetz-Sargent, M., Robb, A., McKinnel, L., Davidson, W. and Koop, B. (2008), 'Functional adaptive diversity of the Atlantic salmon T-cell receptor gamma locus' *Mol Immunol* **45**(8), 2150 - 2157.

Yazawa, R., Yasuike, M., Leong, J., von Schalburg, K. R., Cooper, G. A., Beetz-Sargent, M., Robb, A., Davidson, W. S., Jones, S. R. M. and Koop, B. F. (2008), 'EST and Mitochondrial DNA Sequences Support a Distinct Pacific Form of Salmon Louse, *Lepeophtheirus salmonis*' *Mar Biotechnol* **10**(6), 741-749.

Yu, X. and Gabriel, A. (1999), 'Patching broken chromosomes with extranuclear cellular DNA' *Mol Cell* **4**(5), 873-881.

Zhang, L., Vision, T. and Gaut, B. (2002), 'Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*' *Mol Biol Evol* **19**(9), 1464 - 1473.

Zhang, L. Q., Lu, H. H. S., Chung, W. Y., Yang, J. and Li, W. H. (2005), 'Patterns of segmental duplication in the human genome' *Mol Biol Evol* **22**(1), 135-141.

Zheng, X., Torstensen, B., Tocher, D., Dick, J., Henderson, R. and Bell, J. (2005), 'Environmental and dietary influences on highly unsaturated fatty acid biosynthesis and expression of fatty acyl desaturase and elongase genes in liver of Atlantic salmon (*Salmo salar*)' *Biochim Biophys Acta, BBA* **1734**(1), 13 - 24.

Appendix A

Atlantic Salmon Contig Viewer (Figure 7) - Website Access Statistics

For a one year period (March 15, 2009 – March 7, 2010) the information on the website has been accessed 238,447 times. Of those visits, 7,222 were from a unique IP address. Therefore, each address represents one or more uninterrupted sessions.

Appendix B

Publication List

1. Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., Munro, S., Moore, R., Holt, R. A., Jones, S. J. M., Davidson, W. S. and Koop, B. F. (2010), '*Salmo salar* and *Esox lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome' *BMC Genomics* **11**: 17.

   For this publication I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the full-length gene identification algorithm, the full-length reference gene identification algorithm, the NCBI data formatting, the NCBI data submission, and the NCBI data updating.

2. Koop, B. F., von Schalburg, K. R., Leong, J., Walker, N., Lieph, R., Cooper, G. A., Robb, A., Beetz-Sargent, M., Holt, R. A., Moore, R., Brahmbhatt, S., Rosner, J., Rexroad, C. E., McGowan, C. R. and Davidson, W. S. (2008), 'A salmonid EST genomic study: genes, duplications, phylogeny and microarrays' *BMC Genomics* **9**: 16.

   For this publication I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the cDNA library statistical analysis, the contig assembly statistical analysis, the full-length gene analysis, the web-based contig viewer implementation, assistancing with 32K microarray probe selection, the NCBI data formatting, and the NCBI data submission.

3. von Schalburg, K. R., Cooper, G. A., Leong, J., Robb, A., Lieph, R., Rise, M. L., Davidson, W. S. and Koop, B. F. (2008), 'Expansion of the genomics research on Atlantic salmon *Salmo salar* L. project (GRASP) microarray tools' *J Fish Biology* **72**(9): 2051-2070.

   For this publication I contributed by performing transcript annotation, and the 5K oligo microarray probe selection for 3' sequence bias.

4. von Schalburg, K. R., Leong, J., Cooper, G. A., Robb, A., Beetz-Sargent, M. R., Lieph, R., Holt, R. A., Moore, R., Ewart, K. V., Driedzic, W. R., ten Hallers, B. F. H., Zhu, B., de Jong, P. J., Davidson, W. S. and Koop, B. F. (2008), 'Rainbow smelt (*Osmerus mordax*) genomic library and EST resources' *Mar Biotechnol* **10**(5): 487-491.

For this publication I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the EST and contig statistical analysis, the NCBI data formatting, and the NCBI data submission.

5. Wright, J. J., Lubieniecki, K. P., Park, J. W., Ng, S. H. S., Devlin, R. H., Leong, J., Koop, B. F. and Davidson, W. S. (2008), 'Sixteen Type 1 polymorphic microsatellite markers from Chinook salmon (*Oncorhynchus tshawytscha*) expressed sequence tags' *Animal Genetics* **39**(1): 84-85.

   For this publication I contributed by performing the EST sequence handling, the EST assembly, that database design, the database population, the transcript annotation, the NCBI data formatting, and the NCBI data submission.

6. Yazawa, R., Yasuike, M., Leong, J., von Schalburg, K. R., Cooper, G. A., Beetz-Sargent, M., Robb, A., Davidson, W. S., Jones, S. R. M. and Koop, B. F. (2008), 'EST and Mitochondrial DNA Sequences Support a Distinct Pacific Form of Salmon Louse, *Lepeophtheirus salmonis*' *Mar Biotechnology* **10**(6): 741-749.

   For this publication I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the transcript selection, the NCBI data formatting, and the NCBI data submission.

Appendix C

Presentation List

1. Leong, J., Yasuike, M., Jantzen, S., Marass, F., von Schalburg, K.R., Davidson, W.S., Kay, W., Nilsen, F., Jones, S.R.M., and Koop, B.F. (2010), 'Genomic Resources For Sea Lice: Analysis of Genome Sequence, ESTs and Mitochondrial Genomes', *Sea Lice 2010 – The 8th International Sea Lice Conference*

   For this poster presentation I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the transcript statistical analysis, the whole genome sequencing, the whole genome assembly, the genome-transcript statistical analysis, and the web-interface implementation.

2. Leong, J., von Schalburg, K., Cooper, G., Moore, R., Holt, R., Davidson, W.S., and Koop, B. (2009), 'Identification And Annotation Of Full-Length Atlantic Salmon cDNAs', *Plant & Animal Genome XVII*

   For this poster presentation I contributed by performing the EST sequence handling for full-length libraries, the EST assembly, the database design, the database population, the transcript annotation, the transcript statistical analysis, the detailed full-length reference gene annotation, and the web-interface implementation.

3. Yasuike, M., Yazawa, R., Leong, J., Cooper, G.A., Beetz-Sargent, M., Robb, A., Davidson, W.S., Jones, S.R.M., Koop, B.F. (2009), 'Analysis of EST and mitochondrial DNA from the Pacific Salmon Louse, *Lepeophtheirus salmonis*', *Plant & Animal Genome XVII*

   For this poster presentation I contributed by performing the EST sequence handling, the EST assembly, and the transcript annotation.

4. Leong, J., Cooper, G., von Schalburg, K., Beetz-Sargent, M., Yazawa, R., Brown, G.D., Holt, R., Davidson, W.S., and Koop, B.F. (2007), 'Identification and annotation of Salmonidae genes and the fabrication of a 16,000 cDNA salmonid microarray', *Plant & Animal Genome XV*

   For this poster presentation I contributed by performing the EST sequence handling, the EST assembly, the database design, the database population, the transcript annotation, the transcript statistical analysis, the full-length cDNA prediction, and the web-interface implementation.

5. Yazawa, R., Cooper, G.A., Beetz-Sargent, M., Robb, A., Leong, J., Davidson, W.S., Koop, B.F. (2007), 'High Diversity for Antigen Recognition in the Atlantic Salmon T-cell Receptor Alpha / Delta Locus', *Plant & Animal Genome XV*

   For this poster presentation I contributed by performing the EST sequence handling, and the EST assembly.

Appendix D

Full-Length Reference Gene Data

To retrieve accession codes for all full-length reference genes submitted to GenBank, and discussed in Chapter 3 of my thesis, specific search terms can be used. The complete list consists of 9,057 records for Atlantic salmon (*Salmo salar*), and 1,365 records for northern pike (*Esox lucius*). They can be found in the core nucleotide database under search terms 'salmo salar[orgn] AND leong AND complete cds' and 'esox lucius[orgn] AND leong AND complete cds' for *S. salar* and *E. lucius*, respectively.