THE FATE OF DUPLICATED REGIONS OF THE ATLANTIC SALMON (*Salmo salar*) GENOME

by

Leslie Mitchell
B.Sc. (Hon.), University of Waterloo, 2002

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

MASTER OF SCIENCE

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

December 2004

APPROVAL

Name: **Leslie Mitchell**

Degree: **Master of Science**

Title of Thesis: **The Fate of Duplicated Regions of the Atlantic salmon (*Salmo salar*) Genome**

Examining Committee:

Chair: **Dr. Norbert H. Haunerland**
Professor, Department of Biological Sciences

Dr. William S. Davidson

Senior Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Fiona S.L. Brinkman

Supervisor
Assistant Professor, Department of Molecular Biology and Biochemistry

Dr. David L. Baillie

Supervisor
Professor, Department of Molecular Biology and Biochemistry

Francis Ouellette
Internal Examiner
Adjunct Professor, Department of Molecular Biology and Biochemistry
Associate Professor, Department of Medical Genetics, UBC

Date Defended/Approved: December 13, 2004

ABSTRACT

Gene and genome duplications have played a major role in vertebrate evolution. Salmonids provide a useful resource for studying the consequences of these events as their common ancestor underwent a genome duplication between 25 and 120 million years ago. To understand how a genome reorganizes itself to cope with duplicated chromosomes and the importance of gene duplications for evolution and adaptation, homeologous regions of the Atlantic salmon genome were identified and studied within a large insert, genomic BAC library; these BACs contain the metallothionein loci, a gene known to have remained in duplicate since the tetraploidization event. A BAC from each region was subsequently shotgun subcloned and sequenced. Sequence analysis revealed the presence of 10 genes, retaining their collinearity between the BACs, although pseudogenization events have occurred in one of the duplicate loci in two instances. Comparative genomic analysis revealed the existence of extraordinary conservation of synteny over time.

DEDICATION

For my family; John, Joanne, and Jeff.

ACKNOWLEDGEMENTS

I would sincerely like to thank my supervisor, Dr. Willie Davidson, for giving me the opportunity to work on this project and being so supportive and enthusiastic. I would also like to thank Team Davidson for being such a fantastic group of people to work with. Finally, I would like to thank my family – Mom, Dad, and Jeff – for their support.

TABLE OF CONTENTS

Approval ii

Abstract iii

Dedication iv

Acknowledgements v

Table of Contents vi

List of Figures vii

List of Tables viii

Chapter 1 Introduction 1

1.1 Preamble 1
1.2 Gene Duplication 2
1.3 Evolution and Fish 8
1.4 Salmonid Fishes 11
1.5 Salmonids and Metallothionein 19
1.6 Aim of Thesis 20

Chapter 2 Materials and Methods 21

2.1 Characterization of Salmonid Metallothionein Genes 21
2.2 Identification and Shotgun Sequencing of Metallothionein BACs 23
2.3 Sequence Annotation 30

Chapter 3 Results 33

3.1 Salmonid Metallothionein Genes 33
3.2 Identification of Metallothionein BACs 38
3.3 Confirmation of Duplication 48
3.4 Shotgun Libraries of Metallothionein BAC Clones 51
3.5 BAC Contig Construction 51
3.6 BAC Sequence Annotation 56
3.7 Gene Annotation 78
3.8 Phylogenetic analysis 106

Chapter 4 Discussion 109

4.1 Metallothionein and the Salmonid Tetraploidization Event 109
4.2 Annotation of Homeologous Regions of the Atlantic salmon Genome 114
4.3 Conservation of Synteny 123
4.4 Dating the Tetraploidization Event 123
4.5 Future Directions 125

Reference List 127

LIST OF FIGURES

Figure 1 Salmonid phylogeny. 13
Figure 2 Multiple sequence alignment of partial metallothionein sequences 34
Figure 3 Neighbour-joining tree based on partial metallothionein sequences 36
Figure 4 BAC filter hybridized with a metallothionein probe and the control *C. briggsea* DNA 40
Figure 5 Contig 341. 42
Figure 6 Contig 2172. 43
Figure 7 PCR confirmation and identification of MetA and MetB positive contigs. 45
Figure 8 *Hind*III fingerprints for MetA positive clones. 46
Figure 9 *Hind*III fingerprints for MetB positive clones 47
Figure 10 MetA FISH 49
Figure 11 MetB FISH 50
Figure 12 S0085O16 initial build 52
Figure 13 S0188I22 initial build 54
Figure 14 Dot plot of S0085O16 MetA versus S0188I22 Met B 57
Figure 15 S0085O16 versus S0085O16 Pipmaker Plot 64
Figure 16 S0188I22 versus S0188I22 PipMaker Plot 66
Figure 17 Multiple Sequence Alignment and Consensus Sequence of the Blue Minisatellite 69
Figure 18 Atlantic salmon gene order 76
Figure 19 Conservation of synteny 77
Figure 20 Structure of the Atlantic salmon metallothionein A and B genes 82
Figure 21 Alignment of MetA and MetB regulatory regions 83
Figure 22 Lin10 amino acid sequence alignment. 85
Figure 23 Dead eye amino acid multiple sequence alignment 87
Figure 24 Beta-1,3-Galactosyltransferase multiple sequence alignment 90
Figure 25 CBFB amino acid multiple sequence alignment 92
Figure 26 GNAO splice variant amino acid multiple sequence alignments 94
Figure 27 Herp amino acid multiple sequence alignment 96
Figure 28 Cetp amino acid multiple sequence alignment 98
Figure 29 BBS2 amino acid multiple sequence alignment 100
Figure 30 SLC12A3 amino acid multiple sequence alignment 103
Figure 31 Number of nucleotide changes versus divergence time 107
Figure 32 Number of amino acid changes versus divergence time 108

LIST OF TABLES

Table 1 Metallothionein Positive BAC Clones 39
Table 2 Genes Identified in the BAC Sequences 59
Table 3 Percentage of BAC Sequences that Contain Repetitive DNA 62
Table 4 Minisatellite Repetitive Elements Unique Between BACs 71
Table 5 Percent G+C Content of Exons and Introns 74

CHAPTER 1 INTRODUCTION

1.1 Preamble

"Natural selection merely modified, while redundancy created". This statement was made by Susumu Ohno in his forward-thinking book "Evolution by Gene Duplication" (1970a), where he put forth the theory that most novelty in evolution occurs as a result of genomic redundancy. He proposed that the presence of two copies of a gene allows for one copy to accumulate "forbidden mutations", while the other maintains the original function. In this way, divergent selection may lead to genes with new functions. Furthermore, he speculated on the evolutionary role played by both tandem and whole genome duplications, suggesting that at least two genome wide duplications occurred early in vertebrate evolution, and were overall much more important to evolution than tandem duplications. Since his book was published, its relevance has remained, and evidence for his theories has accumulated based on sequence data from complete genome sequencing projects.

Ohno's theory is most impressive because he presented it at a time when the hypotheses could not be easily tested, when the documentation and quantification of genetic variation within populations and between species was largely restricted to isozyme studies and kayrotype inspections. His ideas were, therefore, ahead of his time, similar to Charles Darwin, who proposed the theory of natural selection before the field of genetics was even created.

Ohno's classical model postulates two major fates for duplicated genes: (1) non-functionalization, where the accumulation of deleterious mutations leads to the formation of pseudogenes and eventual loss of the duplicate copy, and (2) neo-functionalization, where beneficial mutations and positive selection leads to genes with a novel product and function. Additionally, Ohno suggests the possibility of mutations in regulatory regions affecting the roles played by the duplicate genes. This third possibility has been expanded upon by Force et al. (1999) in the Duplication-Degeneration-Complementation (DDC) model, which hypothesizes that degenerative mutations in regulatory regions can lead to subfunctionalization, where the original function is partitioned between the duplicate genes. This would, in turn, result in the preservation of both copies of the duplicate gene. The unique aspect of this model is that deleterious mutations actually lead to duplicate preservation, in contrast to the classical model, where deleterious mutations only result in non-functionalization.

1.2 Gene Duplication

1.2.1 Types of Gene Duplication

Duplication of genetic information can occur in several different ways and can be classified according to the extent of the genomic region that is involved. There are five recognized mechanisms that lead to an increased copy number of DNA segments: (1) partial or internal gene duplication, (2) complete gene duplication, (3) partial chromosome duplication, (4) complete chromosome duplication, and (5) complete genome duplication (Graur and Li, 2000). In principle, all of these mechanisms have the potential to contribute to the evolutionary process by creating a redundant segment of DNA on which natural selection can be relaxed. In reality, however, not all five kinds of duplications have played equal roles in evolution. Partial or internal gene duplications, where the duplicate copy may be missing the regulatory region associated with the gene are unlikely to be functional. Duplications that result in partial or complete polysomy can lead to extensive imbalances in gene products, and therefore are less evolutionarily important. Evidence of this is shown in man, where trisomies of larger chromosomes are lethal, and those of smaller chromosomes (e.g. trisomy 21) cause reduced fitness (Li, 1997). On the other hand, regional duplications, such as those observed when genes are tandemly duplicated, and complete polyploidy play a vastly more important role in evolution, and examples of both are found extensively in nature.

1.2.1.1 Tandem Duplications

Tandem duplications result in adjacent, identical chromosomal segments. One of the best characterized examples of tandem duplication is that of the genes for rRNA. While organisms require four kinds of rRNA (5S, 5.8S, 18S, 28S), very large quantities of each are necessary to carry out translation in a cell. To compensate, organisms contain multiple copies of the rRNA genes. The genes are arranged such that one cistron encodes the tandem repeats of the 18S, 5.8S, and 28S rRNA, and a separate locus codes for multiple copies of the 5S rRNA. For instance, the African clawed frog (*Xenopus laevis*) contains 500-760 tandemly duplicated complete 18S, 5.8S, 28S rRNA gene sets, yeast (*Saccharomyces cerevisiae*) contains about 140, *C. elegans* contains about 55, and humans contain about 300 (Graur and Li, 2000). It has been hypothesized that these tandem repeats have been selected for in order to fulfil the enormous metabolic requirement of the gene product. Other examples of this phenomenon include the histone genes (Rooney et al., 2002) and tRNA genes (Perez et al., 2000). The sequences of these tandemly duplicated genes are believed to be kept homogenous by the process of concerted evolution (Graur and Li, 2000)

Other instances of tandem duplication events result in divergence of the duplicate loci, allowing for gene product diversity, as opposed to large quantities of the same product. The enzymes trypsin and chymostrypsin, which are involved in protein digestion, provide an excellent example of gene product diversity. The main difference between the two proteins is their substrate specificity; chymotrypsin cleaves at the C-terminal side of bulky hydrophobic residues like tyrosine, phenylalanine, and tryptophan, while trypsin cleaves at the C-terminal side of basic residues like lysine and arginine. Maintaining active copies of both genes endows an organism with an increased capability for food digestion. By comparing the amino acid sequences of the proteins, it is quite obvious that the locus for one enzyme evolved from a duplicate copy of the original locus by mutations that affected the active site (Ohno, 1970a). This example of divergent evolution highlights the process of neofunctionalization by tandem gene duplication.

1.2.1.1.1 Mechanisms of Tandem Duplications

Two mechanisms that generate tandemly duplicated stretches of DNA are slipped-strand mispairing and unequal recombination. The latter (also known as unequal crossover) is viewed as the predominant biological mechanism responsible for the production of large, tandemly repeated sequences (Elemento et al., 2002). When chromosomes do not line up properly during meiosis due to the presence of repeated sequences, unequal pairing of the chromatids occurs and the resulting shift leads to one chromatid ending up with a duplicated region, and the other with a deleted region. In turn, tandemly repeated regions increase the likelihood of further duplications because the possibility of mispairing is higher, which may ultimately lead to block duplications (Elemento et al., 2002).

1.2.1.1.2 Disadvantages of Tandem Duplications

Although tandem duplication is a mechanism that has been used extensively during evolution, there are some drawbacks. As mentioned above, the presence of a tandem duplication permits further unequal exchange and unequal crossing over. A second limitation of tandem duplication is that the presence of two (or more) copies of a particular gene that arise via tandem duplication may can disrupt the natural gene dosage ratio. Finally, and probably most importantly, if the tandem duplication of a gene does not include the associated regulatory region that governs the gene, there is little chance of the duplicate gene maintaining its function (Ohno, 1970a).

1.2.1.2 Complete Genome Duplications

Complete genome duplications are much rarer than tandem duplications but are arguably more important in driving major evolutionary change (Ohno, 1970a). As the DNA is duplicated in its entirety and not just specific genes or regions, potential problems with gene dosage ratio and missing regulatory regions are avoided. Furthermore, because the duplicates are carried on separate chromosomes, no instability due to unequal crossing over ensues, as seen with tandem duplicates.

Evidence of polyploidy is particularly abundant in the plant kingdom. Estimates suggest that up to 70% of angiosperms (flowering plants) have experienced one or more episodes of polyploidization (Masterson, 1994). Polyploidy is prolific in plants because so many of these species are hermaphroditic, that is both male and female sex organs (stamens and carpels or pistils, respectively) are present in the same flower. This also helps to explain the scarcity of polyploids in invertebrates and vertebrates. The bisexual nature that categorizes the majority of these species is incompatible with polyploidy. To explain this, consider that a teptraploid male would have to maintain an XXYY constitution, and a tetraploid female would need an XXXX constitution. During meiosis, a mechanism that would ensure exclusive production of two classes of gametes by the male, XX and YY, would have to exist to guarantee the production of viable offspring. Otherwise, XY gametes would be produced by the male, resulting in XXXY offspring, which would generate two possibilities: all male offspring or infertile offspring.

Unfortunately, since no such mechanism exists, those organisms with well established chromosomal sex-determining mechanisms are excluded from polyploidizations. While modern mammals, birds, and reptiles are denied the possibility of evolution by polyploidization, however, two ancient episodes of tetraploidization are hypothesized for the vertebrates (Spring, 1997). Organisms such as fish and amphibians, whose chromosomal sex-determining mechanisms are only in an initial state of differentiation, are some of the few vertebrates still capable of polyploidy. The consequences of polyploidization in the evolutionary trajectory of fish will be discussed in detail below.

Polyploidy is an effective mechanism of speciation because sexually reproducing autoploids are automatically isolated from their diploid progenitors due to the nature of their gametes. Combining a diploid gamete from a tetraploid organism with the haploid gamete from a diploid organism results in a triploid offspring, which, if not lethal, would result in infertility. The effect of polyploidy on speciation therefore is one of reproductive isolation.

1.2.1.2.1 Mechanisms for Complete Genome Duplications

Two well-established mechanisms exist for complete genome duplication. Allopolyploidy is the condition that arises from the combination of genetically distinct chromosome sets, and usually occurs through hybridization events. Autopolyploidy is simply the multiplication of one basic set of chromosomes. This can occur when there is a lack of disjunction between all the daughter chromosomes following chromosome replication, or when two cells fuse together to form a tetraploid cell (Li, 1997).

1.2.1.2.2 Disadvantages of Genome Duplications

Although genome duplication appears to provide an evolutionary benefit, several deleterious effects could accompany a complete doubling of the DNA content in a cell:

(1) cell division time is prolonged, (2) the volume of the nucleus is increased, (3) there is an increase in the number of chromosome disjunctions during meiosis, (4) there is an increased likelihood of genetic imbalances, and (5) possible interferences with sexual differentiation when the sex of the organism is determined by either the ratio between the number of sex chromosomes and the number of autosomes, or by the degree of ploidy (Graur and Li, 2000).

1.2.1.2.3 Detection of Complete Genome Duplications

The availability of complete genome sequence data for several eukaryotes has led to the identification of whole genome duplication events in various species and has refocused the attention of the scientific community on this topic. For instance, a map-based approach has been used to provide strong evidence in favour of genome duplication events in both *Arabidopsis thaliana* and *Saccharomyces cerevisiae* (reviewed by Wolfe, 2001). This approach seeks to identify the chromosomal locations of duplicate genes, looking for chromosomes (or sections of chromosomes) that can be paired up because they contain similar sets of genes (ideally in the same order on the chromosomes). In both yeast and *Arabidopsis*, paired chromosomal regions can be identified that cover more than half the genome. In fact, further work by Kellis et al. (2004) has provided conclusive proof of an ancient genome duplication event in *S. cerevisiae* based on comparison to the complete genome sequence of *Kluyveromoyces waltii*, a closely related, non-duplicated species of yeast. This group demonstrated that a 1:2 relationship exists between non-duplicated and duplicated chromosomes in these two species, such that sister regions in *S. cerevisiae* contain an ordered subsequence of the genes in the corresponding region of *K. waltii* (Kellis et al., 2004).

Phylogenetic trees can also be used to test for evidence of genome duplication events because genes that were duplicated simultaneously should share a common history. Gene pairs that make up a duplicated chromosomal segment should be the same age, and furthermore, following a genome wide duplication event, all of the duplicated segments in the genome should be the same age. Therefore, age estimates made from phylogenetic trees should be consistent among trees drawn from different genes. Taken together with evidence for large scale block duplications, Vandepoele et al (2004) demonstrated that a major portion of the duplicate genes in *Takifugu rubripes* are about the same age (225 to 425 mya divergence times) and therefore suggested the most parsimonious explanation would be an entire genome duplication.

In general, detection of whole genome duplication is confounded by both the loss of duplicate genes as well as chromosomal rearrangements over time such that the more ancient the polyploidization episode, the more difficult it is to delineate. This has resulted in conflicting opinions towards Ohno's suggestion of multiple rounds of complete genome duplication early in the vertebrate lineage. The 2R hypothesis, as it has come to be known, states that two rounds of genome duplication occurred in the vertebrate lineage, one immediately before, and one immediately after the divergence of the lamprey lineage. To date, no genome duplication as ancient as the ones suggested under the 2R hypothesis (430-750 million years ago (Gu et al., 2002)) has been proven.

1.3 Evolution and Fish

1.3.1 Timetable of Evolution

Life on earth is believed to have arisen approximately 4 billion years ago (bya) in the Archean era, which extended until 2.5bya (Graur and Li, 2000). Between 2.5bya and 590 million years ago (mya), during the Proterozoic era, eukaryotes, photosynthetic, and multi-cellular organisms emerged (Graur and Li, 2000). The superphylum Chordata, the chordates, appeared soon after this near the end of the Proterozoic era (Kumar and

Hedges, 1998). The first of these vertebrates were the ostracoderms, or jawless fishes, of the class Agnatha. Subsequently, fish went on to develop jaws, a major step in vertebrate evolution, which preceded the split between ray-finned fishes (Actinopterygii) and the lobe-finned fishes (Sarcoptergyii) about 450mya (Venakatesh, 2003).

The ray-finned fish are the most diverse group of vertebrates and are comprised of ~23,700 extant species. Approximately 99% of ray-finned fish are classified as teleosts (bony fishes), and the remaining 1% belong to the Chondrostei (sturgeons, bichirs), and Holostei (bowfins, gars) (Nelson, 1994). Based on fossil evidence, the "non-teleost" fish diverged from the teleosts at least 235mya (Venkatesh, 2003)

The lobe-finned fishes, on the other hand, include the rest of the bony vertebrates, such as the coelacanths, lungfishes and tetrapods (mammals, birds, reptiles, and amphibians), of which there are ~23,600 living species (Venkatesh, 2003).

1.3.2 Teleost Genome Duplication Events

Teleost fish comprise an extremely diverse and successful group of vertebrates, in that they comprise about fifty percent of all extant vertebrates and they show vast differences in morphology and adaptations. The successful radiation of the teleosts has been attributed to an ancient genome duplication event that occurred in an ancestor of teleost fish after the split from the "non-teleosts" (Postlethwait et al., 2004).

Several lines of evidence have been put forward in support of a teleost-specific genome duplication event, including the observation that teleost fish have more copies of many genes than their tetrapod relatives (van de Peer et al., 2003; Vandepoele et al., 2004); specific attention has centred on the fact that teleosts have up to eight copies of the *Hox* gene clusters, a group of genes that are critical for development in bilaterian vertebrates, whereas all tetrapods have only four copies (Amores et al., 2002).

Additionally, synteny studies have described multi-gene blocks of paralogy that exist between various teleosts including zebrafish, medaka, and *Takifugu* (Taylor et al., 2003).

The recently completed genome sequence of the freshwater pufferfish, *Tetraodon nigroviridis*, provides proof of a whole genome duplication in this species (Jaillon et al., 2004). Using the same technique employed by Kellis et al. (2004) in *S. cerevisiae*, Jaillon and colleagues (2004) demonstrated that (1) nearly every region in the human genome corresponds to two sister regions in the *Tetraodon* genome; (2) the two sisters regions in *Tetraodon* each contain a subset of the genes in the corresponding region of human, with each containing roughly half of the genes and the two subsequences interleaving to account for nearly all the genes; and (3) nearly every region of *Tetraodon* corresponds to one region of the human genome, and is therefore paired to a sister region in *Tetraodon.* The pattern of distribution of duplicated genes in the *Tetraodon* genome shows precisely the signatures expected from a whole genome duplication followed by massive gene loss. In combination with the evidence mentioned above, specifically the presence of supernumerary *Hox* clusters, Jaillon and colleagues (2004) suggest that the duplication event they detected in the *Tetraodon* genome probably affected all teleosts.

1.3.3 Duplicate Genes and the Teleost Radiation

Amores et al. (1998) have proposed that the highly successful teleost radiation was spurred on by the teleost-specific genome duplication event. Myer and Schartl (1999) take this one step further and suggest that there may be a cause-effect relationship between gene copy number and species diversity, which, in this case, is consistent considering that the duplication event may have taken place at about the time of the teleost radiation.

Divergent evolution has been proposed as a mechanism by which gene duplication and genome duplication can contribute to speciation. Divergent evolution occurs when different copies of a duplicated gene are lost or changed in geographically separated populations and could genetically isolate these populations should they become re-united (Lynch and Force, 2000; Lynch and Conery, 2000).

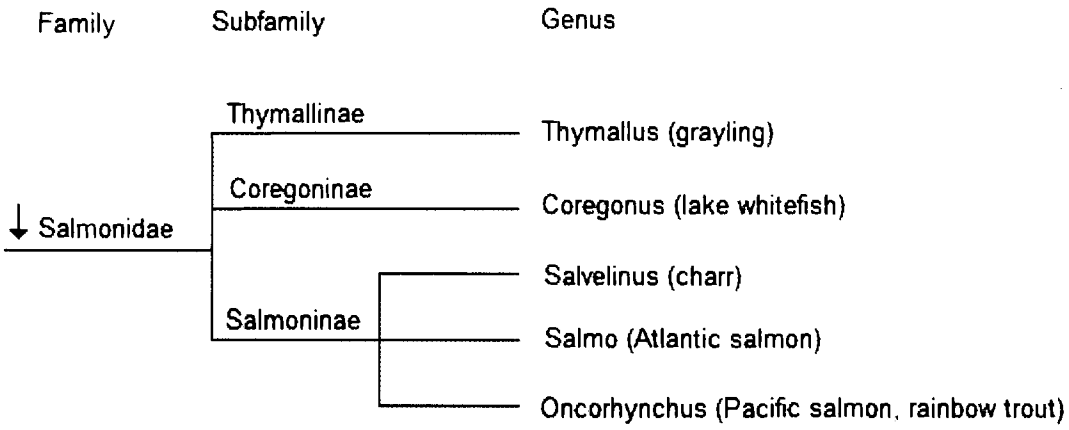
1.4 Salmonid Fishes

1.4.1 Introduction to Salmonids

The Salmonidae family includes the subfamilies Coregoninae (whitefishes, ciscos), Thymallinae (graylings), and Salmoninae (trouts, salmons, charrs) (Figure 1). These fish have been further classified into nine genera and roughly sixty eight species (Nelson, 1994). This family has been studied extensively due to the commercial importance of many of its members. The Salmonidae are native to the Northern hemisphere, but have been introduced to many areas of the world. While members of this family may be anadramous or freshwater, spawning events always take place in fresh water. Some species (sockeye and chinook salmon) die after spawning, whereas others (Atlantic salmon, trout, and charr) can spawn several times.

Notably, members of the Salmonidae family diverged from a common ancestor that is believed to have undergone a tetraploidization event between 20 and 120 million years ago, after the teleost radiation (Allendorf and Thorgaard, 1984). It has been suggested that the diploid ancestor of salmonids possessed 48 acrocentric chromosomes, based on the fact that this karyotype is widespread among other fish groups (reviewed in Phillips and Rab, 2001). Thus, the ancestral salmonid karyotype probably had 96 acrocentric chromosomes. Extant salmonids have 52 to 102 chromosomes per 2N cell (of which more than half are metacentric or submetacentric) and genome sizes about twice that of most other fish (Ohno et al., 1968; Phillips and Rab, 2001).

Figure 1 **Salmonid phylogeny.** The relationship between the three subfamilies within the Salmonidae family. The genome wide duplication event is denoted by an arrow.



1.4.2 Evidence for an ancestral genome duplication event in salmonids

Four major lines of evidence substantiate the hypothesis of an ancestral tetraploidization event in salmonid evolution (Ohno, 1970b; reviewed in Allendorf and Thorgaard, 1984). First, salmonid fish, with about 80% as much DNA per cell as mammals, have roughly twice the amount of DNA per cell as closely related fish such as smelt, herring, and anchovy. Second, salmonids have approximately twice as many chromosome arms (100) as closely related fish. Third, meiotic preparations from several salmonid species have revealed multivalent chromosomes lining up. Finally, duplicated enzyme loci are prevalent in salmonids, and moreover, many of the duplicated genes that have been sequenced fall into two clades.

An example of a duplicated locus is that of the growth hormone genes, GH1 and GH2. The GH1 genes in the genera *Brachymystax*, *Hucho*, *Salvelinus*, *Salmo*, and *Oncorhynchus* form a clade, and the GH2 genes for these same taxa form another clade, supporting the hypothesis of an ancestral tetraploidization followed by divergence (Oakley and Phillips, 1999).

The salmonid tetraploidization event was probably autotetraploid in nature. This is supported by the presence in current salmonids of multivalents at meiosis, by the existence of some duplicate loci pairs with no evidence of divergence, and by apparent examples of tetrasomic inheritance of some of these loci (reviewed in Allendorf and Thorgaard, 1984). Taken together, this makes a segmental allopolyploidy event unlikely for the salmonids.

1.4.3 Time of occurrence

Determining the date of the tetraploidization event with a reasonable degree of confidence has proven to be difficult. Estimates of 25-120mya have been made based on the principle of the molecular clock, i.e., homologous proteins tend to evolve at similar rates in different lineages (reviewed in Allendorf and Thorgaard, 1984). This type of estimation has been called into question because it is based on the assumption that orthologous genes in separate species will evolve at the same rate as duplicate genes in the same species. Duplicate genes, however, would be expected to evolve more quickly because they would not face the same pressures of natural selection, provided that one locus kept its' original function while the other locus was allowed to accumulate "forbidden mutations". Another factor that makes the tetraploidization event hard to date is that it is difficult to associate the time of tetraploidization with the divergence of duplicate loci, in that divergence cannot begin until disomic inheritance is re-established. The length of time of divergence is therefore a minimum estimate of the time since the tetraploidization event. The fossil records for salmonids, which would help to unravel the mystery, are extremely sparse for the time period in question.

1.4.4 Current salmonid genetic system

1.4.4.1 Sex chromosomes and sex determination

In all salmonids studied to date, the male is the heterogametic sex. Some salmonids, such as the rainbow trout (*Oncorhynchus mykiss*) and sockeye salmon (*Oncorhynchus nerka*) have heteromorphic sex chromosomes, while in others such as the Atlantic salmon (*Salmo salar*), the sex chromosomes are indistinguishable from one another (reviewed in Phillips and Rab, 2001). Woram et. al (2003) examined the linkage groups containing the SEX locus of Artic charr (*Salvelinus alpinus*), brown trout (*Salmo trutta*), Atlantic salmon, and rainbow trout, and found that SEX is associated with different linkage groups in each of the representative species of the three major subfamilies of salmonids. Furthermore, the arrangement of the markers close to the sex-determining locus were found to be preserved on homologous but different linkage groups across the four species, indicating that a small region of DNA has been involved in the rearrangement of the sex-determining region (Woram et al., 2003). The discovery of the sex-determining factor, however, still remains a highly coveted topic in the fields of salmonid research and sex-determination.

1.4.4.2 State of diploidization

Although disomic inheritance is prevalent, currently the salmonids are still in the process of returning to a diploid state, at least for a fraction of their genome, evidenced by the fact that extant males exhibit quadrivalents in meiosis (Allendorf and Thorgaard, 1984).

Following a tetraploidization event, a newly arisen tetraploid organism is expected to display multivalent formation and tetrasomic segregation. In order for the tetraploid state to contribute to the evolution of new gene loci, a state of disomy needs to be re-established. This can occur by functional diversification of the four original homologues, so that one of the original linkage groups is transformed into two separate linkage groups (Ohno, 1970b). The term diploidization, then, is used to describe the return to a disomic state, whereby the preferential formation of two separate bivalents is the pre-requisite. Without disomic re-establishment, the four homologues would be expected to form a quadrivalent during meiosis from which the four would segregate randomly at the end of the first meiosis, allowing no possibility of functional diversification.

The diploidization process is driven by the fact that non-disjunction of chromosomes during meiosis leads to decreased fertility. Since the likelihood of a non-disjunction event is increased for multivalents, selection should favour a reduction of multivalent pairing and a restoration of disomy (Ohno et al., 1969). In addition to this, disomy might also be favoured because it allows the structural and regulatory divergence of newly duplicated gene loci (Allendorf and Thorgaard, 1984).

Robertsonian fusions, which involve the fusion of two acrocentric chromosomes at their centromeres, are likely to be the major mode of karyotype evolution in the Salmonidae (Phillips and Rab, 2001). As it has been suggested that the diploid ancestor of salmonids possessed a karyotype with 48 acrocentric chromosomes, the resulting 96 acrocentrics that were present in the original tetraploid karyotype could quite possibly have undergone extensive fusions (Phillips and Rab, 2001). Selection for Robertsonian fusions and the resulting metacentric chromosomes probably aided in the diploidization process by effectively eliminating the presence of multivalents at meiosis (Roberts, 1970; Ohno et al., 1969).

The evolutionary changes that the karyotypes of the three subfamilies of the salmonid fishes have undergone were extensively reviewed by Phillips and Rab (2001). The *Thymallinae* have evolved by inversions, retaining chromosome numbers close to the karyotype of the hypothetical tetraploid ancestor but substantially increasing the chromosome arm numbers. The chromosomes of the *Corregoninae* and *Salmoninae*, on the other hand, have evolved by centric fusions, leading to a decrease in chromosome numbers while retaining chromosome arm numbers close to the hypothetical ancestor. The exception to this is the Atlantic salmon, where the chromosome arm number changed by subsequent tandem fusions.

1.4.5 Fate of Duplicate Loci

The evolution of duplicate loci following a tetraploidization event can be divided into three time periods. First of all, the re-establishment of disomy must occur. Second, a period of divergence occurs, when the original locus is functionally duplicated and the two copies can diverge from one another. Finally, the third period begins when a substantial degree of structural (fixed inheritance of alleles) or regulatory divergence (tissue-specific patterns of expression) of the duplicate loci has taken place. These three time periods are non-discrete, but nevertheless, it is important to distinguish among them.

One of the best studied duplicate loci in the salmonids, and the first to be described, is lactate dehydrogenase (LDH). The diploid ancestor of the salmonids apparently possessed three LDH loci: a muscle-specific locus (LDH-A), an eye-specific locus (LDH-C), and a third locus found in all tissues (LDH-B). Wright et al. (1975) undertook a study examining the genetic control of LDH in *Salmo* and *Salvelinus*, where they found five loci coding for LDH activity in species of these genera. The two loci (*Ldh-1* and *-2*) that descend from the ancestral LDH-A have different common alleles, but exhibit the same expression patterns. The two loci (*Ldh-3* and *-4*) that descend from LDH-B also have different isozymes, but additionally exhibit different tissue-specific expression patterns. *Ldh-3* is expressed predominantly in the heart, while *Ldh-4* is expressed mainly in the liver. Finally, only one locus coding for the eye-specific form of LDH was found. A similar situation is seen in the whitefish, in which LDH is also encoded by five loci, however, the two descendants of LDH-B do not show tissue specific expression (Massaro, 1972). It may be concluded that the similar expression pattern of two LDH-B duplicates in salmonids is the primitive condition, whereas tissue specific regulation is a derived characteristic that is likely to have evolved only once. Taken together, this indicates that the *Salmo* and *Salvelinus* lineages shared a common ancestor longer after the tetraploidization event than they did with whitefish.

1.5 Salmonids and Metallothionein

Metallothionein is another well-characterized salmonid protein. Metallothioneins are non-enzymatic proteins involved in binding primarily divalent cations and play an essential role in homeostasis and detoxification of heavy metals in a large variety of organisms (Cousins, 1985). Sequence conservation, structure, and function among metallothoinein proteins across a wide variety of taxa are remarkable. Some properties of metallothionein proteins include low molecular weight (~6kda) and an unusual amino acid composition of about 30% cysteine resides organized into characteristic motifs (Cys-X-Cys and Cys-X-X-Cys). The cysteine resides engage in the formation of metal thiolate clusters. Induction of metallothionein by metals is mediated by multiple copies of metal-responsive elements (MREs) in the MET gene 5' regulatory region.

To date, several salmonid species, including rainbow trout, Atlantic salmon, and Artic charr, have been found to contain duplicate metallothionein genes. The genes, which are known as MetA and MetB, display a tripartite exon-intron structure and the differences between the two copies include an expanded second intron in MetB (roughly 450 base pairs longer), and the presence of an indel resulting in the addition of one amino acid at position 31 in MetA. The duplicate gene pairs are hypothesized to be the result of the salmonid-specific tetraploidization event, seeing as many other tèleosts including the closest relative of the salmonids, the Northern pike (*Esox lucius*), possess only a single copy of the metallothionein gene. The metallothionein genes, therefore, were used as a starting point for identifying homeologous regions of the Atlantic salmon genome that arose through tetraploidization. First, however, further study of the salmonid MetA and MetB genes was carried out to better characterize the nature of the duplication event.

1.6 Aim of Thesis

The goal of this work is to compare large, homeologous regions of the Atlantic salmon genome to one another as well as to corresponding regions in other genomes to better understand the evolution of duplicated genomic regions, and the specific genes that lie therein.

CHAPTER 2 MATERIALS AND METHODS

2.1 Characterization of Salmonid Metallothionein Genes

2.1.1 Sequencing Salmonid Metallothionein Genes

Metallothionein specific primers (Forward 5' ATG GAT CCG TGT GAA TGC TC; Reverse 5' GAT ACC AGC TGT TGT CAG TGA) were used to amplify genomic DNA from several salmonid species including *Oncorhynchus gorbuscha* (pink salmon), *Oncorhynchus keta* (chum salmon), *Oncorhynchus nerka* (sockeye salmon), *Oncorhynchus tshawytscha* (chinook salmon), and *Thymallus arcticus* (Arctic grayling). In a volume of 25uL, the following procedure was performed: initial denaturation step, 95°C for 5 minutes, 35 cycles with a denaturation step at 95°C for 1 minute, annealing at the Tm of 50°C for 1 minute, extension at 72°C for 5 minute, followed by a final extension at 72°C for 5 minutes. For DNA amplification, 0.05U Taq, 12.5 pmoles of each specific primer, 2.5µL 10X buffer containing 1.5mM $MgCl_2$, and 12.5µmol of dNTPs and 50ng of genomic DNA was mixed together. Following amplification, the products were electrophoresed on a 1% agarose gel made with 1X TBE. Multiple bands in each lane, corresponding to MetA and MetB, were separately excised from the gel using the Qiagen Gel Extraction Kit and subcloned using the pGEM-T Easy Vector System (Promega, Madison, WI, USA) and XL1-Blue Competent Cells (Stratagene, La Jolla, CA, USA) following the protocols recommended in the kits.

A colony PCR method using the metallothionein primers was employed to check the insert size of several white clones from each subcloning reaction. In a volume of 25µL, the following procedure was performed: initial denaturation step, 95°C for 5 minutes, 35 cycles with a denaturation step at 95°C for 1 minute, annealing at the Tm of 50°C for 1 minute, extension at 72°C for 5 minute, followed by a final extension at 72°C for 5 minutes. For each DNA amplification, 0.05U Taq, 12.5 pmoles of each specific primer, 2.5µL 10X buffer containing 1.5mM $MgCl_2$, and 12.5µmol of dNTPs was mixed together. An isolated, white colony was picked with a yellow tip and dipped and swirled in the PCR mix to provide the template for amplification. 5µL of PCR product was electrophoresed on a 1% agarose gel made with 1X TBE. Clones representing MetA and MetB for each salmonid species were chosen for sequencing. The remaining 20µL of selected PCR product from above was purified using the Qiagen PCR Cleanup Kit and used for sequence analysis with the fluorescent dideoxy terminator method. Briefly, 3µL of cleaned-up PCR product was mixed with 5pmol of primer (M13F or M13R), and 4µL of sequencing premix in a final volume of 10µL. Sequencing reactions were carried out with the following conditions: 94°C 20sec, Tm (50°C) 20sec, 60°C 1min, for 30 cycles. Following amplification, unincorporated dideoxynucleotides were removed from each reaction mixture by ethanol precipitation, and the reactions were resuspended in 2uL of diformamide loading dye. Sequence analysis was carried out on an ABI 377 DNA Sequencer (Applied Biosystems).

Additional metallothionein sequence data was downloaded from GenBank. Accession numbers are as follows: *Oncorhynchus mykiss* (rainbow trout) M81800.1 and M22487.1; *Salvelinus alpinus* (Arctic charr) AY267819.1 and AY267818.2; Esox lucius (Northern pike) X70042.1.

2.1.2 Phylogenetic Analysis

A multiple sequence alignment of all MetA and MetB sequences was performed using Clustal W (Thompson et al., 1994). Sequence alignments were corrected by eye in Seaview (Galtier et al., 1996). Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 3.0 (Kumar et al., 2004).

2.2 Identification and Shotgun Sequencing of Metallothionein BACs

2.2.1 BAC Library

An Atlantic salmon BAC library was available for this work (Thorsten et al., submitted). Briefly, the library consists of approximately 300,000 clones with an average insert size of 190 kilobase pairs (kb), and provides an 18.8X coverage of the Atlantic salmon genome. The clones have been stamped onto 17 nylon membranes for hybridization purposes. Each filter contains 18,432 clones stamped in duplicate, as well as control anchor spots consisting *C. briggsae* DNA, to be utilized to orient the filters after hybridization. Additionally, a physical map of the BAC library has been created by *Hind*III fingerprinting the first two thirds of the library and placing the clones into contigs (Artieri et al., submitted). The physical map provides an 11.8X coverage of the Atlantic salmon genome and contains 4,353 contigs and 27,824 singletons.

2.2.2 Probe Design

A 40mer oligonucleotide probe was designed based on the coding sequence in the third exon of the Atlantic salmon metallothionein genes. The metallothionein probe sequence is 5' TCA CTG ACA ACA GCT GGT ATC ACA AGT CTT GCC CTT GCA A. A second probe was designed to hybridize to the *C. briggsea* control anchor spots. The sequence of this 50mer oligonucleotide is 5' GTT GCC AAA TTC CGA GAT CTT GGC GAC GAA GCC ACA TGA T.

2.2.3 BAC Library Screening

Both the metallothionein probe and the *C. briggsea* anchor probe were simultaneously hybridized onto the first six filters in the BAC library, providing approximately 6X coverage of the genome.

2.2.3.1 Probe labelling

Oligonucleotide probes were 5'-end-labeled with ATP γ-^{32}P using T4 Polynucleotide Kinase from Invitrogen (Burlington, ON, Canada). The labelling reactions included 10μmol of probe, 2μL of 5X forward reaction buffer, 10U of T4 Polynucleotide Kinase, 2μL of ATP γ-^{32}P (3000Ci/mmol, NEN), and 4μL of water. The reactions were incubated at 37°C for 1 hour.

2.2.3.2 Hybridization

Hybridization of the probes onto the BAC filters was performed in a Techne Roller-Blot Hybridizer HB-3D hybridization oven using 45mm diameter hybridization tubes.

2.2.3.2.1 Prehybridization

50mL of prehybridization solution (5X SSC, 5X Denhardt's, 0.1% SDS) was added to each of two hybridization tubes containing 3 filters separated by mesh inserts. The tubes were allowed to rotate at 65°C for 3 hours.

2.2.3.2.2 Hybridization

Half of each probe solution was added to each of two hybridization tubes. The tubes were allowed to rotate overnight (18 hours) at 65°C.

2.2.3.2.3 Washes

To remove unhybridized labelled probe, three 30 minute washes were performed at 50°C, the first consisting of 2X SSC and 0.1%SDS, and the next two consisting of 1X SSC and 0.1% SDS. The BAC filters were then taken out of the hybridization tubes, wrapped in saran wrap, and exposed to Phosphor screens (Amersham Biosciences, New Jersey, USA) for 18 hours.

2.2.3.3 Positive Identification

Phosphor screens were scanned using a Storm Phosphoimager (Amersham Biosciences, New Jersey, USA).

2.2.4 PCR Confirmation of Positives

In order confirm each potential positive, one set of PCR primers able to distinguish between the duplicate metallothionein genes was used to perform colony PCR on each of the clones. All PCR amplifications were performed in a T3 Thermocycler (Biometra, Goettingen, Germany), in a volume of 25µL, using the following procedure: initial denaturation step, 95°C for 5 minutes, 35 cycles with a denaturation step at 95°C for 1 minute, annealing at the Tm of 55°C for 1 minute, extension at 72°C for 1 minute, followed by a final extension at 72°C for 5 minutes. For each DNA amplification, 0.05U Taq, 12.5 pmoles of each specific primer, 2.5µL 10X buffer containing 1.5mM $MgCl_2$, and 12.5µmol of dNTPs was mixed together with 5µL of BAC DNA template (overnight culture grown in LB medium with 20ug/ml chloramphenicol, diluted 1/40).

2.2.5 Confirmation of Duplicated Regions

A representative MetA (S0085O16) and MetB (S0188I22) clone were chosen for fluorescent *in situ* hybridization (FISH).

2.2.5.1 Fluorescent *in situ* Hybridization

FISH was performed by Dr. Ruth Phillips at Washington State University.

2.2.6 Shotgun Library Creation

2.2.6.1 Preparation of BAC DNA

BAC DNA for was isolated using the Qiagen Maxiprep kit.

2.2.6.2 Shearing DNA

15µg of BAC DNA was sheared by sonication for each sample. Sonication efficiency was verified by electrophoresis.

2.2.6.3 End-repairing DNA

End-repair reactions were carried out using the Epicentre End-It DNA End-Repair Kit (Madison, Wisconsin, USA). All end-repair reactions were carried out at room temperature in a 50µL final volume. 5µL each of dNTP mix, ATP, and 10X buffer from the kit were mixed with 1µL of End-Repair Enzyme Mix and 34µL of DNA, and allowed to incubate for 45 minutes. The reactions were terminated by heating the samples at 70°C for 10 minutes.

2.2.6.4 Size Fractionation and DNA Extraction from Gel

Sonicated, end-repaired BAC DNA was size fractionated by gel electrophoresis, using a 1% agarose gel made with 1X TAE. Ethidium bromide was not added to the gel, and upon completion of the electrophoresis, the ladder lane was cut off from the rest of the gel and allowed to stain in a 0.5µg/ml ethidium bromide solution for 30 minutes. The ladder was then visualized under ultraviolet light, and notches were made in the gel between 2-5Kb. The remainder of the original gel was then compared to the notched ladder portion, and DNA between 2 to 5kb was cut out from the gel. DNA was then purified from the gel using the Qiagen Gel Extraction Kit. DNA concentration was then measured by spectrophotometry and the samples were diluted to 10ng/µL.

2.2.6.5 Ligation

DNA was ligated into *Sma*I-digested, phosphatase-treated, pUC19 and used to transform *E. coli.*

2.2.6.5.1 Preparation of Vector

10µg of circular pUC19 vector was digested with SmaI (20U) by incubating at 37°C for 2 hours. The sample was heated at 65°C for 15 minutes, placed on ice for 5 minutes, and then purified with the Qiagen PCR purification kit. The ends of the vector were then dephosphorylated with 28.5U of Shrimp Alkaline Phosphatase (USB) by incubating at 37°C for 1 hour. The sample was heated at 65°C for 15 minutes, placed on ice for 5 minutes, and then purified with the Qiagen PCR purification kit. A test ligation and transformation was done to ensure vector quality.

2.2.6.5.2 Ligation

Ligation reactions were carried out using T4 DNA Ligase from Invitrogen in 20µL reaction volumes. To each reaction tube was added 10U of ligase, 100ng of insert DNA, 20ng of *Sma*I digested, phosphatase-treated pUC19 vector, and 4µL of 5X Reaction buffer. The samples were incubated overnight (18-20 hours) at 14°C, and then stored at -20°C until they were used for transformation reactions.

2.2.6.6 Transformation

2.5μL of each ligation reaction was used to transform 100μL of XL-1 Blue Competent Cells (Stratagene, La Jolla, California, USA) using a heat shock method. The cells were first allowed to thaw while sitting on ice. 4μL of beta-mercaptoethanol was added to each tube, followed by a 10 minute incubation on ice. 2.5μL of each ligation reaction was mixed with 100μL of competent cells and then incubated on ice for 30 minutes. The cells were heat shocked for 45 seconds at 42°C and then incubated on ice for 2 minutes. After the addition of 900μL of SOC medium, an outgrowth step was performed by shaking at 250rpm for 1 hour at 37°C. Finally, 250μL of each sample was transferred onto a 14cm LB plate containing 200μg/mL ampicillin. 375μL of X-gal (20mg/ml) and 225μL of IPTG (200mg/ml) were also transferred onto the plates, and the resulting mixture was spread evenly on the plate surface. The plates were then incubated for 20 hours at 37°C.

2.2.6.7 Insert Size Check

A colony PCR method, using an M13 forward (M13F) primer (5' GTA AAA CGA CGG CCA GT) and an M13 reverse (M13R) primer (5' CAG GAA ACA GCT ATG AC), was employed to check the insert size of several colonies for both libraries. In a volume of 25μL, the following procedure was performed: initial denaturation step, 95°C for 5 minutes, 35 cycles with a denaturation step at 95°C for 1 minute, annealing at the Tm of 50°C for 1 minute, extension at 72°C for 5 minute, followed by a final extension at 72°C for 5 minutes. For each DNA amplification, 0.05U Taq, 12.5 pmoles of each specific primer, 2.5μL 10X buffer containing 1.5mM $MgCl_2$, and 12.5μmol of dNTPs was mixed together. An isolated, white colony was picked with a yellow tip and dipped and swirled in the PCR mix to provide the template for amplification. 5μL of PCR product was electrophoresed on a 1% agarose gel made with 1X TBE.

2.2.6.8 Sequencing to Determine Library Quality

The percent of *E. coli* genomic DNA contamination in each library was estimated by sequencing several representative clones. To do so, the remaining 20μL of PCR product from above was purified using the Qiagen PCR Cleanup Kit and used for sequence analysis with the fluorescent dideoxy terminator method. Briefly, 3μL of cleaned-up PCR product was mixed with 5pmol of primer (M13F or M13R), and 4μL of sequencing premix in a final volume of 10μL. Sequencing reactions were carried out with the following conditions: 94°C 20sec, Tm (50°C) 20sec, 60°C 1min, for 30 cycles. Following amplification, unincorporated dideoxynucleotides were removed from each reaction mixture by ethanol precipitation, and the reactions were resuspended in 2uL of diformamide loading dye. Sequence analysis was carried out on an ABI 377 DNA Sequencer (Applied Biosystems) using DYEnamic ET Terminator chemistry (Amersham Biosciences). The resulting sequences were then compared to the non-redundant nucleotide database at GenBank using BLASTN and the number of clones containing E. coli genomic DNA was determined.

2.2.7 Large Scale Sequencing of Libraries

Approximately 2000 randomly selected subclones were sequenced in both directions from each BAC shotgun library, giving a total of 4000 sequence reads for each BAC clone. Assuming an average read length of 500 bp and a BAC insert size of 200 kb, this corresponds to a 10-fold sequence coverage for each BAC. The sequencing was performed at the Genome British Columbia sequencing platform at the University of Victoria.

2.2.8 Building Contigs

The sequence reads were analyzed using the Phred/Phrap/Consed suite of programs. Crossmatch was run to filter out sequence reads containing *E.coli* genomic DNA as well as the vector sequence (pTARBAC2.1).

2.2.9 Sequence Finishing

For gap filling, a primer-walking procedure was employed. Unique primers were designed at the both ends of each contig and tested for their ability to amplify corresponding BAC DNA. PCR products were then purified using the Qiagen PCR Cleanup Kit, and sequencing reactions were carried out as detailed above.

2.3 Sequence Annotation

2.3.1 Pairwise Dot Plot Alignment

To analyze the similarity between the two BAC sequences, a dot plot was created using the program Dotter (Sonnhammer and Durbin, 1995) with the window length set to 22.

2.3.2 Gene Identification

Two gene predication programs were utilized to predict genes in the BAC sequences: GENSCAN (Burge and Karlin, 1997; http://genes.mit.edu/GENSCAN.html) and Eukaryotic GeneMark (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi). The deduced protein sequences were compared to the RefSeq protein database (Pruitt et al., 2003) using BLASTP (Altschul et al., 1990). Each BAC sequence was also compared to the non-redundant nucleotide database at GenBank via BLASTN searches. Finally, each BAC sequence was compared to the non-redundant protein database at GenBank via BLASTX searches. The BAC sequences were then compared to the expressed sequence tag (EST) databases for rainbow trout (release 4.0) and Atlantic salmon (release 2.1; Rise et al., 2004) available at The Institute of Genomic Research (TIGR) (http://www.tigr.org/) via BLASTN searches. All searches utilized BLAST version 2.2.10 released October 19. 2004. The searches were performed between November 1st and 18th, 2004.

Exons for each gene were annotated based on similarity with known and predicted gene in the genomes of other sequenced organisms. The genome annotations available through the Ensembl Genome Browser (Stalker et al., 2004; http://www.ensembl.org/) were used for this work. Genomes included in the analysis were human (release 26.35.1), mouse (release 26.33b.1), rat (release 26.3d.1), chick (release 26.1c.1), zebrafish (release 26.4a.1), *Takifugu* (release 26.2c.1), and *Tetraodon* (release 26.1a.1).

2.3.3 Repeat Analysis

Three databases available from RepeatMasker (www.repeatmasker.org) were used to identify known repetitive elements in the BAC sequences: (1) Zebrafish, (2) Pufferfish, and (2) Fish other than Zebrafish and Pufferfish.

A salmonid specific database that is currently being developed in our lab by Siemon Ng was then utilized to identify salmonid specific repeats in the BAC sequences. Briefly, this repeat database was built using approximately twenty Atlantic salmon complete BAC sequences and all known salmonid ESTs.

Finally, to identify internally repeated sequences, each BAC sequence was aligned against itself and visualized using PipMaker plots (Schwartz et al., 2000).

2.3.4 Conservation of Synteny

Gene order and transcriptional orientation of orthologous genes of interest in human, mouse, rat, chick, zebrafish, *Takifugu*, and *Tetraodon* were determined using the annotated genomes available through the Ensembl Genome Browser (see section 2.3.2 for the genome release).

2.3.5 Multiple Sequence Alignments

All nucleotide and amino acid sequence alignments were generated with Clustal W using default parameters (Thompson et al., 1994). Sequence alignments were corrected by eye in Seaview (Galtier et al., 1996).

CHAPTER 3 RESULTS

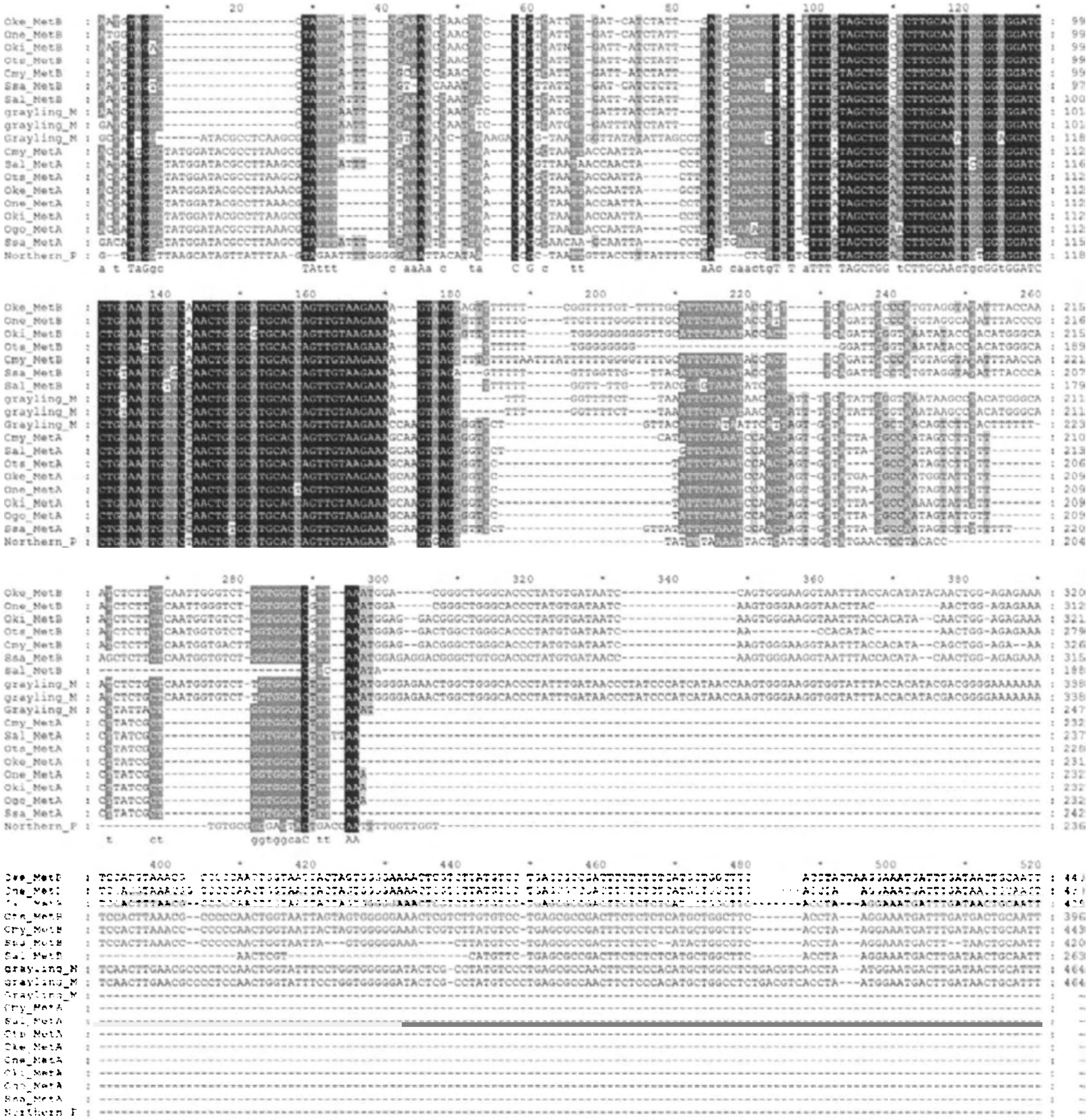
3.1 Salmonid Metallothionein Genes

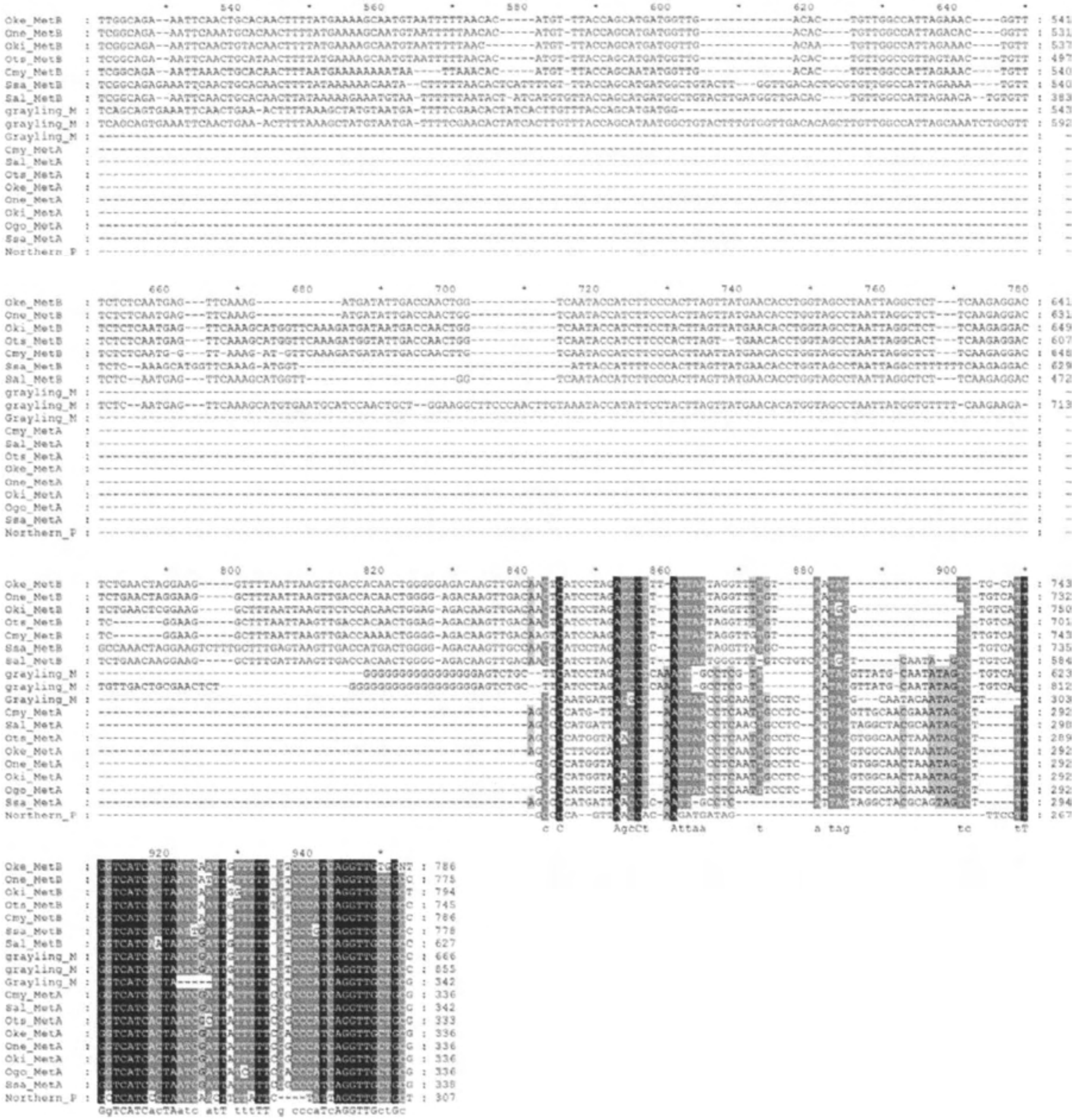
Metallothionein-specific primers were designed in order to highlight the differences in PCR amplification products for the two loci such that they spanned the single amino acid coding change in exon two as well as the longer intron 2 characteristic of MetB. Genomic DNA from various salmonid species was used for PCR amplification, including chum salmon (*Oncorhynchus keta*), sockeye salmon (*O. nerka*), pink salmon (*O. gorbuscha*), coho salmon (*O. kitsutch*), chinook salmon (*O. tshawytscha)* and grayling (*Thymallus arcticus*). Each of the *Oncorhynchus* species gave two amplification products of the expected sizes (~400bp and ~900bp), which were subsequently subcloned and sequenced. Amplification of the grayling genomic DNA yielded three bands of 400bp, 900bp, and 1000bp. Cloning and sequencing revealed that both of the larger bands were, in fact, representative of a metallothionein B gene.

Several metallothionein sequences were also available in GenBank; MetA and MetB sequences for Arctic charr (*Salvelinus alpinus*) and rainbow trout (*O. mykiss*) were downloaded, as was the single metallothionein sequence for Northern pike (*Esox lucius*). The Atlantic salmon sequences were determined from the BAC shotgun library sequencing (see section 3.7.1.1).

A multiple sequence alignment was performed (Figure 2) and a neighbour joining tree was drawn using Northern pike as the outgroup (Figure 3).

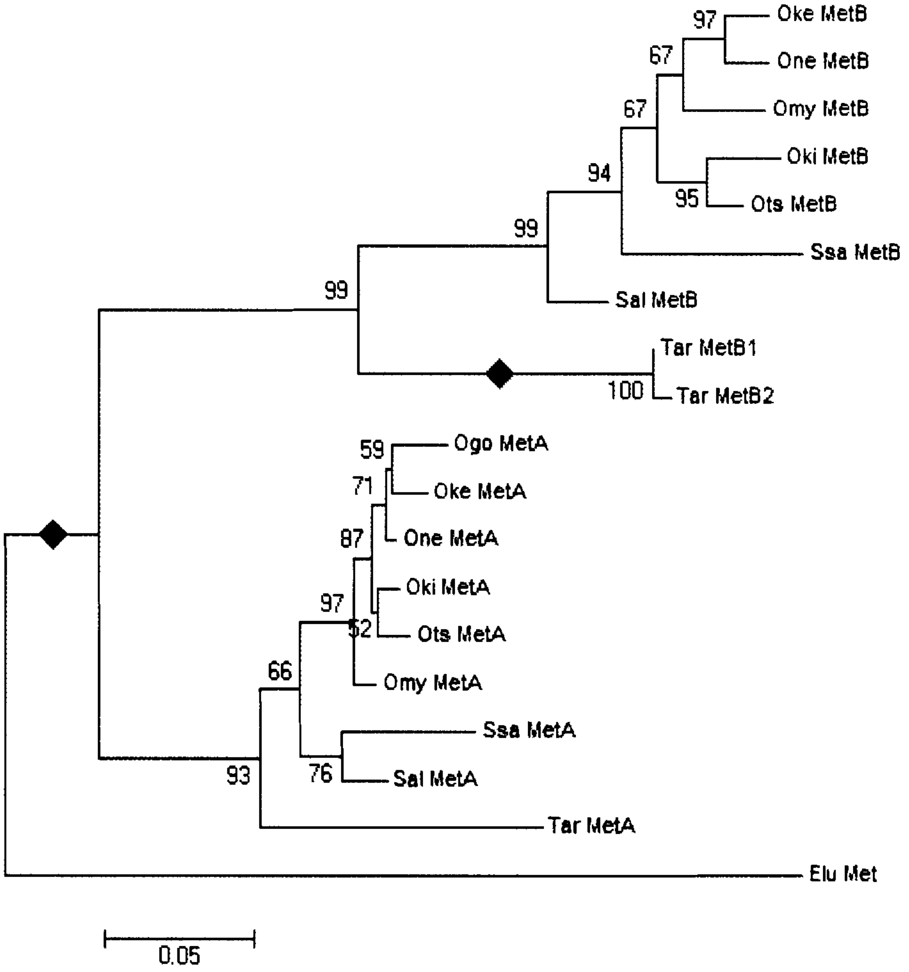
Figure 2 **Multiple sequence alignment of partial metallothionein sequences.** Partial MetA and MetB sequences, representing intron 1, exon 2, and intron 3, from various salmonids as well as the single metallothionein sequence from Northern pike were aligned.





(Legend: Ogo: *Oncohynchus gorbuscha*, pink salmon; Oke: *Onchorhynchus keta*, chum salmon; Omy: *Oncorhynchus mykiss*, rainbow trout; One: *Oncorhynchus nerka*, sockeye salmon; Ote: *Oncorhynchus tshawytscha*, chinook salmon; Tar: *Thymallus arcticus*, Artic grayling; Ssa: *Salmo salar*, Atlantic salmon; Sal: *Salvelinus alpinus*; Arctic charr; Elu: *Esox lucius*, Northern Pike.)

Figure 3 Neighbour-joining tree based on partial metallothionein sequences. Based on the topology of the tree, it is believed that the metallothionein gene was duplicated in the ancestor common to all salmonids (black diamond). This event is consistent with the timing of the salmonid-specific tetraploidization event. An additional duplication event occurred in the grayling lineage leading to the duplicate copies of metallothionein B (grey diamond).



Ogo: *Oncorhynchus gorbuscha*, pink salmon

Oke: *Oncorhynchus keta*, chum salmon

Oki: *Oncorhynchus kitsutch*, coho salmon

Omy: *Oncorhynchus mykiss*, rainbow trout

One: *Oncorhynchus nerka*, sockeye salmon

Ots: *Oncorhynchus tshawytscha*, chinook salmon

Tar: *Thymallus arcticus*, Arctic grayling

Ssa: *Salmo salar*, Atlantic salmon

Sal: *Salvelinus alpinus*, Artic charr

Elu: *Esox lucius*, Northern pike

The topology of the tree revealed two separate clusters of the salmonid MetA and MetB genes. The grayling sequences were the farthest removed within their respective groups and the *Onchorhynchus* species all grouped together. The placement of the Atlantic salmon (Ssa) and Artic charr (Sal) branches is not consistent between the MetA and MetB groups, but the bootstrap values in the MetA branch are not particularly high. Interestingly, the two grayling MetB sequences (Grayling MetB1 and MetB2) grouped together with a bootstrap value of 100, indicating the occurrence of a MetB duplication event specific to the grayling lineage. The topology of the tree can be taken as evidence for an early duplication of the metallothionein gene in the ancestor of the salmonids, suggestive of and consistent with a whole genome duplication.

3.2 Identification of Metallothionein BACs

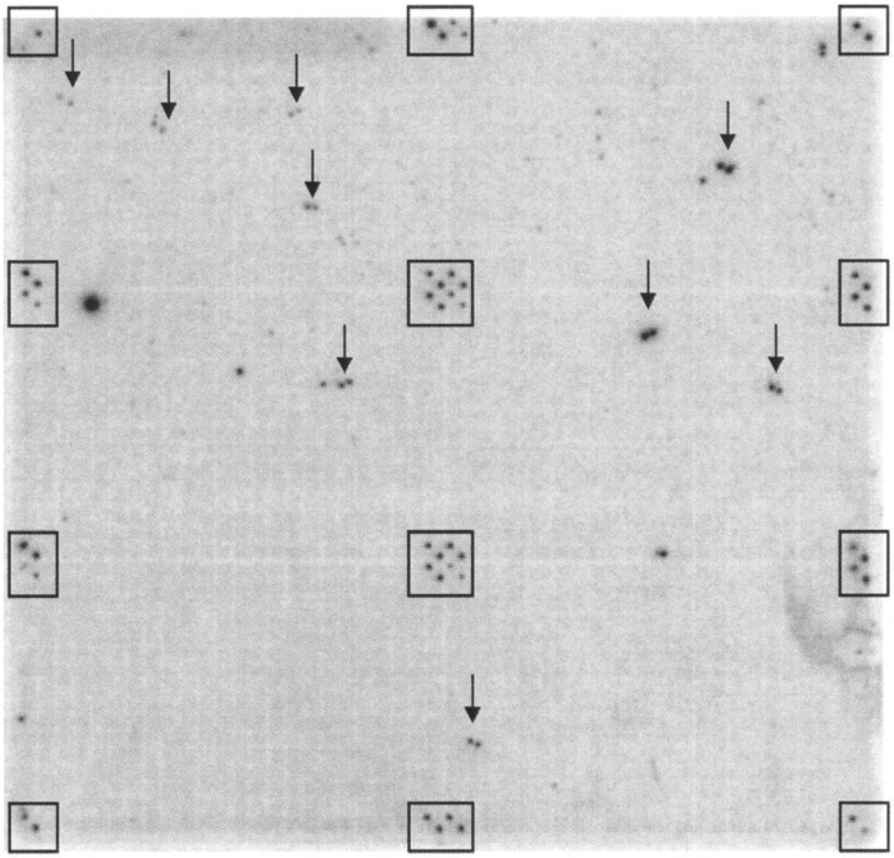
A large insert, Atlantic salmon genomic BAC library was available as a tool for isolating large segments of DNA that contain the duplicate metallothionein genes. The complete BAC library has been stamped onto 17 nylon filters for hybridization purposes and provides an 18.8-fold coverage of the Atlantic salmon genome. The first two-thirds of the library have been *Hind*III fingerprinted and placed into contigs to form a physical map with an approximately 12-fold genomic coverage.

For this work, the first six filters in the library, corresponding to the first 288 plates of the library and providing a roughly 6-fold genome coverage, were screened with a probe designed in the third exon of the Atlantic salmon metallothionein sequences; this probe was designed to hybridize to both MetA and MetB sequences. The hybridization resulted in the identification of thirty-three BAC clones, nineteen belonging to contig 341 in the physical map, nine falling into contig 2172, three clones that were singletons, and two whose digests were unsuccessful (Table 1). A typical BAC filter hybridized with the metallothionein probe and the control *C. briggsae* DNA is shown in Figure 4.

Table 1 **Metallothionein Positive BAC Clones.** 33 BAC clones falling into two contigs were identified as metallothionein-positive by hybridization. Classification as MetA or MetB was carried out using PCR.

BAC Clone	Contig	Gene
S0001H02	341	MetB
S0002G01	341	MetB
S0030G12	Undigested	MetB
S0043E01	341	MetB
S0049K22	341	MetB
S0054B07	341	MetB
S0078B05	Singleton	MetA
S0085O16	2172	MetA
S0093H07	Singleton	MetA
S0100C24	2172	MetA
S0100O22	341	MetB
S0111O06	341	MetB
S0114E19	341	MetB
S0116L01	Singleton	MetA
S0120L02	341	MetB
S0123O23	341	MetB
S0124A24	341	MetB
S0130J09	341	MetB
S0134N02	341	MetB
S0139F17	2172	MetA
S0161B10	341	MetB
S0167B06	2172	MetA
S0169C02	2172	MetA
S0171O05	2172	MetA
S0177C09	2172	MetA
S0188I22	341	MetB
S0217I23	2172	MetA
S0224F05	341	MetB
S0231K08	2172	MetA
S0242O03	341	MetB
S0277K06	341	MetB
S0284H09	Undigested	MetA
S0288K06	341	MetB

Figure 4 BAC filter hybridized with a metallothionein probe and the control *C. briggsea* DNA. Positive clones (denoted by arrows) are recognizable because they appear as duplicates in one of eight set orientations. The *C. briggsea* orientation spots (boxes) allow for membrane orientation after hybridization.



The locations of the metallothionein positive clones found in contigs 341 and 2172 were examined. In contig 341, the 19 positive clones fall on the right hand side of the contig, one on top of another (Figure 5). Having hybridized the metallothionein probe to only a subset of the BAC library (the first six filters), only positive clones from the first 288 plates in the library were identified, which explains why several clones that are also in this region were not identified as positives by hybridization. Additionally, the fact that the S0205M03 clone was not found to be positive allowed for the location of the metallothionein gene to be narrowed down; the metallothionein gene must lie in the overlapping area of S0188I22 and S0130J09. Similarly, contig 2172 (Figure 6) shows the positive clones highlighted in yellow. In this case, two clones that were screened for (S0170K02 and S0177G07) were not identified as positive, and are probably false negative.

Figure 5 **Contig 341.** BAC clones in contig 341 that were identified by hybridization are highlighted in yellow.

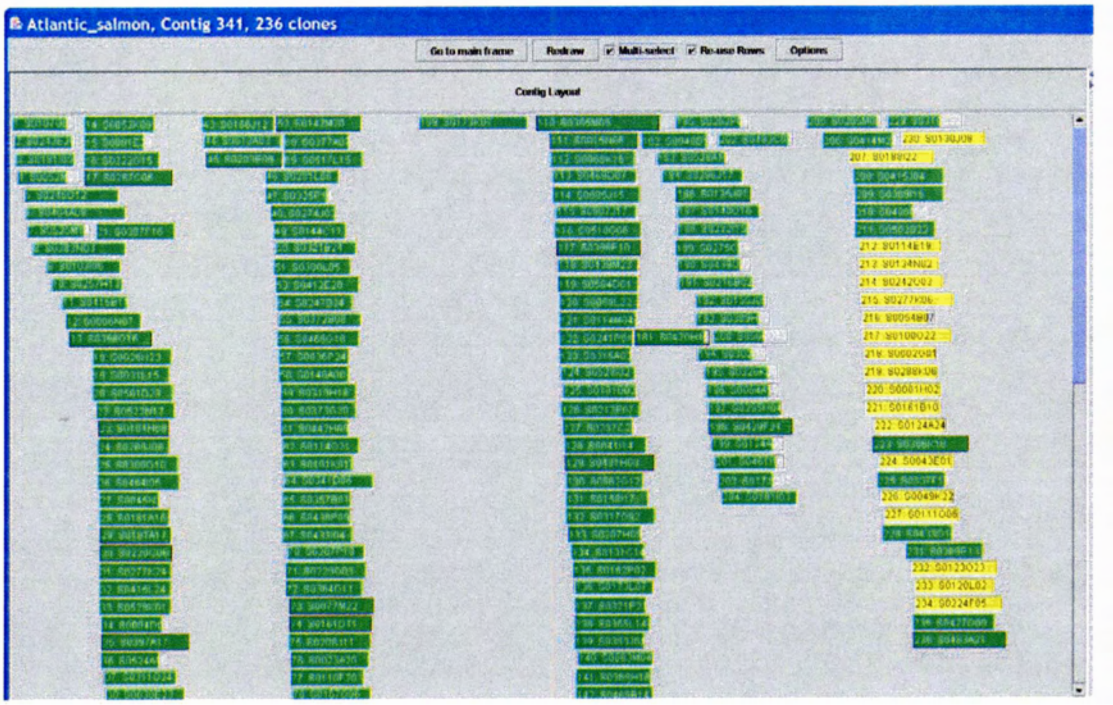
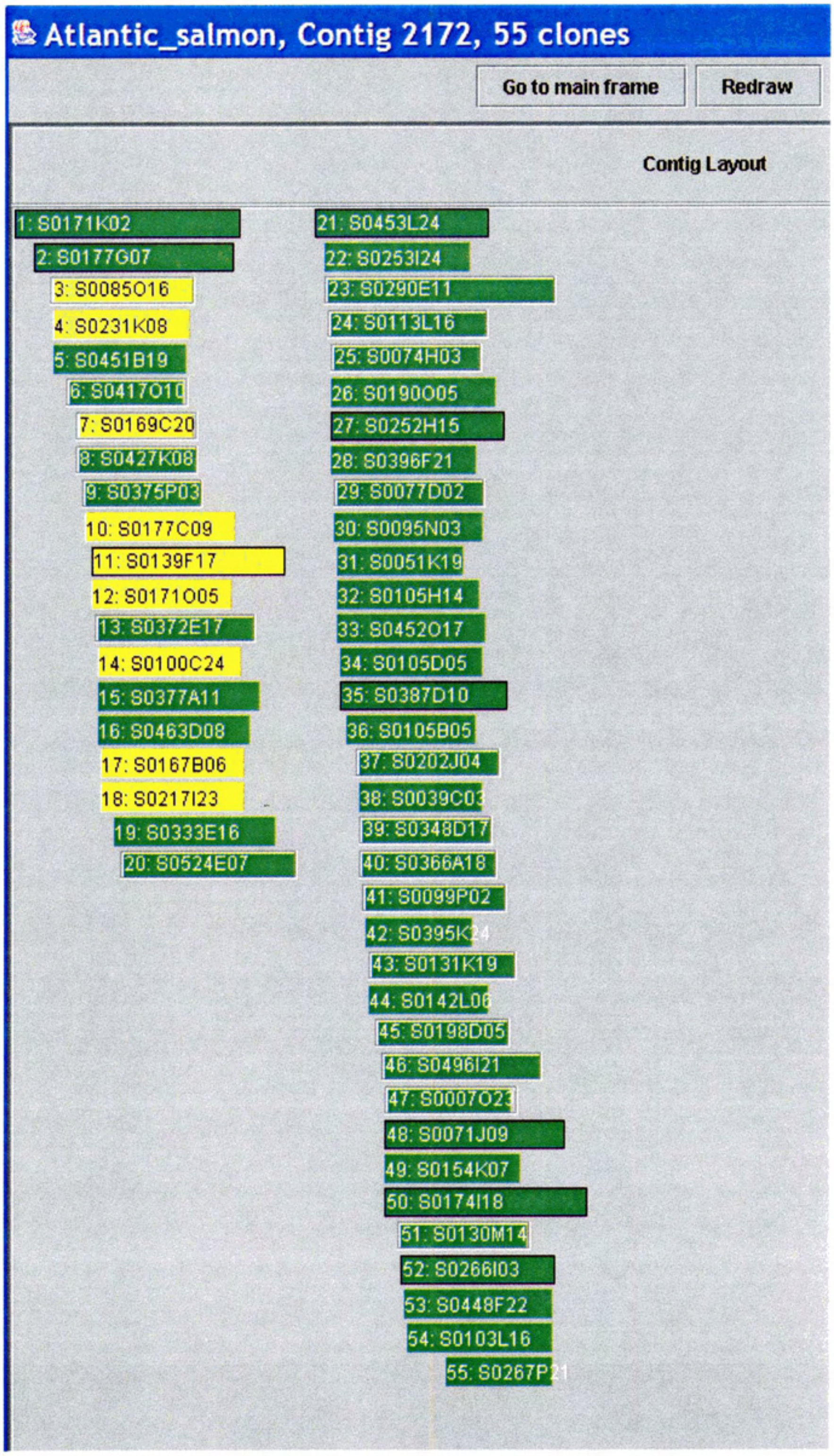


Figure 6 **Contig 2172.** BAC clones in contig 2172 that were identified by hybridization are highlighted in yellow.



Positive BAC clones were then PCR amplified with metallothionein-specific primers for two purposes: first, to provide confirmation of the presence of the metallothionein gene in each positive BAC, and second, to distinguish between MetA and MetB-containing BACs. All of the positive clones identified in contig 2172 had amplification products of the size expected for MetA (~450 base pairs bp) (Figure 7). Similarly, the positive BAC clones from contig 341 had PCR products corresponding to MetB (~900 bp) (Figure 7). Furthermore, the three singletons and one of the non-digested clones corresponded to MetA, and the other non-digested clone corresponded to MetB. The *Hind*III fingerprint of each positive clone was examined with respect to the other corresponding positives to verify the contig builds (Figure 8 and Figure 9). Special attention was paid to the *Hind*III digest patterns of the three singletons with respect to the fingerprints of the clones in contig 2172. Multiple bands in common between each of the singletons and the BACs in contig 2172 provided further evidence for the inclusion of the three singletons in this contig.

Figure 7 PCR confirmation and identification of MetA and MetB positive contigs. Representative MetB positive BAC clones from contig 341 have amplification products of approximately 900bp, whereas the MetA positive BACs only have products that are 450bp in size.

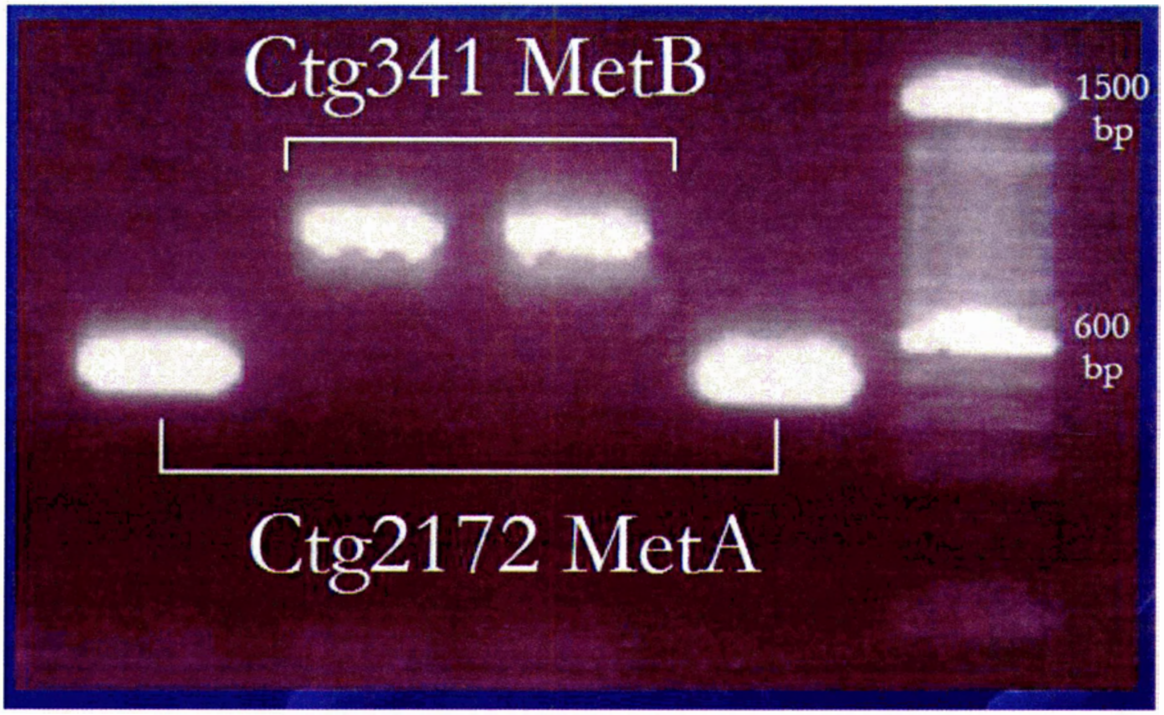
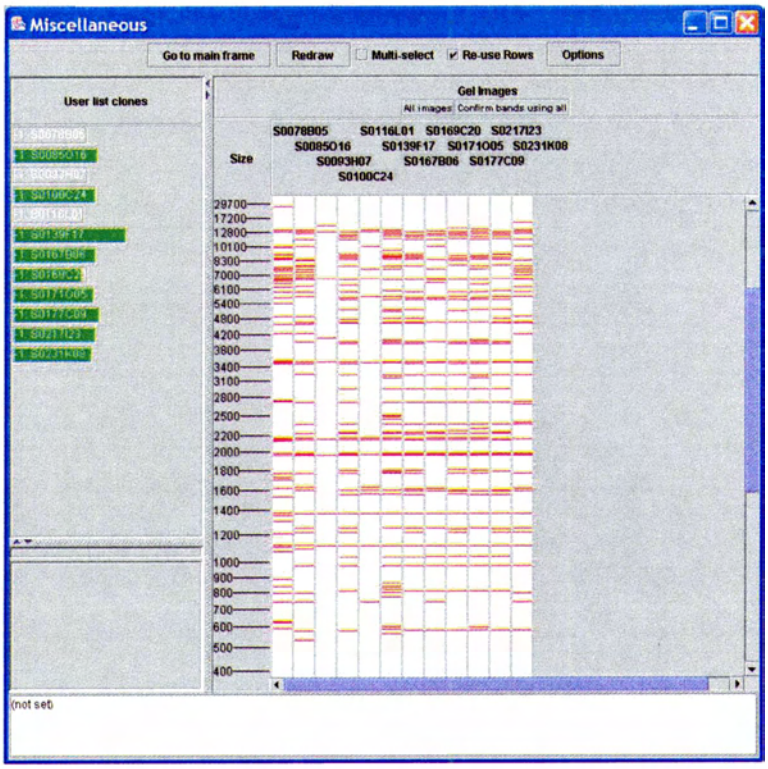
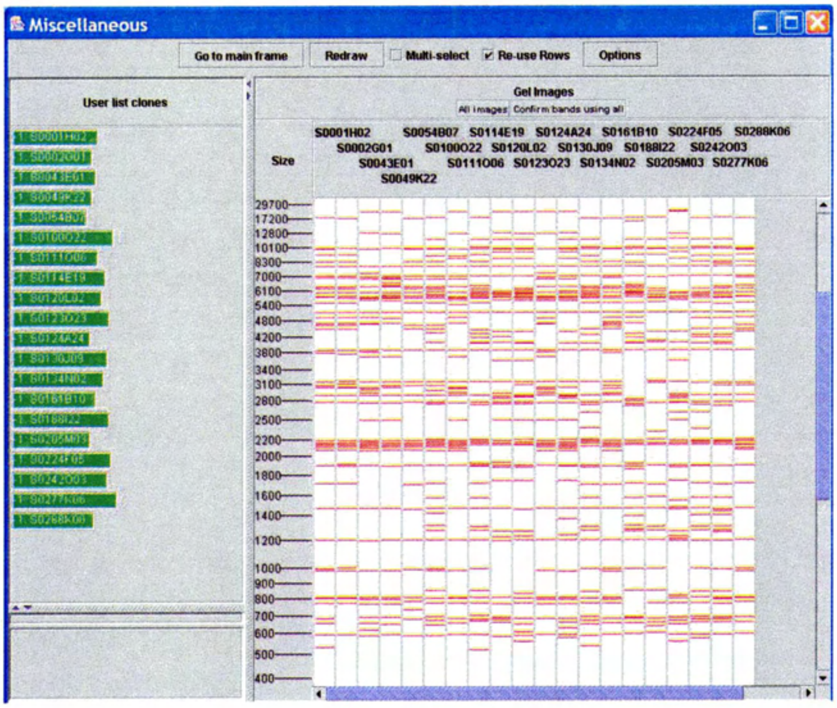


Figure 8 ***Hind*****III fingerprints for MetA positive clones.** All clones with PCR amplification products corresponding to MetA are shown.



☐ Multi-select ☑ Re-use Rows

Figure 9 ***Hind*III fingerprints for MetB positive clones.** All clones with PCR amplification products corresponding to MetB are shown.



3.3 Confirmation of Duplication

To test that the metallothionein genes are present on different chromosomes, fluorescent *in situ* hybridization (FISH) was performed on Atlantic salmon metaphase chromosomes by Dr. Ruth Phillips at Washington State University. A representative clone was selected from each contig for FISH analysis; S0085O16 was selected from contig 2172 to represent MetA, and S0188I22 was chosen from contig 341 to represent MetB. These BACs were chosen because they fall in the middle of the metallothionein positive region in their respective contigs. FISH results confirmed that the duplicate metallothionein genes do indeed reside on different chromosomes; MetA is on chromosome 17 (Figure 10) and MetB is on chromosome 11 (Figure 11).

Figure 10 MetA FISH. S0085O16, which is representative of the MetA BAC positive clones, hybridizes to chromosome 17.

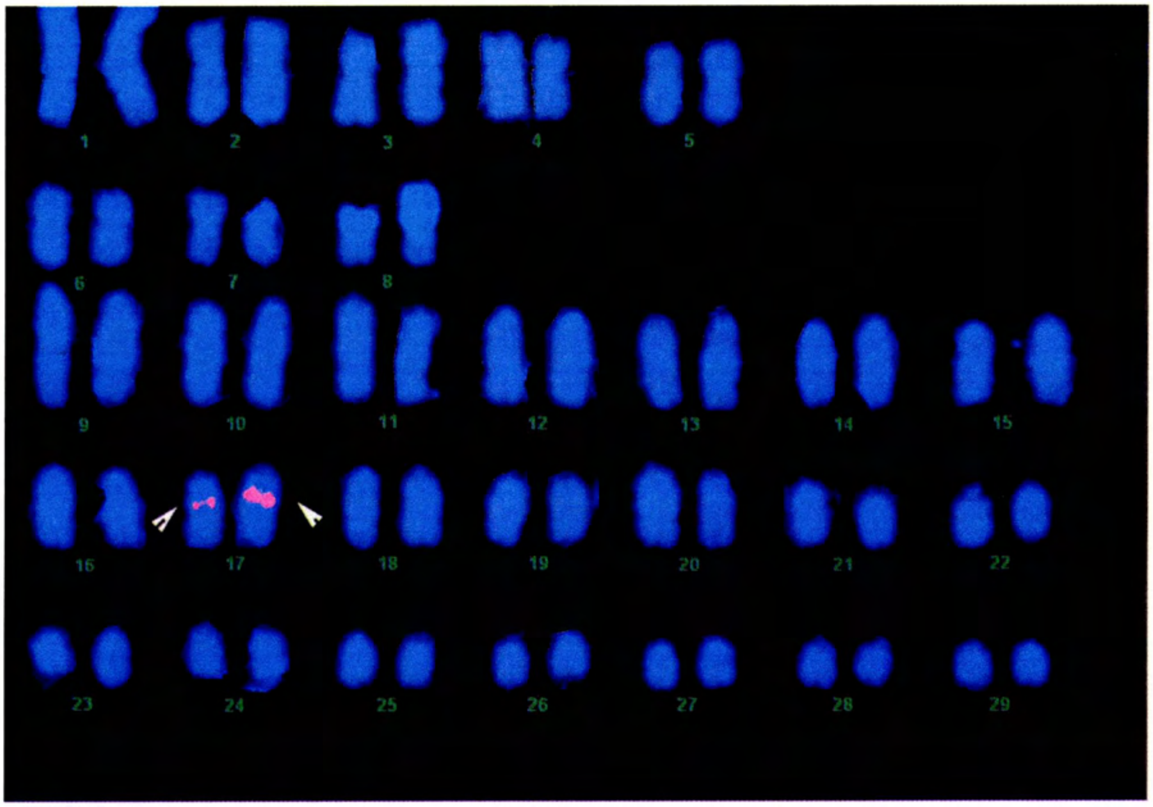
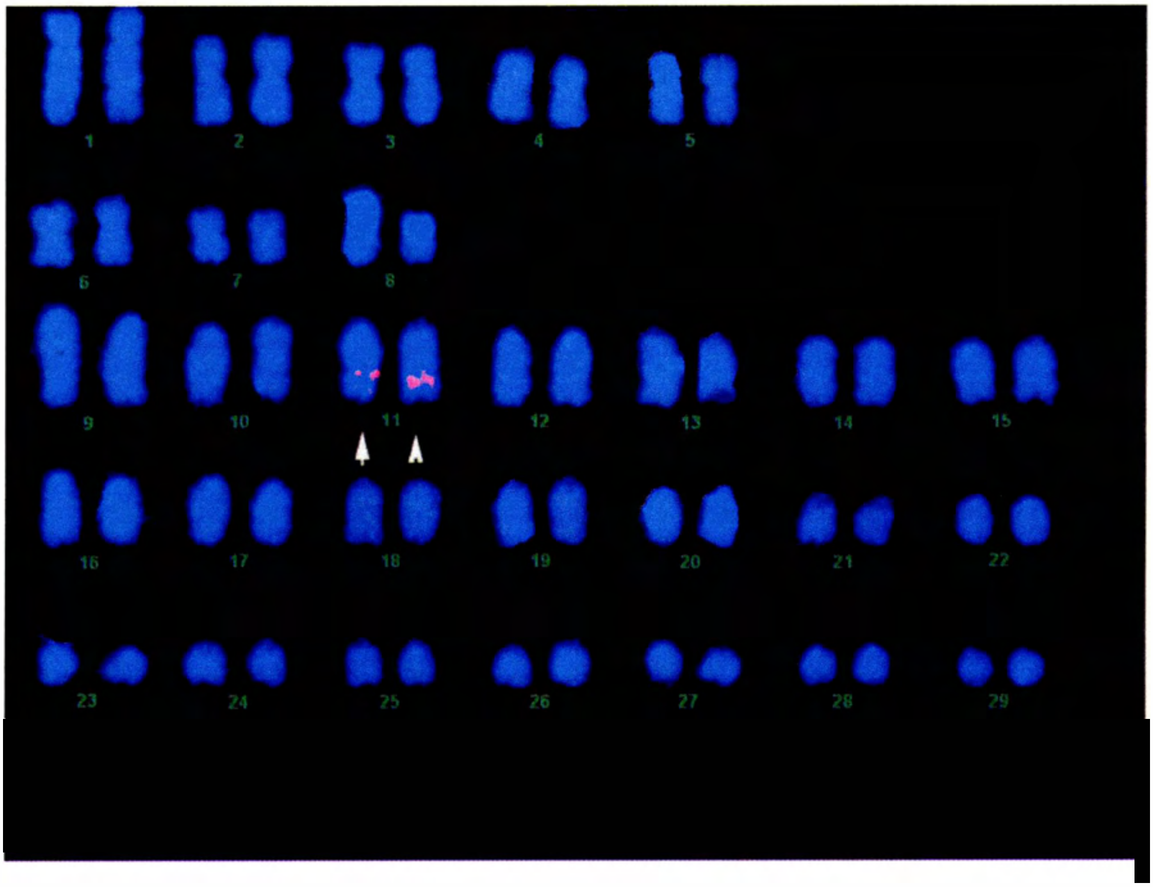


Figure 11 MetB FISH. S0188I22, which is representative of the MetB positive BAC clones, hybridizes to chromosome 11.



3.4 Shotgun Libraries of Metallothionein BAC Clones

The two BACs that were used for FISH analysis, S0085O16 (MetA) and S0188I22 (MetB), were subsequently selected for shotgun library creation. After BAC DNA isolation, shearing by sonication, and end-repair, the DNA was size fractionated by agarose gel electrophoresis to isolate DNA fragments ranging in size from 2 to 5kb. These fragments were then ligated into *SmaI*-digested, phosphatase-treated, pUC19 and used to transform *E. coli*.

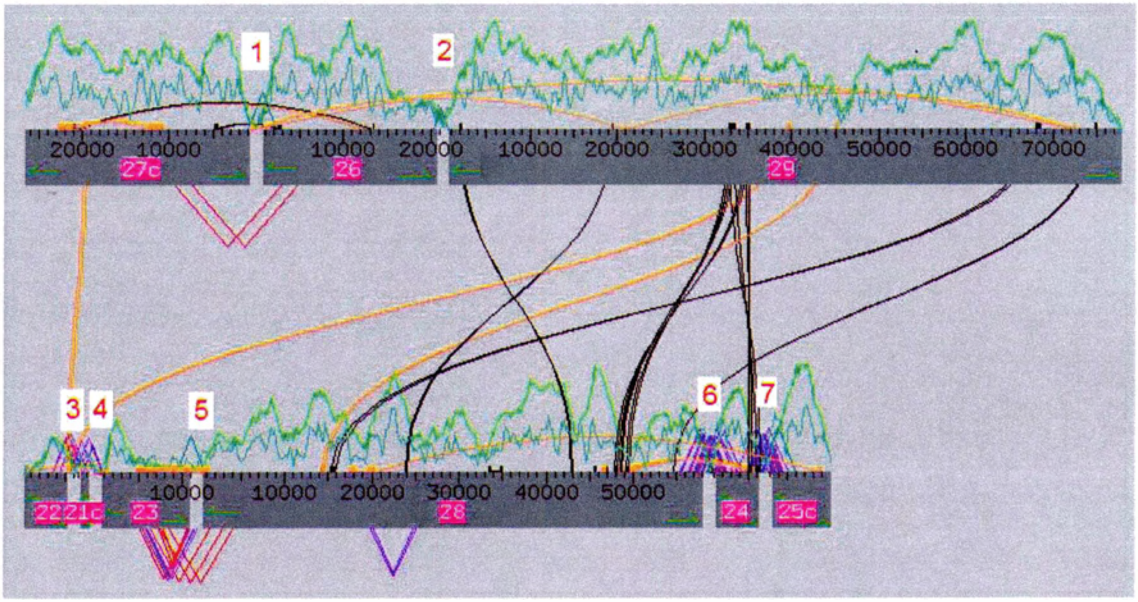
Approximately 2000 clones were picked randomly from each BAC shotgun library and end-sequenced from both ends at the Genome British Columbia sequencing platform at the University of Victoria. This gave approximately 4000 sequence reads for each library. Assuming an average read length of 500 bp and a BAC insert size of 200 kb, this corresponds to a 10-fold sequence coverage for each BAC.

3.5 BAC Contig Construction

The Phred/Phrap/Consed suite of programs was used to build contigs from the sequence data. Contigs were ordered and oriented based on read-pair associations of gap-spanning subclones. Antisense primers were designed to anneal to the 5' end of each contig and sense primers were designed to anneal to the 3' end of each contig. Primer pairs were then used to amplify BAC DNA in order to establish the orientation of the contigs relative to one another and to enable sequencing of these gaps.

For S0085O16, the initial build by Phrap yielded nine contigs that were ordered into two supercontigs based on forward and reverse read consistency (Figure 12).

Figure 12 S0085O16 initial build. Using the Assembly View function in Consed, the orientation of the nine contigs (grey bars) is visualized. Purple and red lines located on top of the contig bars represent forward and reverse read pairs that are consistent distances from one another, and inconsistent distances when shown below the contig bar. Orange lines represent repeats going in the same direction and black indicate inverted repeat sequences. Green arrows represent the locations of primers designed to amplify across the sequence gaps. Numbers one through seven indicate the gaps. Finally, depth of sequence coverage is shown by the duller green line, and forward/reverse read consistency by the brighter green line.

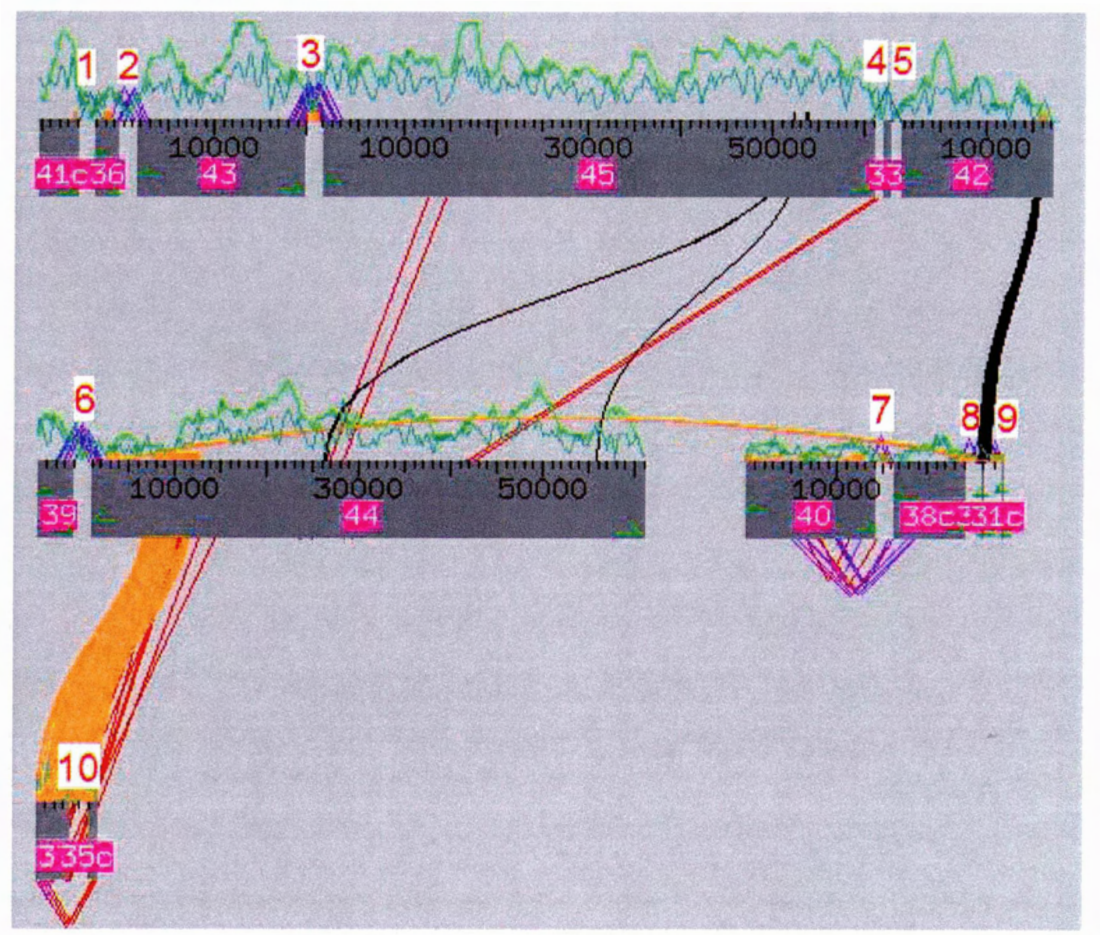


Using the specifically designed primers, gaps 1, 2, and 3, were successfully confirmed and the intervening DNA sequenced. Approximately 200 bp of sequence data were required to fill each of the three gaps. Gap 7, on the other hand, was confirmed by PCR to be about 300 bp, but due to the repetitive nature of the ends of both contigs, the entire sequence of the gap has not yet been determined. 300 Ns have been inserted into the overall BAC sequence to represent this gap. The primers designed to amplify gaps 4 and 6 were unsuccessful in amplifying the BAC DNA, and therefore should be redesigned. 100 bp of Ns have been inserted into each of these gaps to represent the unknown sequence. For gap 5 it was not possible to design specific primers since the gap sits in the middle of a10 kb stretch of repetitive DNA. This bit of sequence will be filled in with 100 bp of Ns to represent the unknown gap.

The two supercontigs for S0085O16 were then oriented with respect to one another by testing PCR primer combinations from the four possible contig orientations. It was determined that the primer set corresponding to the 3' end of contig 29 and the 5' end of contig 22 yielded a 3kb amplification product, therefore allowing overall contig orientation. The complete sequence of this 3kb gap is not yet known, and therefore is represented in the final BAC sequence by Ns. In this way, a contiguous stretch of 222,799 bases of DNA sequence has been generated for the S0085O16 MetA-containing BAC.

For the MetB clone, S0188I22, the initial assembly performed by Phrap resulted in 14 contigs oriented into four supercontigs (Figure 13).

Figure 13 S0188I22 initial build. Using the Assembly View function in Consed, the orientation of the fourteen contigs (grey bars) is visualized. Purple and red lines located on top of the contig bars represent forward and reverse read pairs that are consistent distances from one another, and inconsistent distances when shown below the contig bar. Orange lines represent repeats going in the same direction and black indicate inverted repeat sequences. Green arrows represent the locations of primers designed to amplify across the sequence gaps. Numbers one through seven indicate the gaps. Finally, depth of sequence coverage is shown by the brighter green line, and forward/reverse read consistency by the duller green line.



Primers specific for the ends of each contig enabled joining of seven of the gaps (1-7), of which four have successfully been sequenced (1, 2, 3, and 6). The primers for gaps 8, 9, and 10 failed due to the presence of repetitive DNA, where it is extremely difficult to design specific primers. These sequence gaps will be filled with Ns at this time.

Two of the supercontigs have been oriented with respect to one another; primers corresponding to the 3' end of contig 42 and the 3' end of contig 31 yielded a PCR product of 1.6kb. In order to align the remaining two contigs, a few assumptions were made. First of all, the smallest supercontig, composed of contigs 35 and 37, contains the same repetitive DNA as that found in contig 44. As shown in Figure 13 there is a large amount of forward/reverse read inconsistency within this repetitive region, and it is assumed that that this region will collapse into a single, 5kb stretch of repetitive DNA. Secondly, based on gene order (see section 3.6.6), the 3' ends of contigs 40 and 44 will be joined, separated by 1000 Ns. In this way, a contiguous stretch of 205,993 bases of DNA sequence has been built for the S0188I22 MetB-containing BAC.

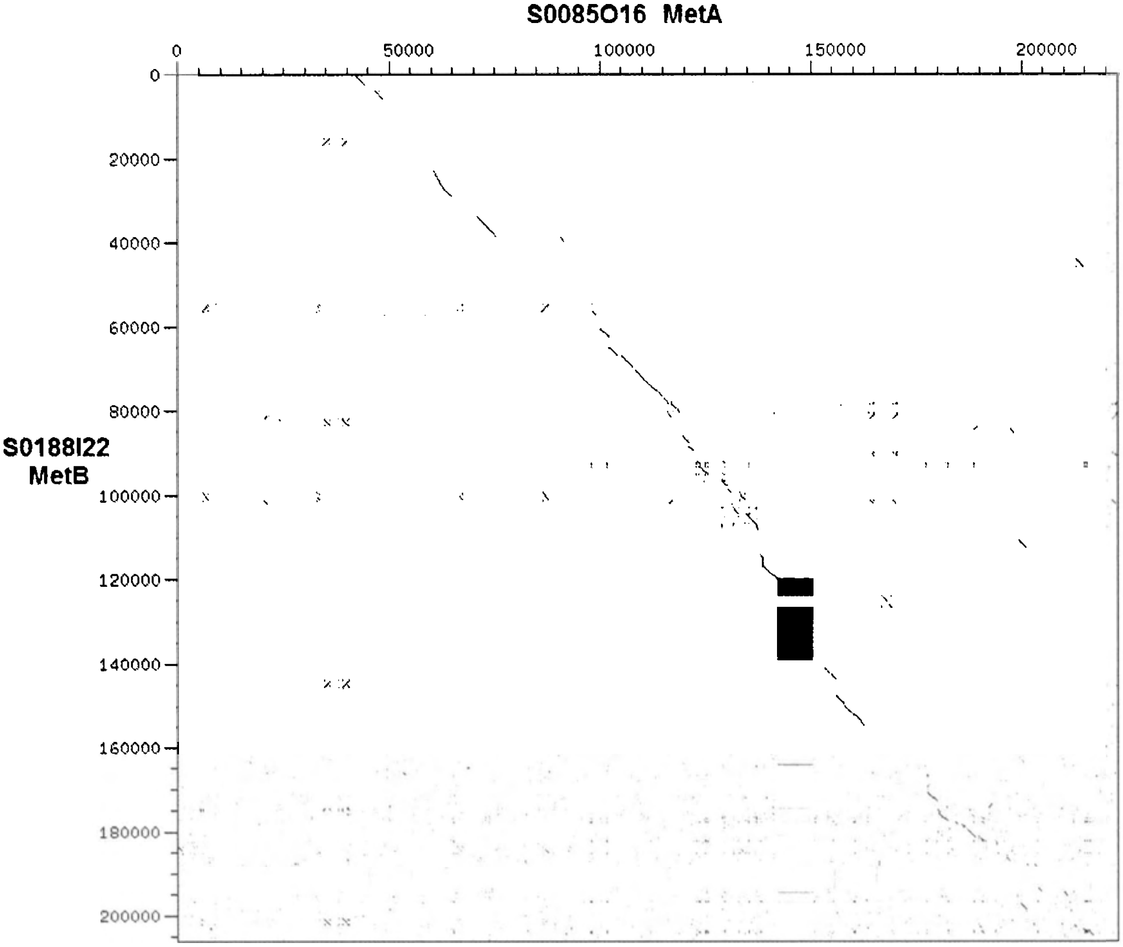
Based on the *Hind*III fingerprints, it was estimated that the S0085O16 and S0188I22 clones would be 211kb and 239kb in length, respectively. After sequence assembly, however, the total number of bases (including Ns) in the S0085O16 sequence was found to be slightly higher than expected, at 222,799 bases. The sequence for S0188I22 was lower than expected at 205,993 bases. The discrepancy between the length of the sequence data and expected sizes of these two clones may be partly attributed to the fact that the sequence gaps, currently filled by Ns, may be either under or over-representing the actual genomic sequence length. Additionally, the large repetitive sequences contained within these BACs may have been built incorrectly.

3.6 BAC Sequence Annotation

3.6.1 Pairwise Dot Plot Alignment

To examine the extent of sequence conservation that exists between the two BACs, the complete S0085O16 MetA and S0188I22 MetB sequences were compared to one another in the form of a dot plot (Figure 14). The diagonal line in the dot plot indicates a high degree of similarity between the two sequences across the full length of S0188I22, and beginning about 50 kb into the S0085O16 sequence. The black square indicates a repetitive region that is common to both BACs. This repeat is about twice as long in the S0188I22 MetB sequence compared to the S0085O16 MetA sequence.

Figure 14 Dot plot of S0085O16 MetA versus S0188I22 Met B. S0085O16 and S0188I22 were compared to one another along the length of each sequence. The diagonal line, indicates that the sequences are quite conserved across the length of S0188I22, starting just before 50kb in S0085O16.



3.6.2 Gene Identification

Initial annotation of the BAC sequences involved identifying the genes existing in each BAC. To begin with, two gene prediction programs were used to execute this task: Genscan and GeneMark. The translated output from each program was subjected to a BLAST search against the non-redundant protein database in GenBank, resulting in hits to several genes that have been characterized or predicted in other organisms; segments of the genes corresponding to SLC12A3 (solute carrier 12, family 3), deadeye (nuclear pore complex protein, 93 kda), BBS2 (Bardet-Biedl Syndrome protein 2), beta-1,3-galactosyltransferase, GNAO1 (guanine-nucleotide-binding protein G(o) subunit 1), CBFB (core-binding factor, beta subunit), Lin10 (unknown function), and Herp (homocysteine-responsive endoplasmic reticulum-resident ubiquitin-like domain member 1) were predicted in both BAC sequences by both Genscan and Genemark, while Cetp (cholesteryl ester transfer protein precursor) was detected only in the S0085O16 sequence, also by both programs (Table 2). Neither of the gene prediction programs predicted the metallothionein gene in either of the BAC sequences.

Comparing the BAC sequences to the non-redundant nucleotide database at GenBank via BLASTN searches, however, identified the metallothionein locus in each sequence. The S0085O16 sequence contains the MetA gene, as predicted by PCR with metallothionein-specific primers. Likewise, the S0188I22 sequence contains the MetB gene.

Table 2 Genes Identified in the BAC Sequences

Acronym	Gene Description	Ensembl Gene ID
BBS2	Bardet-Biedl Syndrome protein 2	Human: ENSG00000125124 Mouse: ENSMUSG00000031755 Rat: ENSRNOG00000019020 Zebrafish: ENSDARG00000032844
CBFB	Core-binding factor, beta subunit	Human: ENSG00000067955 Mouse: ENSMUSG00000031885 Rat: ENSRNOG00000014647 Chick: ENSGALG00000003169 Zebrafish: ENSDARG00000033632
Cetp	Cholesteryl ester transfer protein precursor	Human: ENSG00000087237 Zebrafish: ENSDARG00000030872 Chick: ENSGALG00000001234 *Takifugu*: SINFRUG00000127845
Dead eye (Nup93)	nuclear pore complex protein, 93kda	Human: ENSG00000102900 Mouse: ENSMUSG00000032939 Rat: ENSRNOG00000018564 Chick: ENSGALG00000003010 Zebrafish: ENSDARG00000003487
Gal	Beta-1,3-galactosyltransferase	Human: GENSCAN00000053607 Mouse: ENSMUSG00000051418 Rat: ENSRNOG00000015172
GNAO1	Guanine-nucleotide-binding protein G(o), alpha subunit 1	Human: ENSG00000087258 Mouse: ENSMUSG00000031748 Rat: ENSRNOG00000019482 Chick: ENSGALG00000003163
Herp	Homocysteine-responsive ER-resident ubiquitin-like domain member 1	Human: ENSG00000051108 Mouse: ENSMUSG00000031770 Rat: ENSRNOG00000018796 Zebrafish: ENSDARG00000032723

Lin10	unknown function	Human: ENSG00000125149 Mouse: ENSMUSG00000031889 Rat: ENSRNOG00000014668 Chick: ENSGALG00000003183 Zebrafish: ENSDARG00000033312
Metallothionein	Metal homeostasis	Human: ENSG00000125144* Mouse: ENSMUSG00000031762* Rat: ENSRNOG00000018756* Chick: ENSGALG00000014616 Zebrafish: ENSDARG00000033416 Fugu: SINFRUG00000137342 *Tetraodon*: GSTENG00016791001
SLC12A3	Thiazide sensitive Na-Cl co-transporter; solute carrier family 12, member 3	Human: ENSG00000070915 Mouse: ENSMUSG00000031766 Rat: ENSRNOG00000018607 Chick: ENSGALG00000002957 Zebrafish: ENSDARG00000013855 Fugu: SINFRUG00000127805

*A single metallothionein Ensembl gene ID was chosen to represent the metallothionein multi-gene locus in each of these species.

3.6.3 Gene Identification of Salmonid Specific Genes

The sequences for each BAC were subjected to BLAST searches of the Atlantic salmon and rainbow trout EST databases at The Institute for Genomics Research (TIGR). No further genes were identified.

3.6.4 Repetitive Elements

The percentage of repetitive DNA was estimated to be 30.4% in S0085O16 and 35.7% in S0188I22 (Table 3) by comparing the BAC sequences to various databases containing known repeats as well as by comparing the BAC sequences to themselves in order to identify internal repetitive elements not present in the databases.

3.6.4.1 Known Fish Repetitive Elements

RepeatMasker, a program designed to identify known interspersed repeats and low complexity DNA was used to identify known repetitive elements in the two BAC sequences. Three of the repeat databases available through RepeatMasker were used for this task: (1) Zebrafish, (2) Pufferfish, and (3) Fish Other than Zebrafish and Pufferfish. In this way, 13.8% of the S0085O16 and 13.6% of the S0188I22 sequences were identified as being similar to known repetitive sequences such as retroelements (LINEs, SINEs, and LTRs), transposable elements, simple sequence repeats, and low complexity DNA (Table 3). The repetitive DNA identified by this search was replaced by Xs in each of the S0085O16 and S0188I22 sequences.

Table 3 Percentage of BAC Sequences that Contain Repetitive DNA

Repeat Element	S0085O16	S0188I22
Retroelements	2.31%	4.93%
- SINEs	- 0.10%	- 0.04%
- LINEs	- 1.8%	- 4.56%
- LTRs	- 0.41%	- 0.33%
Transposons	9.08%	6.19%
Simple Sequence Repeats	1.8%	1.47%
Low Complexity DNA	0.48%	0.93%
Small RNA	0.07%	0.03%
Other	0.06%	0%
Salmonid Specific	13%	13%
Minisatellite DNA	3.8%	9.2%
Total Bases	222,799	205,993
Total Bases excluding Ns	219,298	203,433
Total Repetitive DNA	**30.4%**	**35.7%**

3.6.4.2 Salmonid Specific Repeats

The masked BAC sequences were then compared to a salmonid specific repeat database that is currently being developed in our lab. This search identified a further 13% of each of the BAC sequences as sharing similarity with sequences that have been determined to be repetitive in the genomes of various salmonid species.

3.6.4.3 Internal Repeats

Each BAC sequence was aligned to itself in order to identify internal repeats; percent identity plots (PipMaker plots, Schwartz et al., 2000) were used to analyze the results (Figure 15 and Figure 16). In doing so, several repetitive elements not initially picked up by the previous analyses were identified, and the location of these elements was compared between the BACs.

Figure 15 **S0085O16 versus S0085O16 Pipmaker Plot**

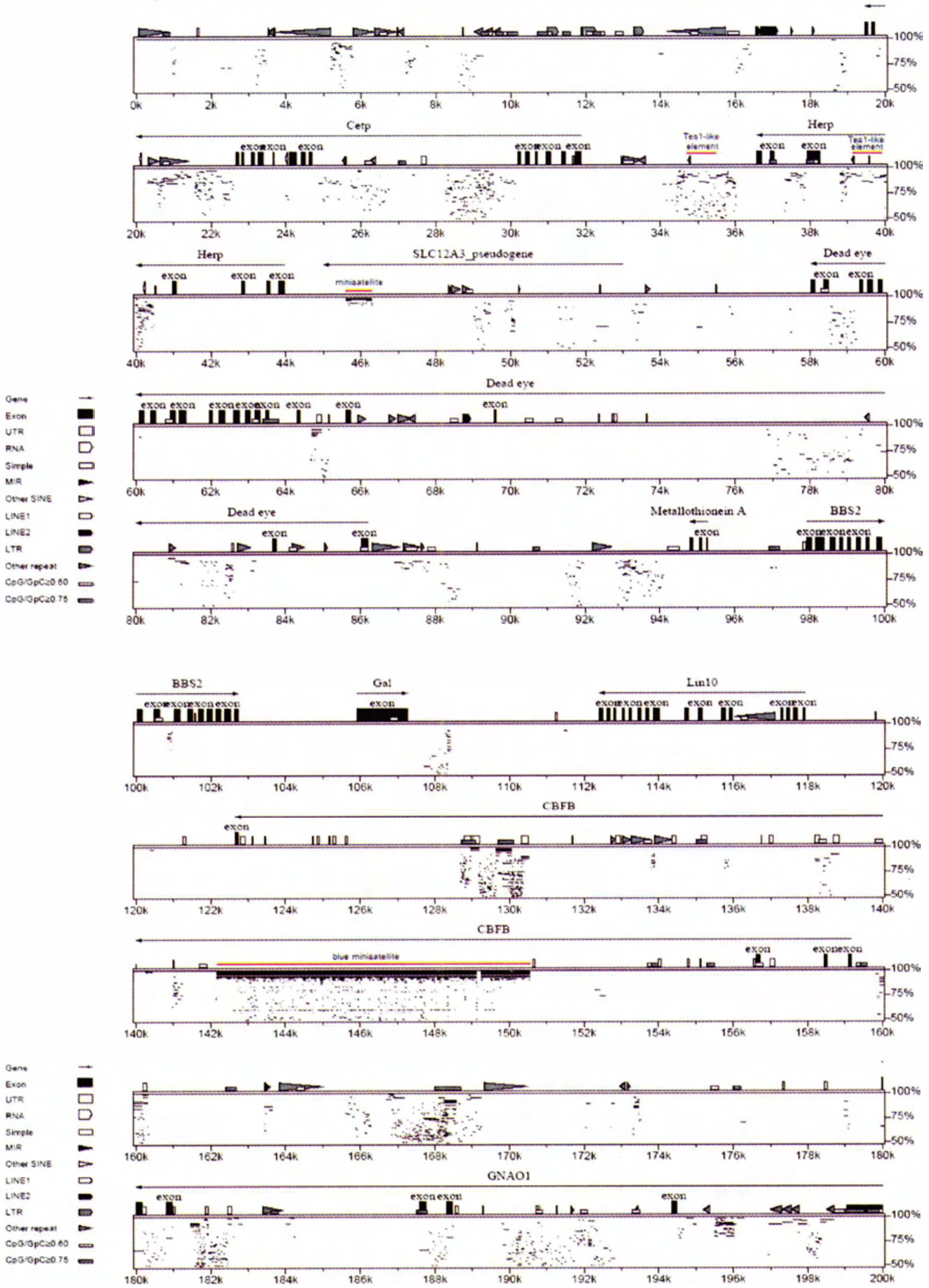


Figure 15 Continued.

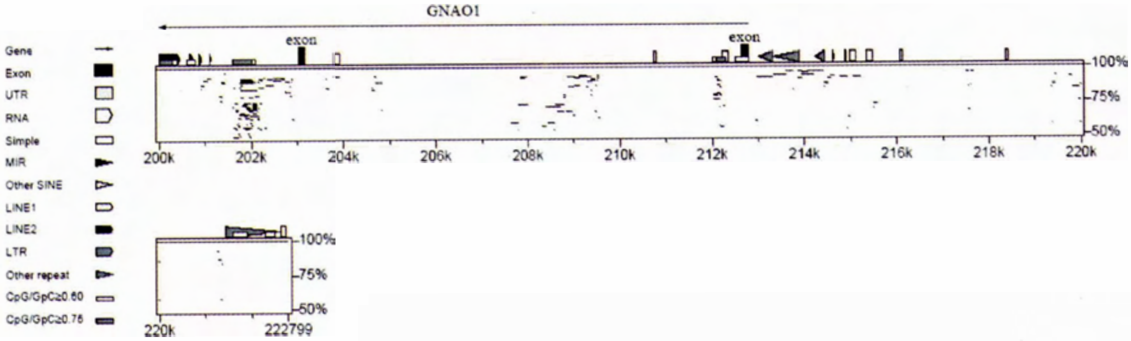


Figure 16 S0188I22 versus S0188I22 PipMaker Plot

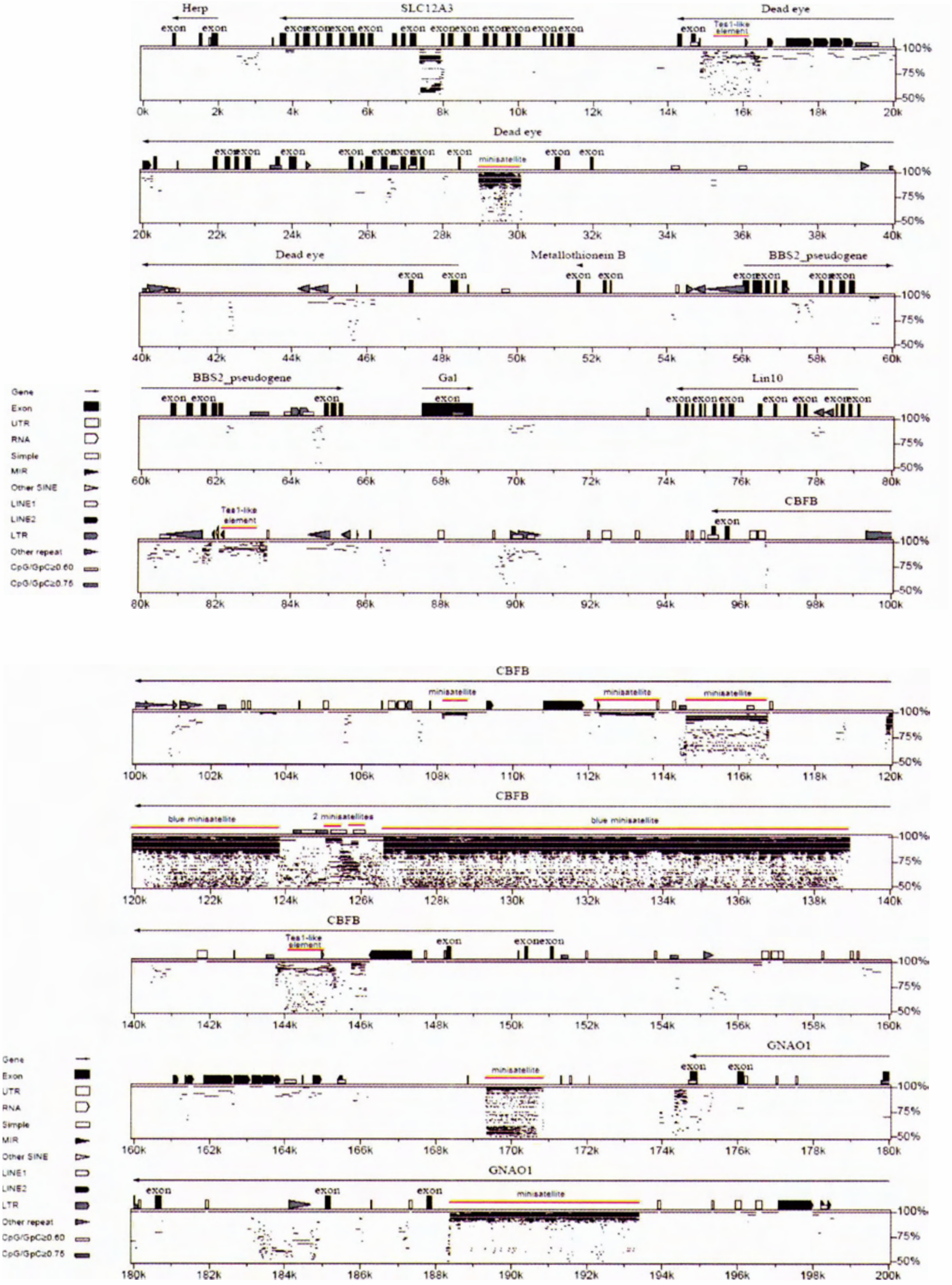
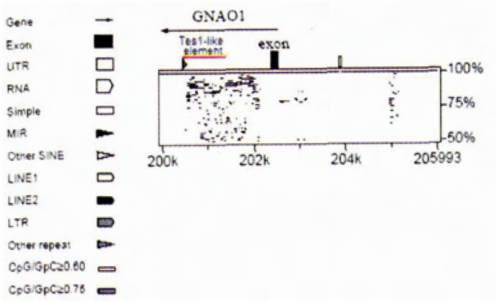


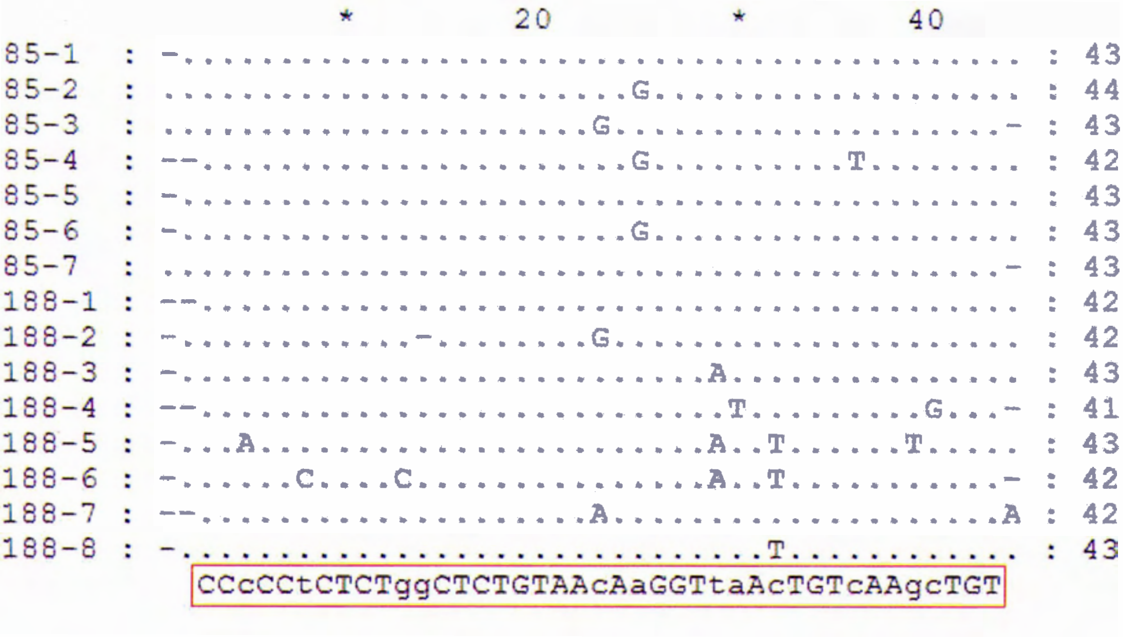
Figure 16 Continued



3.6.4.3.1 Unique Minisatellite Repeat

A large minisatellite repeat was identified in each of the two BAC sequences and has been tentatively named the blue minisatellite. It is composed of an imperfect repeating unit of 41-44 nucleotides and is quite G+C rich at 50%. In the S0085O16 sequence, the blue minisatellite spans 8kb (142 to150kb), or 3.6% of the total sequence. In the S0188I22 sequence, the minisatellite spans 120 to 124kb and 126.6 to 139kb, totalling 16.4kb of repetitive DNA, or 7.0% of the complete BAC sequence. The interruption in the blue minisatellite in this sequence is due to the presence of an even more G+C rich region (54%), partially composed of two different, G+C rich minisatellites (125.1 to 125.5kb and 125.5 to 126kb). The blue minisatellite shares no sequence similarity with any repetitive elements in the salmonid specific database. In both sequences, the blue minisatellite lies within the intron of a gene (CBFB). A random sampling of repeating units within the blue minisatellite from both BAC sequences was selected and subjected to a multiple sequence alignment to determine the consensus sequence (Figure 17).

Figure 17 Multiple Sequence Alignment and Consensus Sequence of the Blue Minisatellite



```
                 *        20         *        40
85-1   : -........................................... : 43
85-2   : ........................G................... : 44
85-3   : ......................G....................- : 43
85-4   : --......................G..........T........ : 42
85-5   : -........................................... : 43
85-6   : -.......................G................... : 43
85-7   : ...........................................- : 43
188-1  : --.......................................... : 42
188-2  : -............-........G..................... : 42
188-3  : -...........................A............... : 43
188-4  : --...........................T.........G...- : 41
188-5  : -...A.......................A..T......T..... : 43
188-6  : -......C....C...............A..T...........- : 42
188-7  : --....................A....................A : 42
188-8  : -..............................T............ : 43
           CCcCCtCTCTggCTCTGTAAcAaGGTtaAcTGTcAAgcTGT
```

3.6.4.3.2 Other Internal Repeating Units

Several more minisatellites were identified within the BAC sequences, two of which are present in the salmonid specific database, the rest of which are not (Table 4). Of note, all of the minisatellites are unique to their respective BAC sequence, and therefore likely arose after the duplication event.

One of the two minisatellites found in the salmonid repeat database is quite large and spans 8kb in the S0188I22 sequence (188.5 to 193.5kb); it is comprised of a 54 nucleotide repeating unit. The other minisatellite that the salmonid database recognized is only 1kb in length, is comprised of a 22 nucleotide repeating unit, and is only found in the S0188I22 sequence (29 to 30kb).

The remaining minisatellites that were identified were not found in the salmonid specific repeat database. In S0085O16, a 29 nucleotide long repeating unit spanning 45.6 to 46.1kb was found. In S0188I22, 19 and 42 nucleotide repeats were found from 7.3 to 7.9kb and 108.2-108.8kb plus 114.5-116.7kb, respectively. Additionally, a 39 nucleotide long repeating unit minisatellite was found in the S0188I22 sequence from 169.3-170.4kb that was very G+C rich (58%). At the corresponding region in S0085O16 a sequence gap exists (currently represented by 3000 Ns), followed by the tail end of what appears to be a G+C rich region. This 33 nucleotide minisatellite might be found within the unknown sequence. With the possible exception of the 33 nucleotide repeating unit minisatellite, each of the other four minisatellites is both unique within and between these two BAC sequences.

The identification of these additional minisatellites results in 0.2% more of the S0085O16 and 2.2% more of the S0188I22 sequences being classified as repetitive.

Table 4 Minisatellite Repetitive Elements

Location		Nucleotides per Repeating Unit	Consensus Sequence	Salmonid Specific Repeat Database
S0085O16	S0188I22			
45.6 to 46.1kb	---	29	CCCAATATGTTCCCCCTGTT GATGATGCA	No
---	7.4 to 7.9kb	19	TGGTCGTGATTCCATGTCC	No
---	28 to 29kb	22	TAGAAATGTGTTCATTTGCC AG	Yes
---	108.2 to 108.8kb and 114.5 to 116.7kb	42	TAATGGTAACTACAGGACAG AATACCCACACACTACAGTC TG	No
---	125.1 to 125.5kb	30	CCCGCCAGTCAACAGTCAT CGTCAGAGCTG	No
---	125.5 to 126	27	CCAGACTGCGCTGAACTGC CGGAGTGG	No
---	169.3 to 170.4kb	33	TAATAGATAGGTGTCAGGTG CCAGAGGACAGTG	No
---	188.5 to 193.5kb	54	TGTGGTAGAGAGGATCATCA CACCCAAGACTAACCAACAG TCAGCACCCAGTAA	Yes

3.6.4.3.3 Additional Transposable Elements

Tes1-like elements, which are Tc1-like elements first identified in the Pacific hagfish (*Eptatretus stouti*) (Heierhorst et al., 1992), were also found in non-conserved locations in the Atlantic salmon BAC sequences. Two are present on the minus strand in S0085O16 at 35kb and 39kb, and four are found in the S0188I22 sequence at 16kb (plus), 82.5kb (minus), 145kb (plus), and 201kb (minus) (Figure 15 and Figure 16) . Each Tes1-like element is about 800 bp in length in the Atlantic salmon sequences.

3.6.4.4 Comparative Locations of Repetitive Sequences

The relative location of the various repetitive elements between the BAC sequences was compared. The majority of the known repetitive elements within the sequences are transposable elements and LINES, where S0188I22 has more LINES than S0085O16, and more repetitive DNA overall even though the total sequence length is shorter. In general, very few of the repeat locations are conserved between the homeologous loci, indicating a great deal of activity since the duplication event. The exception to this statement, where a repetitive element was presumably in place before the genome duplication, is the blue minisatellite, although it has either expanded or contracted in one of the two BAC sequences since divergence.

3.6.5 G+C Content

The average G+C content across the length of the BAC sequences is 42.52% and 43.06% for S0085O16 and S0188I22, respectively. Examining the G+C content of the exonic portions of the genes in each sequence shows a significantly higher value than the overall average, varying from 50% (Lin10 duplicates) to 56% (dead eye duplicates) (Table 5). In contrast, the G+C content of the introns (37 to 42.6%) is quite a bit lower than the corresponding value in the exons (Table 5). Furthermore, the G+C content between each pair of duplicate genes is relatively constant.

Table 5 **Percent G+C Content of Exons and Introns**

Gene	Exons G+C (%)		Introns G+C (%)	
	S0085O16	**S0188I22**	**S0085O16**	**S0188I22**
BBS2	51.5	50.4 (pseudogene)	38.6	43.1 (pseudogene)
Dead eye	56	55.6	41	41.9
Galactosyltransferase	51.4	51.5	n/a[1]	n/a[1]
Herp	56.2	n/a[2]	39.1	n/a[2]
Lin10	50	50.3	42.6	41.1
Metallothionein	52.7	51.4	37	39.9
SLC12A3	n/a[3]	51.9	n/a[3]	40.3
Average[4]	**53.2**	**52.7**	**40.7**	**41.6**

[1] single exon gene, therefore no introns

[2] incomplete gene

[3] not possible to determine G+C content of pseudogene

[4] excluding values for BBS2 pseudogene; calculated over total length of exon or intron sequence data

3.6.6 Conservation of Synteny

The BAC sequences were compared to one another to determine the extent of collinearity – the conservation of gene order and content between the two genomic segments. The gene order was found to be identical between the two loci and is as follows: GNAO1, CBFB, Lin10, beta-1,3-galactosyltransferase, BBS2, Metallothionein A or B, dead eye, SLC12A3, and HERP (Figure 18). The gene for CETP is found only in the S0085O16 MetA sequence and falls at the 3' end of the HERP gene. Because the S0188I22 sequence does not extend that far in the 3' direction (it ends within the third intron of the Herp gene), there is no comparable CETP sequence for S0188I22. The transcriptional orientation has remained conserved between each pair of duplicates.

In comparison to the genomes of other fully sequenced organisms, synteny has been conserved to a remarkable degree with respect to these ten genes (Figure 19). In the human genome, all ten of these genes are found on the long arm of chromosome 16, although the gene order is different from that seen in Atlantic salmon. In human, the deadeye homolog is called Nup93, and is part of a segment of six genes that have retained their collinearity since the common ancestor with Atlantic salmon; BBS2, the metallothionein multi-gene locus, Nup93, SLC12A3, Herp, and Cetp all lie in the region of 55.5Mb on chromosome 16. Interestingly, in human, the metallothionein locus has undergone multiple tandem duplications, resulting in twelve paralogous copies of the metallothionein gene at this locus. Also found in this region of chromosome 16 is the GNAO1 gene, which is separated from BBS2 by the presence of three intervening genes (AMFR, CPSF5, and NP_060703). The remaining three genes – CBFB, Lin10 and beta-1,3-galactosyltransferase - have been involved in a chromosomal rearrangement relative to the salmon gene order, and now sit in the region of 65.5Mb on chromosome 16; again, the gene order and directionality has been conserved.

Figure 18 Atlantic salmon gene order

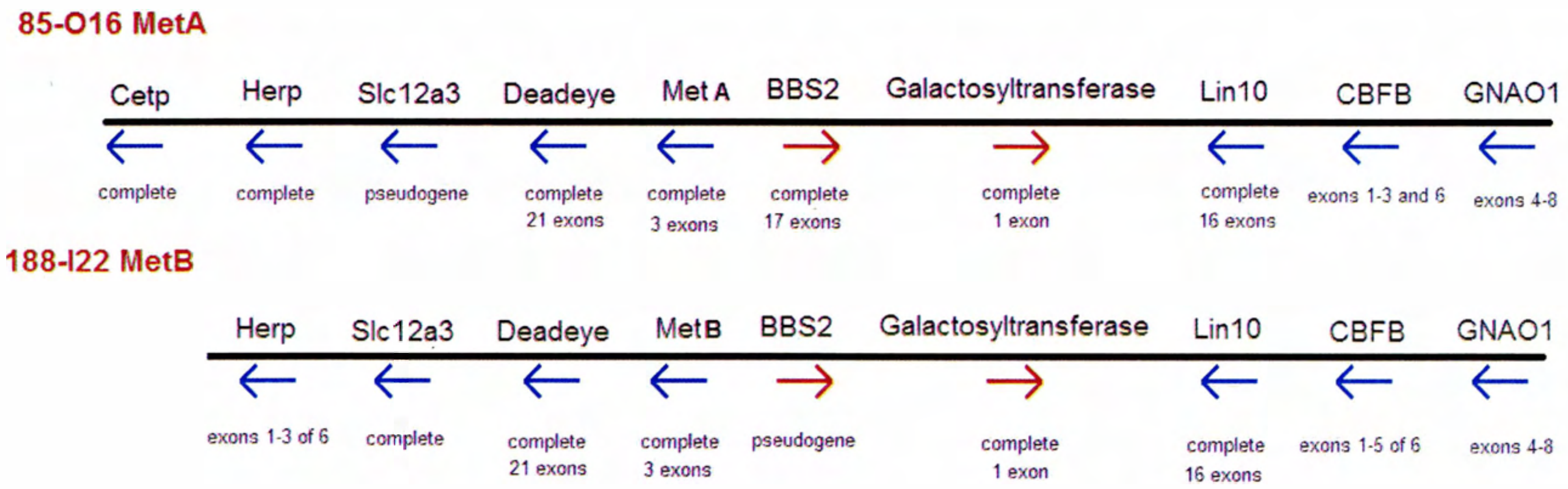
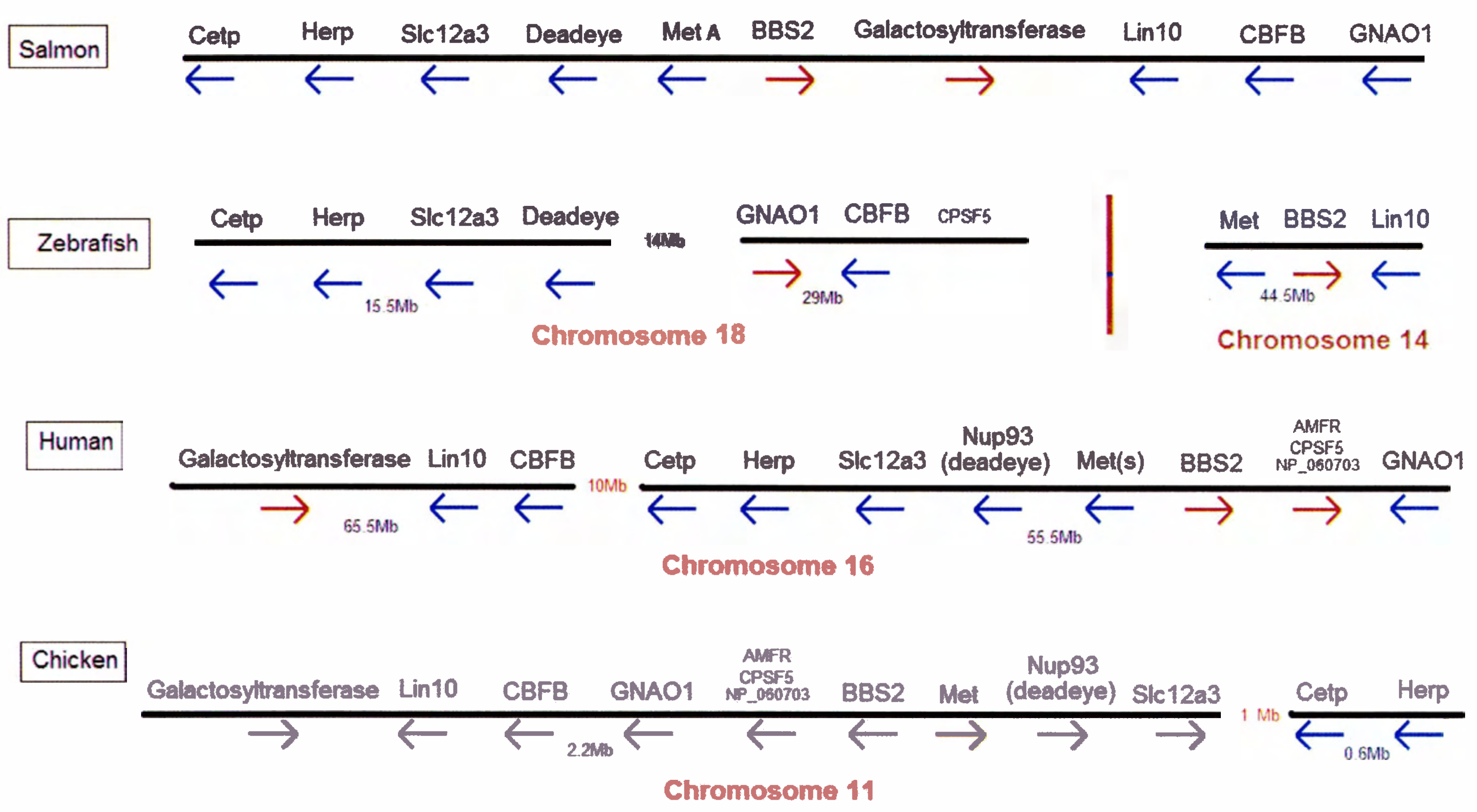


Figure 19 Conservation of synteny



Other mammalian genomes have very similar gene structure with respect to the ten genes identified in the Atlantic salmon BAC sequences. In mouse, these ten genes are found on chromosome 8 and exhibit the identical arrangement as that seen in human. In rat, on the other hand, an inversion has taken place with respect to human and mouse, leading to a reversal in direction for the block of genes containing Cetp, Herp, SLC12A3, Nup93, the metallothionein gene cluster, BBS2, the three intervening genes (AMFR, CPSF5, and NP_060703) and GNAO1, which all lie on chromosome 19. CBFB, Lin10, and the galactosyltransferase homologs, which are located at 34Mb on chromosome 19, have retained their collinearity. Finally, in the dog genome, the same large block of genes has also experienced an inversion on chromosome 2, and furthermore, a translocation event has resulted in CBFB, Lin10, and galactosyltransferase ending up on chromosome 5.

In chick, these ten genes are found on chromosome 11. Again, segments containing multiple genes that have conserved order and transcriptional direction relative to Atlantic salmon can be seen. BBS2, metallothionein, Nup93, and SLC12A3 lie adjacent to one another, and are separated from galactosyltransferase, Lin10, CBFB and GNAO1 by the presence of the same three genes that are present in human (AMFR, CPSF5, and NP_050703). The genes coding for CETP and Herp are located 1.6Mb down chromosome 11 from the other eight genes.

In zebrafish, both chromosomal rearrangement and translocation events have taken place relative to Atlantic salmon. Six of the genes are found in two clustered segments on chromosome 18 (Cetp, Herp, SLC12A3 and dead eye at 15.5Mb; CBFB and GNAO1 at 19Mb), and three other genes are found on chromosome 14 (Lin10, BBS2, and metallothionein). A gene coding for beta-1,3-galactosyltransferase could not be located anywhere in the vicinity of any of these three chromosomal regions. Once again, the directionality of the genes within these regions is conserved relative to salmon with the exception of CBFB and GNAO1 where chromosomal rearrangements have presumably resulted in the orientation of one of the genes being reversed.

Examining other fully sequenced organisms, such as *Takifugu* and *Tetraodon*, the two sequenced pufferfish genomes, is less informative in terms of examining the degree of conservation of synteny because many scaffolds have yet to be anchored to chromosomes. In the current *Takifugu* genome build (release 26.2c.1), for instance, homologs of SLC12A3, Herp, and Cetp lie adjacent to one another on a single scaffold, but the location of this scaffold within the genome is unknown. Similarly, dead eye is beside the metallothionein locus, which has two metallothionein genes, presumably the result of a tandem duplication, on a different, unanchored scaffold. Homologs for CBFB, Lin10, BBS2, Herp, and beta-1,3-galactosyltransferase were not found, possibly because they have not been annotated, or alternatively because they do not exist. In *Tetraodon*, there are two metallothionein genes adjacent to one another that are found on chromosome 13 one gene away from the SLC12A3 homolog; the intervening gene, a novel gene prediction, was not been identified in any of the orthologous chromosomal regions that were examined. Additionally, a Cetp homolog was found on an unanchored scaffold, and interestingly, this scaffold also contains an SLC12A3 homolog. Finally, GNAO1 is found on chromosome 5, separated by two genes from the homolog of NP_050703, one of the intervening genes found in the tetrapod orthologous chromosomal regions. Homologs of CBFB, Lin10, BBS2, and Herp, beta-1,3-galactosyltransferase were not identified in the current *Tetraodon* release (26.1a.1). With the exception of Herp, none of these genes have been annotated in *Takifugu* either, suggesting that they might simply not exist in the pufferfish lineage.

3.7 Gene Annotation

Each of the ten genes identified in the previous section was further annotated on an individual basis in both BAC sequences. With the exception of metallothionein, none of these genes has been previously characterized in salmonids and therefore no full-length coding sequence exists for any of them. Therefore, coding sequences were predicted based on sequence similarity to known exons from human, zebrafish, and chick, as well as by incorporating the intron/exon boundaries predicted by Genscan and Genemark. Additionally, the salmonid ESTs found in the database at TIGR were used to discern gene structure when possible. The ten genes that have been identified in the homeologous BAC sequences may be grouped into three categories: complete genes, incomplete duplicate pairs due to missing sequence data, and complete gene/pseudogene duplicate pairs. The exons in all of the multiple sequence alignments in this section are distinguished by alternating underlines.

3.7.1 Complete Genes

In salmonids, the complete coding sequence is only known for metallothionein with respect to this group of ten genes, and so the genes that have been classified as complete are based on predictions. The annotation of a complete gene, in this thesis, is based on the presence of consensus splice sites on either side of segments of DNA that share high sequence similarity with exons from orthologous genes in other organisms, that when translated, result in complete amino acid sequences (i.e. no stop codons, missense mutations, etc.). Along with metallothionein, complete gene sequences were found for Lin10, dead eye, and beta-1,3-galactosyltransferase in both BAC sequences.

3.7.1.1 Metallothionein

The metallothionein loci have been well studied in salmonids, as well as throughout the vertebrate lineage. Both the MetA and MetB genes in the Atlantic salmon BAC sequences display the expected tripartite gene structure, coding for proteins of about 60 amino acids in length with a high (30%) cysteine content. Atlantic salmon MetA is 61 amino acids long (the extra residue being coded for in exon 2 relative to MetB) and has a relatively short second intron compared to the MetB locus, which is 461 bp longer than MetA (Figure 20).

The BAC sequence data provide, for the first time, the opportunity to examine the regulatory region of MetA and MetB in Atlantic salmon. Five metal responsive elements (MREa-e) were identified in the 5' flanking region of each of the genes (5' TGGRCNC), as well as a TATA box. Four of the MREs (a-d) lie in conserved areas between the two regulatory regions, whereas MREe, is located further upstream in the MetB sequence than the MetA sequence (Figure 21). In MetA, the MREs are located at -140, -195, -673, -781 and -831. In MetB, the MREs are located at -135, -172, -646, -751, and -1467.

Figure 20 Structure of the Atlantic salmon metallothionein A and B genes

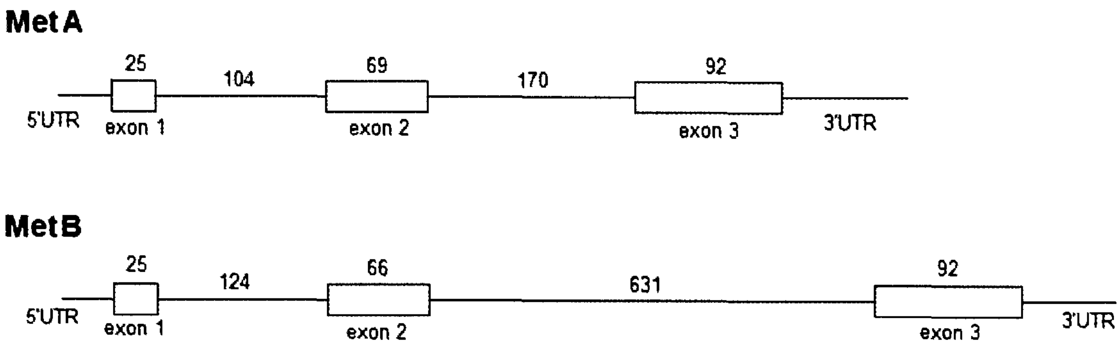
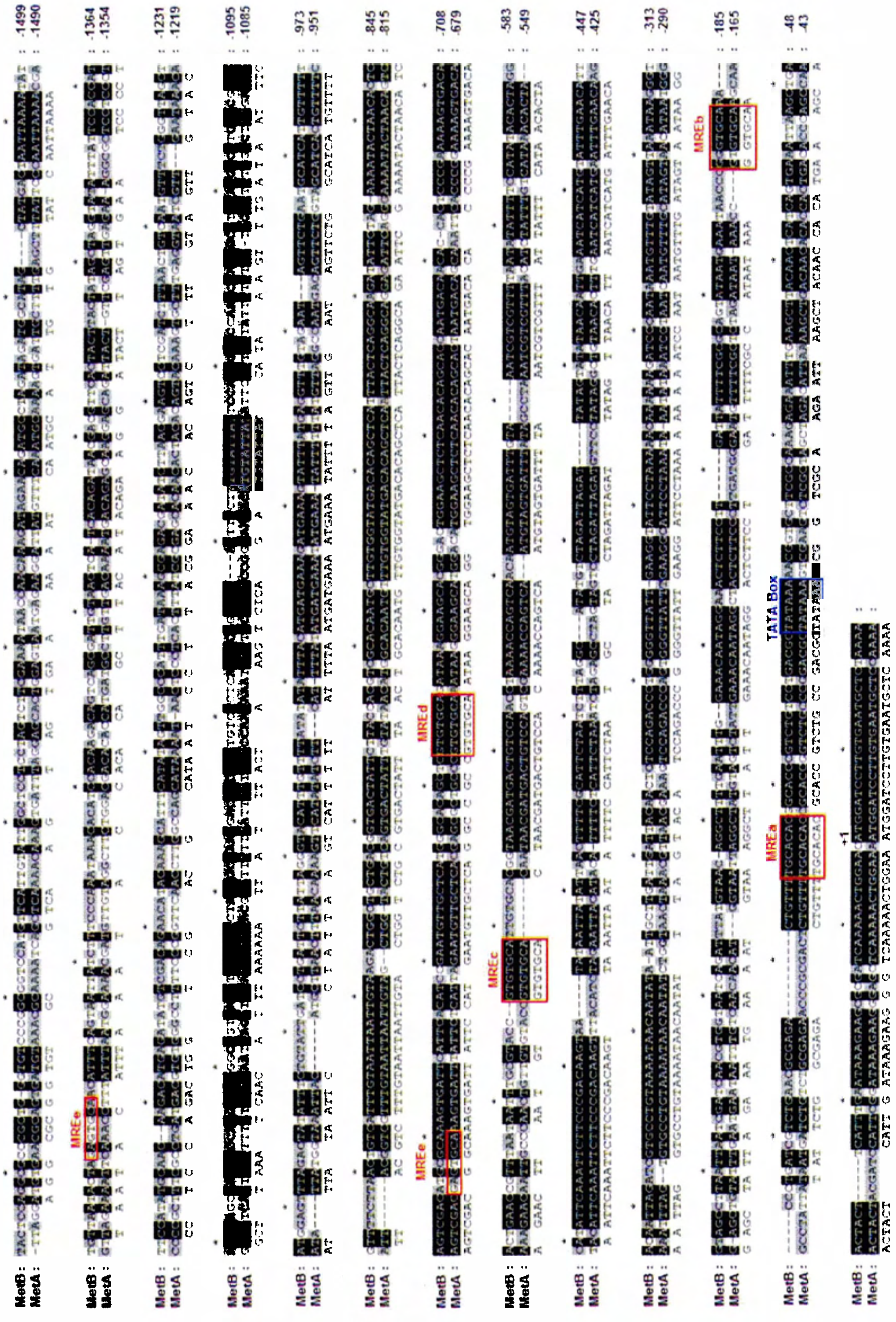


Figure 21 Alignment of MetA and MetB regulatory regions



3.7.1.2 Lin10

The Atlantic salmon duplicate Lin10 loci are situated on the minus strand in each BAC sequence and are composed of 16 exons of identical length. In S0085O16, the gene spans 5.4 kb, from 112.5 through to 117.9 kb while in S0188I22 the Lin10 gene is slightly shorter, spanning 4.7 kb from 74.3 to 79 kb. The average intron size is quite similar between the paralogs except for intron 4, which is 1311 in S0085O16, and only 780 bp in S0188I22. The size difference may be accounted for by the presence of a longer transposable element in the S0085O16 intron compared to the S0188I22 intron.

In both human and zebrafish the Lin10 homologs are also composed of 16 exons, but the genomic region covered by each gene is quite a bit longer due to larger intron sizes. In human, the gene spans 38.5kb on chromosome 16 with an average intron size of 2.4kb, and in zebrafish the gene spans 13.3kb on chromosome 14 with an average intron size of 800bp. In chicken, the Lin10 homolog exists on chromosome 11 is annotated as a novel gene prediction made by Ensembl, where the first exon from human and zebrafish is not annotated. Once again, however, the gene covers quite a long genomic distance at 27.7kb, with an average intron size of 1.7kb.

The coding sequences for the Atlantic salmon Lin10 duplicates are 1278 bp in length and share 95% sequence similarity at the nucleotide level (61 nucleotide changes overall). The protein sequences are 426 amino acids long and share 96% sequence similarity (15 amino acid changes). An amino acid sequence alignment between the Lin10 homologs of Atlantic salmon, zebrafish, human, mouse, chick, and rat shows that this protein sequence is highly conserved (Figure 22). The Atlantic salmon paralogs share 94% amino acid sequence similarity with zebrafish, and 86 to 88% similarity with the three tetrapod sequences.

Figure 22 Lin10 amino acid sequence alignment.

```
Mouse            MLDLEVVPERSLGNEQWEFTLGMPLAQAVAILQKHCRIIRNVQVLYSEQSPLSHDLILNL 60
Rat              MLDLEVVPERSLGNEQWEFTLGMPLAQAVAILQKHCRIIKNVQVLYSEQSPLSHDLILNL 60
Human            MLDLEVVPERSLGNEQWEFTLGMPLAQAVAILQKHCRIIKNVQVLYSEQSPLSHDLILNL 60
Chick            ------------------FFAGMPLAQAVAILQKHCRIIKNVQVLYSEQSPLSHDLILNL 42
S0085016         MLDLEVVPERSLGHEQWEFALGMPLAQAISILQKHCRIIKNVQVLYSEQTPLSHDLILNL 60
S0188I22         MLDLEVVPERSLGNEQWEFALGMPLAQAISILQKHCRIIKNVQVLYSEQTPLSHDLILNL 60
Zebrafish        MLDLEVVPERSLGNEQWEFALGMPLAQAISILQKHCRIIKNVQVLYSEQMPLSHDLILNL 60
                                   *  *******::*********.********* **********

Mouse            TQDGITLLFDAFNQRLKVIEVCELTKVKLKYCGVHFNSQAIAPTIEQIDQSFGATHPGVY 120
Rat              TQDGIKLLFDAFNQRLKVIEVYDLTKVKLKYCGVHFNSQAIAPTIEQIDQSFGATHPGVY 120
Human            TQDGIKLMFDAFNQRLKVIEVCDLTKVKLKYCGVHFNSQAIAPTIEQIDQSFGATHPGVY 120
Chick            TQDGIKLLFDAFNQRLKVIEVYDLTKVKLKYCGVHFNSQAIAPTIEQIDQSFGATHPGVY 102
S0085016         TQDGIKLLFDACNQRLKVIEVYDLSKVKLKYCGVHFNTQAIAPTIEQIDQSFGATHPGVY 120
S0188I22         TQDGIKLLFDACNQRLKVIEVYDLSKVKLKYCGVHFNTQAIAPTIEQIDQSFGATHPGVY 120
Zebrafish        TQDGIKLLFDACNQRLKVIEVYDLTKVKLKYCGVHFNSQAIAPTIEQIDQSFGATHPGVY 120
                 *****.*:*** ********* :*:************:**********************

Mouse            NSTEQLFHLNFRGLSFSFQLDSWTEAPKYEPNFAHGLASLQIPHGATVKRMYIYSGNSLQ 180
Rat              NSAEQLFHLNFRGLSFSFQLDSWTEAPKYEPNFAHGLASLQIPHGATVKRMYIYSGNSLQ 180
Human            NSAEQLFHLNFRGLSFSFQLDSWTEAPKYEPNFAHGLASLQIPHGATVKRMYIYSGNSLQ 180
Chick            NSAEQLFHLNFRGLSFSFQLDSWTETPKYEPNFAHGLASLQIPHGATVKRMYIYNGNSLQ 162
S0085016         NAAEQLFHLNFRGLSFSFQLDSWNEAPKYEPNFAHGLASLQIPHGATVKRMYIYTGNNLQ 180
S0188I22         NAAEQLFHLNFRGLSFSFQLDSWNEAPKYEPNFAHGLASLQIPHGATVKRMYIYAGNNLQ 180
Zebrafish        NAAEQLFHLNFRGLSFSFQLDSWSEAPKYEPNFAHGLASLQIPHGATVKRMYIYSGNNLQ 180
                 *::********************.*:**************************** **.**

Mouse            DTKAPVMPLSCFLGNVYAESVDVLRDGTGPSGLRLRLLAAGCGPGVLADAKMRVFERAVY 240
Rat              DTKAPMMPLSCFLGNVYAESVDVLRDGTGPSGLRLRLLAAGCGPGVLADAKMRVFERAVY 240
Human            DTKAPMMPLSCFLGNVYAESVDVLRDGTGPAGLRLRLLAAGCGPGLLADAKMRVFERSVY 240
Chick            DTKAPLMPLSCFLGNVYAENVDVLRDGTGPSGLRLRLLTAGCGPGVLADAKMRVFERCVY 222
S0085016         DTKAPVMPLACFLGNVYAECVDVLKNGAGPLGLRLRILTAGCGPGVMADAKVRAVERNIY 240
S0188I22         DTKAPVMPLACFLGNVYAECVDVLRDGVGPLGLRLRLLTAGCGPGVMADAKVRAVERNIY 240
Zebrafish        ETKAPAMPLACFLGNVYAECVEVLRDGAGPLGLKLRLLTAGCGPGVLADTKVRAVERSIY 240
                 :**** ***:********* *:**::*.** **:**:*:******::**:*:*..** :*

Mouse            FGDSCQDVLSMLGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDILFDAN 300
Rat              FGDSCQDVLSMLGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDILFDAN 300
Human            FGDSCQDVLSMLGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDILFDAN 300
Chick            FGDSCQDVLSTLGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDILFDAN 282
S0085016         FGDSCQDVLSALGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDILFDST 300
S0188I22         FGDSCQDVLSALGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYFTLGVDLLFDSS 300
Zebrafish        FGDSCQDVLSALGSPHKVFYKSEDKMKIHSPSPHKQVPSKCNDYFFNYYILGVDILFDST 300
                 ********** *************************************: ****:***:.

Mouse            THKVKKFVLHTNYPGHYNFNIYHRCEFKIPLAIKKENAGGQTEIC--TTYSKWDSIQELL 358
Rat              THKVKKFVLHTNYPGHYNFNIYHRCEFKIPLAIKKENAGGQTEIC--TTYSKWDSIQELL 358
Human            THKVKKFVLHTNYPGHYNFNIYHRCEFKIPLAIKKENADGQTETC--TTYSKWDNIQELL 358
Chick            THKVKKFVLHTNYPGHYNFNIYHRCEFKIPLVIKRDSADSQTETC--TTYSKWDSIQDLL 340
S0085016         THLVKKFVLHTNFPGHYNFNIYHRCDFKIPLVIKKEAADAQWEDCILTTYSKWDQIQELL 360
S0188I22         THLVKKFVLHTNFPGHYNFNIYHRCDFKIPLVIKKEGAAAQREDCTLTTYSKWDQIQELL 360
Zebrafish        THLVKKFVLHTNFPGHYNFNIYHRCDFKIPLIIKKDGADAHSEDCILTTYSKWDQIQELL 360
                 ** *********:************:***** **:: * .: * *  *******.**:**

Mouse            GHPVEKPVVLHRSSSPNNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGPPRPGAHLRT 418
Rat              GHPVEKPVVLHRSSSPNNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGPPRPGAHLRT 418
Human            GHPVEKPVVLHRSSSPNNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGPPRPGSHLRT 418
Chick            GHPVEKPVVLHRSSSPNNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGPTRPSSQLRT 400
S0085016         GHPMEKPVVLHRSSSANNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGAPRPNSRARA 420
S0188I22         GHPMEKPVVLHRSSSANNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGAPRPSSRARA 420
Zebrafish        GHPMEKPVVLHRSSSANNTNPFGSTFCFGLQRMIFEVMQNNHIASVTLYGAPRPSSLAR- 419
                 ***:***********.**********************************..**.:  * 

Mouse            AELP- 422
Rat              AELP- 422
Human            AELP- 422
Chick            SDLP- 404
S0085016         EASAH 425
S0188I22         EAIGH 425
Zebrafish        -----
```

3.7.1.3 Dead eye

The Atlantic salmon dead eye duplicates are composed of 20 exons of identical lengths, located on the minus strand of both BAC sequences. Due to the presence of several large introns, the genomic region covered by the dead eye duplicates is larger than the Atlantic salmon Lin10 paralogs; in S0085O16, dead eye spans 28kb (58-86kb) while in the S0188I22 sequence the gene covers 34kb (14-48kb), with the second intron being 14 and 15kb, respectively. Overall, the average intron length is comparable between the paralogs with the exception of intron 18, where the insertion of several LINEs in the S0188I22 sequence has increased the intron length to 7kb compared to 1kb in S0085O16.

Relative to the known human and zebrafish homologs, which are composed of 21 exons each, the Atlantic salmon genes are one exon shorter because human and zebrafish exons 14 and 15 have condensed into a single exon the Atlantic salmon duplicates. In chick, the dead eye homolog is a predicted gene, where exons 1 and 2 from human and zebrafish are not annotated. Similarly, in rat, exons 1 to 4 from human and zebrafish are not annotated.

The salmon dead eye coding sequences are composed of 2463 nucleotides and are 94% similar (137 differences). The proteins are 820 amino acids in length and share 96% sequence similarity (27 changes). An amino acid alignment of the Atlantic salmon dead eye paralogs with the orthologs from human, mouse, rat, chick, zebrafish, and *Takifugu* reveals these proteins to be highly conserved; the Atlantic salmon sequence are 90 to 93% similar to zebrafish and *Takifugu*, and 80 to 82% similar to the tetrapod sequences (Figure 23).

Figure 23 Dead eye amino acid multiple sequence alignment

```
S0188I22            MEAEGFGELLQQAEQLAAETEAVSELPHVERNLQEIQQAGERLRSRTLTRTSQDAADVKA 60
S0085O16            MEAEGFGELLQQAEQLAAETEAVSELPHVERNLQEIQQAGERLRSRTLTRTSQDAADVKA 60
Takifugu            MDAEGFGELLQQAEQLAAETEAVSELPHVERNLQEIQQAGERLRSRTLTRTSQDAADVKA 60
Zebrafish           MDTEGFGELLQQAEQLAAETEAVSELPHVERNLQEIQQAGERLRSRTLTRTSQDTADVKA 60
mouse               MDTEGFGELLQQAEQLAAETEGISELPHVERNLQEIQQAGERLRSRTLTRTSQETADVKA 60
rat                 ------------------------------------------------------------
human               MDTEGFGELLQQAEQLAAETEGISELPHVERNLQEIQQAGERLRSRTLTRTSQETADVKA 60
chick               ----------------------------------------------------------

S0188I22            SILLGSRGLDIFHISQRLESLSAATTFEPLEPVKDTDIQGFLKNERDNALLSAIEESRRR 120
S0085O16            SILLGSRGLDIFHISQRLESLSAATTFEPLEPVKDTDIQGFLKNERDNALLSAIEESRRR 120
Takifugu            SILLGSHGLDIFHISQRLESLSAATTFEPLEPVKDTDIQGFLKNERDNALLSAIEESRRR 120
Zebrafish           SILLGSRGLDIFHISQRLESLSAATTFEPLEPVKDTDIQGFLKNERDNALLSAIEESRRR 120
mouse               SVLLGSRGLDISHISQRLESLSAATTFEPLEPVKDTDIQGFLKNEKDNALLSAIEESRKR 120
rat                 ------------------------------------------------------------
human               SVLLGSRGLDISHISQRLESLSAATTFEPLEPVKDTDIQGFLKNEKDNALLSAIEESRKR 120
chick               -------------------------------------LQGFLKNEKDNALLSAIEESRKR 23

S0188I22            TFLLAEEYHRESMLVQWEQVKQRVLHTLLGAGEDALDFSQDIEPS-FVSDAAVPGRSALD 179
S0085O16            TFLLAEEYHRESMLVQWEQVKQRVLHTLLGAGEDALDFSQDVEPS-FVSDAGTPGRSALD 179
Takifugu            TFLLAEEYHRESMLVQWEQVKQRVLHTLLGAGEDALDLTQDLEVQNFVSAVTAPGRSALD 180
Zebrafish           TFLLAEEYHRDSMLVQWEQVKQRVLHTLLGAGEDALDFSQEVEPS-FVSEVGVPGRSALD 179
mouse               TFGMAEEYHRESMLVEWEQVKQRILHTLLASGEDALDFTQESEPS-YIGDVNPPGRSSLD 179
rat                 ------------------------------------------QPS-YVSDVSPPGRSSLD 17
human               TFGMAEEYHRESMLVEWEQVKQRILHTLLASGEDALDFTQESEPS-YISDVGPPGRSSLD 179
chick               TFAMAEEYHRESMLVEWEQVKQRILHTLLASGEDALDFTQESEPS-YVSESGPPGRSSLD 82
                                                              : . ::.    ****.**

S0188I22            SVEVAYGRQIYIFNEKIVNGHVQPNLGDLCASVADNLDDKNVSDMWLMVKQMTDVLLVPA 239
S0085O16            SVEVAYGRQIYVFNEKIVNGHVQPNLGDLCASVADSLDDKNVSDMWLMVKQMTDVLLVPA 239
Takifugu            SVEVAYGRQIYIFNEKIVNGHIQPNLGDLCASVAESLEDKNVSDMWLMVKQMTDVLLVPA 240
Zebrafish           SVEVAYSRQIYVFNEKIVNGHLQPNLGDLCASVAESLDDKNVSEMWLMVKQMTDVLLVPA 239
mouse               SIEMAYARQIYIYNEKIVSGHLQPNLVDLCASVAE-LDDKSISDMWAMVKQMTDVVLTPA 238
rat                 SIEMAYARQIYIYNEKIVSGHLQPNLVDLCASVAE-LDDKSISDMWAMVKQMTDVVLTPA 76
human               NIEMAYARQIYIYNEKIVNGHLQPNLVDLCASVAE-LDDKSISDMWTMVKQMTDVLLTPA 238
chick               SVEMAYARQIYIYNEKIVNRHLQPNLVDLCAAVAE-LDDKNISEMWAMVKQMTSVRLVSA 141
                    .:*:**.****::*****. *:**** ****:**: *:**.:*:** ******.* *..*

S0188I22            KDTLKSRTALDMQVAFVTQALQFLQNSYKNYTMVTVFGNLHQAQLGGVPGTYQLVRSFLN 299
S0085O16            KDTLKSRTSVDMQMAFVTQALQFLENSYKNYTMVTVFGNLHQAQLGGVPGTYQLVCSFLN 299
Takifugu            KDALKSRTSVEMQMAFVRQALSCLENSYKNYTMVNVFGNLHQAQLGGVPGTYQLVRSFLN 300
Zebrafish           KDTLKSRVSVDMQMAFVRQALQFLENSYKNYTLVTVFGNLHQAQLGGVPGTYQLVCSFLN 299
mouse               TDALKSRSSVEVRMDFVKQALGYLEQSYKNYTLVTVFGNLHQAQLGGVPGTYQLVRSFLN 298
rat                 TDALKSRSSVEVRMDFVKQALGYLEQSYKNYTLVTVFGNLHQAQLGGVPGTYQLVRSFLN 136
human               TDALKNRSSVEVRMEFVRQALAYLEQSYKNYTLVTVFGNLHQAQLGGVPGTYQLVRSFLN 298
chick               SDVLNVRTNLEVRMEFVRQALRYLEQSYKNYTFLTVFGNLHQAQLGGVPGTYQLVRSYLN 201
                    .*.*: *  ::::: ** ***  *::******::.******************** *:**

S0188I22            IKLPGPLPGMQDGEVEGHPVWALIYYCLRCGDLGAAMQVVNRAQHQLGDFKTWFQEYMNS 359
S0085O16            IKLPGPLPGMQDGEVEGHPVWALIYYCLRCGDLGAAMQVVNRAQHQLGDFKTWFQEYMNS 359
Takifugu            IKLPGPLPGMQDGEIEGHPVWAVIYYCLRCGDLNAAMQVVNRVQHQLGDFKTWLQEYMNS 360
Zebrafish           IKLPTPLPGRQDGEVEGHPVWALIYFCLRCGDLSAAMQVVNKAQHQLGDFKIWFQEYMNS 359
mouse               IKLPAPSPGLQDGEVEGHPVWALIYYCMRCGDLLAASQVVSRAQHQLGEFKTWFQEYMNS 358
rat                 IKLPAPSPGLQDGEVEGHPVWALIYYCMRCGDLLAASQVVSRAQHQLGEFKTWFQEYMNS 196
human               IKLPAPLPGLQDGEVEGHPVWALIYYCMRCGDLLAASQVVNRAQHQLGEFKTWFQEYMNS 358
chick               IKLPAPVPGLQDGEVEGHPVWALIYYCMRCGDLTAAMHVVKRAQHQLGEFKTWFQEYMNS 261
                    **** ● ** ****:*******:**:*:***** ** :**.:.*****:** *:******

S0188I22            --PDRRLAPATENKLRLHYRRVLRNCADPYKRAVYCLIGKCDIADNHGDVADKTEDYLWL 417
S0085O16            --PDRRLAPATENKLRLHYRRVLRNSADPYKRAVYCLIGKCDIADNHGEVADKTEDYLWL 417
Takifugu            PTPDRRLSPNSENKLRLHYRRVLRNSADPYKRAVYCLIGKCDISDNHGEIADKTEDYLWL 420
Zebrafish           --PDRRLSPATENKLRLHYRRVLRNSADPYKRAVYCLIGKCDIGDNHGEVADKTEDYLWL 417
mouse               --KDRRLSPATENKLRLHYRRALRNNTDPYKRAVYCIIGRCDITDNQSEVADKTEDYLWL 416
rat                 --KDRRLSPATENKLRLHYRRALRNNTDPYKRAVYCIIGRCDITDNQSEVADKTEDYLWL 254
human               --KDRRLSPATENKLRLHYRRALRNNTDPYKRAVYCIIGRCDVTDNQSEVADKTEDYLWL 416
chick               --KDRRLSTATENKLRLHYRRALRNNTDPYKRAVYCIIGRCDITDNQSEVADKTEDYLWL 319
                       ****:. :**********.*** :*********:**:**: **:.::**********
```

```
S0188I22         KLNQVCFDDDGSSAPQDRLTLAQLQKQLLEDYGESHFSASQQPFLYFQVLFLTAQFEAAV 477
S0085O16         KLNQVCFDDDGSSAPQDKLTLPQLQKQLLEDYGESHFSAIQQPFLYFQVLFLTAQFEAAV 477
Takifugu         KLNQVCFDDDGSSS-QDRMTLPQLQKQ-LEDYGESHFLASQQPFLYFQVLFLTAQFEAAL 478
Zebrafish        KLNQVCFDEDGSSSPQDRMTLAQLQKQLLEDYGESHFSASHQPFLYFQVLFLTAQFEAAI 477
mouse            KLNQVCFDDDGTSSPQDRLTLSQFQKQLLEDYGESHFTVNQQPFLYFQVLFLTAQFEAAI 476
rat              KLNQVCFDDDGTSSPQDRLTLSQFQKQLLEDYGESHFTVNQQPFLYFQVLFLTAQFEAAI 314
human            KLNQVCFDDDGTSSPQDRLTLSQFQKQLLEDYGESHFTVNQQPFLYFQVLFLTAQFEAAV 476
chick            KLNQVCFDDDGASSSQDRLTLSQFQKQLLEDYGESHFAVNQQPFLYFQVLFLTAQFEAAI 379
                 ********:**:*:* **::**.*:*** ********* . :******************:

S0188I22         AFLFRVERLRSHAVHVALVLYELRLLLKSSGQSAQLLSQEAGDPPMVRRLNFIRLLMLYT 537
S0085O16         AFLFRVERLRSHAVHVALVLYELGLLLKSSGQSAQLLSQEAGDPPMVRRLNFIRLLMLYT 537
Takifugu         AFLFRVERLRSHAVHVALVLHELQLLLKSSGQSAQLLSQEPGDPPMVRRLNFIRLLMLYT 538
Zebrafish        AFLFRVERLRSHAVHVALVLYELKLLLKSSGQSAQLLSQEAGDPPMVRRLNFIRLLMLYT 537
mouse            AFLFRMERLRCHAVHVALVLFELKLLLKSSGQSAQLLSHEPGDPPCMRRLNFVRLLMLYT 536
rat              AFLFRMERLRCHAVHVALVLFELKLLLKSSGQSAQLLSHEPGDPPCMRRLNFVRLLMLYT 374
human            AFLFRMERLRCHAVHVALVLFELKLLLKSSGQSAQLLSHEPGDPPCLRRLNFVRLLMLYT 536
chick            AFLFRTERLRCHAVHVALVLFELKLLLKSSGQSAQLLSHEVGDPPGMRRLNFVRLLMLYT 439
                 ***** ****.*********.** **************:* **** :*****:*******

S0188I22         RKFESTDPREALQYFYFLRNEKNNQGENMFMCCVSELVIESRE-FDMLLGRLEKDGSRKP 596
S0085O16         RKFESTDPREALQYFYFLRNEKNSQGENMFMCCVSELVIESRE-FDMLLGRLEKDGSRKP 596
Takifugu         RKFESTDPREALQYFYFLRNEKDSQGENMFLRCVSELVIESREAIDMLLGRLEKDGSRKP 598
Zebrafish        RKFESTDPREALQYFYFLRNEKDSQGENMFMRCVSELVIESRE-FDMLLGRLEKDGSRKP 596
mouse            RKFESTDPREALQYFYFLRDEKDSQGENMFLRCVSELVIESRE-FDMILGKLENDGSRKP 595
rat              RKFESTDPREALQYFYFLRDEKDSQGENMFLRCVSELVIESRE-FDMILGKLENDGSRKP 433
human            RKFESTDPREALQYFYFLRDEKDSQGENMFLRCVSELVIESRE-FDMILGKLENDGSRKP 595
chick            RKFESTDPREALQYFYFLRNEKDSQGENMFLRCVSELVIESRE-FDMILGKLENDGSRKP 498
                 *******************:**:.******: *********** :**:**:**:******

S0188I22         GVIDKFAGDTRAIITKVALEAENKGLFEEAVRLYELAKNPDKVLELMNRLLSPVIAQVSA 656
S0085O16         GVIDKFAGDTRAIITKVALEAENKGLFEEAVRLYELAKNPDKVLELMNRLLSPVIAQVGS 656
Takifugu         GVIDKFAGDTRAIISKVALEAENKGLFEEAVRLYELAKNPDKVLELMNRLLSPVIAQVSA 658
Zebrafish        GVIDKFAGDTRAIITKVASEAENKGLFEEAVKLYELAKNADKVLELMNKLLSPVIAQVSE 656
mouse            GVIDKFTSDTKPIINKVASVAENKGLFEEAAKLYDLAKNADKVLELMNKLLSPVVPQISA 655
rat              GVIDKFTSDTKPIINKVASVAENKGLFEEAAKLYDLAKNADKVLELMNKLLSPVVPQISA 493
human            GVIDKFTSDTKPIINKVASVAENKGLFEEAAKLYDLAKNADKVLELMNKLLSPVVPQISA 655
chick            GVIDKFNSDTKPIINKVASAAENKGLFEEAAKLYDLAKNPDKVLELMNKLLSPVVPQIST 558
                 ****** .**:.**.***  **********.:**:****.********:*****:.*:.

S0188I22         AQSNKERLKNTAVAIAERYRTQGVAGEKSADSTFYLLLDLMTFFDEYHAGNIDRAYDVME 716
S0085O16         TQSNKERLKNTAVAIAERYRTQGVAGEKSVDSTFYLLLDLMTFFDEYHVGHIDQAYDVME 716
Takifugu         PQSNKERLKNAAVAIAERYRCQGITAEKSIDSTFYLLLDLMTFFDEYHAGHVDRAYDVME 718
Zebrafish        PQSNKERLKNMAVAIAERYRANGVAGEKSVDNTFYLLLDLMTFFDEYHAGHIDRAYDVIE 716
mouse            PQSNKERLKNMALSIAERYRAQGISANKFVDSTFYLLLDLITFFDEYHSGHIDRAFDIID 715
rat              PQSNKERLKNMALSIAERYRAQGISANKFVDSTFYLLLDLITFFDEYHSGHIDRAFDIID 553
human            PQSNKERLKNMALSIAERYRAQGISANKFVDSTFYLLLDLITFFDEYHSGHIDRAFDIIE 715
chick            PQSNKERLKNMAHSIAERYKAQGISAKKSIDSTFYLLLDLITFFDEYHAGHIDRAFDIIE 618
                 .********* * :*****: :*::.:*  *.********:******* *::*:*:*:::

S0188I22         RLKLVPLSQDSVEERVAAFRNFSDEVRHNLSEVLLATMNILFTQCKRLKGATAGTPGRPQ 776
S0085O16         RLKLVPLSQDSVEERVAAFRNFSDEVRHNLSEVLLATMNILFMQYKRLKGATAGTPGRPQ 776
Takifugu         HLKLLPLSQDSVGERVSAFRNFSDEVRHNLSEVLLATMNILFTQHKRLKGAPAGTPGRSQ 778
Zebrafish        RLKLVPLSQDSVEERVAAFRNFSDEVRHNLSEVLLATMNILFTQYKRLKGAAAGTPGRPQ 776
mouse            RLKLVPLNQESVEERVAAFRNFSDEIRHNLSEVLLATMNILFTQFKRLKGTSPSSATRPQ 775
rat              RLKLVPLNQESMEERVAAFRNFSDEIRHNLSEVLLATMNILFTQFKRLKGTSPSSATRPQ 613
human            RLKLVPLNQESVEERVAAFRNFSDEIRHNLSEVLLATMNILFTQFKRLKGTSPSSSSRPQ 775
chick            RLKLVPLSQDCVEERVAAFRNFSDEIRHNLSEVLLATMNILFTQYKRMKGTSPATPARPQ 678
                 :***:**.*:.: ***.********.**************** * **:**:...:. *.*

S0188I22         RSIEDRDSQLRSQARALITFAGIIPYRMAGDTNARLVQMEVLMN 820
S0085O16         RSMEDRDSQLRSQARALITFAGIIPYRMAGDTNARLVQMEVLMN 820
Takifugu         RSMEDKD-MLRIQARALITFAGMIPYNMAGDTNARLVQMELLMN 821
Zebrafish        RTLEDRDMLLRIQARALITFAGMIPYRMAGDTNARLVQMEVLMN 820
mouse            RVIEDRDSQLRSQARALITFAGMIPYRTSGDTNARLVQMEVLMN 819
rat              RVIEDRDSQLRSQARALITFAGMIPYRTSGDTNARLVQMEVLMN 657
human            RVIEDRDSQLRSQARTLITFAGMIPYRTSGDTNARLVQMEVLMN 819
chick            RVIKDRDSQLRSQARALITFAGMIPYRTSGDTNARLVQMEVLMN 722
                 * ::*:*  ** ***:******:***. :***********.***
```

3.7.1.4 Beta-1,3-galactosyltransferase

In Atlantic salmon, the beta-1,3-galactosyltransferase duplicate genes are single exon genes. In S0085O16, the coding sequence is 1362 bp in length, resulting in a protein sequence of 453 amino acids, and in S0188I22, the coding sequence is 1368 nucleotides and 455 amino acids long. The beta-1,3-galactosyltransferase gene spans 105.9 to 107.3 kb in the S0085O16 sequence and 67.4 to 68.8 kb in the S0188I22 sequence, and lies on the plus strand in both cases.

Human, mouse, rat, and chick also have single exon beta-1,3-galactosyltransferase genes that are predicted to be orthologs of the Atlantic salmon duplicates based on their chromosomal location beside Lin10 (Figure 24). While the Atlantic salmon beta-1,3-galactosyltransferase proteins are 88% similar, they share approximately 50% sequence similarity with the tetrapod orthologs. In human, the gene is 377 amino acids long, in chick it is 415 amino acids, and in mouse and rat it is 397 and 399 amino acids, respectively. Sequence identity is higher in the latter portion of the homologs, likely reflecting the fact that the galactosyltransferase activity of the protein is found in this region.

Figure 24 **Beta-1,3-Galactosyltransferase amino acid multiple sequence alignment**

```
Rat             ---RRRPRLCRDAWLTLLLSAALGLLLYAQ-------------------RDGVSP----- 33
mouse           MRRRRRPRLCPDAWLTLLLSAALGLLLYAQ-------------------RDVASP----- 36
human           -----------------MRSATARPRRRAR-------------------REGEGG----- 19
Chick           ----MRVHLKGDAICTLFLVVALCSVLYTQ-------------------LEYLAPKGDKQ 37
S0085016        ---MRRIYINGDVLCTLLMLGLLCLMLYAHQGFTSTWDTWHLEQGCTSSRALLGPPPESH 57
S0188I22        ---MRRIHIKGDVLCTLLMLGLLCLLLYAHQCFTPTWDTWHLEQGSTSSRALLGALPESH 57
                                 :          ::                       .

Rat             -TTRAPPARGRQLPRP-------------TPGRRALELPNTAHAAPPAYEGETP------ 73
mouse           -TTR-PPARGPQLPRP-------------TPSLRARELPNTARAAPLAYEGDTP------ 75
human           -RHRGPPPDPARSSYP-------------TRVQPRRPTKGTHRRR--------------- 50
Chick           PTQKKPSVTQRIFSDPRAP---GRPTVTEATVPRPRLIPITRRTEVANSKVQTA------ 88
S0085016        VTKRVPSMPPDKTKCQSARQSHSKSQTHPKSKLQSKSKPQAKTKS--KSKKDDG------ 109
S0188I22        VTKLVPSMHPDKTKCQ------SEPRSHPKSKLQSKSKPKAKTKSRSKSKKDEGTKKVAP 111
                     *.

Rat             ---VPPTPTD-PFDFRRYLRAKDQRRFPLLINQPRKCHSDGASGGSLDLLIAVKSVAADF 129
mouse           ---VPPTPTD-PFDFGGYLRAKDQRRFPLLINQRRKCRSDGASGGSPDLLIAVKSVAADF 131
human           -----PRLRD-PFDFARYLRAKDQRRFPLLINQPHKCRGDGAPGGRPDLLIAVKSVAEDF 104
Chick           ---TPVPSTDSAFNFKRYFLNKDNRNFNLLINQPKKCRR--IPGG-PFLLIAIKSVVEDF 142
S0085016        TKAVPVLPTRPPFDFEGYLRDKDNRDFILLMDQPGKCSGE------PYMLIAIKSVVADF 163
S0188I22        TKAAPVLPTQPPFDFEGYLRDKDNRDFRLLIDQPGKCSGE------LYMLITIKSVAADF 165
                           .*:*  *:  **:* * **::*  **           :**::***. **

Rat             ERREAVRQTWGAEGRVQ-GALVRRVFLLGVPKGAGSG-----GAGTRTHWRALLEAESRA 183
mouse           ERREAVRQTWGAEGRVQ-GALVRRVFLLGVPKGAGSG-----GAGTRSHWRTLLEAESRA 185
human           ERRQAVRQTWGAEGRVQ-GALVRRVFLLGVPRGAGSGGADEVGEGARTHWRALLRAESLA 163
Chick           DRREIVRKTWGREGLVN-GEQIQRVFLLGTPK----------NRTSLATWETLMHQESQT 191
S0085016        ERRQVVRHTWGREGVFQDGQTVKTVFLLGVPR----------NKTALPLWDRLLAYESHT 213
S0188I22        ERRQVVRHTWGREGVLQDLQTVKTVFLLGVPR----------NKTALPLWDRLLAYESHT 215
                :**: **:*** ** .:    :: *****.*:          .  : . *  *:  ** :

Rat             YADILLWAFEDTFFNLTLKEIHFLSWASAFCPDVHFVFKGDADVFVHVRNLLQFLEPRDP 243
mouse           YADILLWAFEDTFFNLTLKEIHFLSWASAFCPDVHFVFKGDADVFVHVRNLLQFLELRDP 245
human           YADILLWAFDDTFFNLTLKEIHFLAWASAFCPDVRFVFKGDADVFVNVGNLLEFLAPRDP 223
Chick           YRDILLWDFMDTFFNLTLKEIHFLNWAAEFCHNVKFIFKGDADVFVNIENIVDFLERHNP 251
S0085016        FGDILLWDFDDTFFNLTLKETHFLQWVNDSCSNVQFIFKGDTDVYVNIENILEMVKGQKP 273
S0188I22        FGDILLWDFDDTFFNLTLKETHFLQWVNDSCSNVQFIFKGDADVYVNIDNILQMLKGQKP 275
                : ***** * ********** *** *.   * :*:*:****:**:*:: *:::::  :.*

Rat             AQDLLAGDVIVQARPIRARASKYFIPQAVYGLPVYPAYAGGGGFVLSGATLHRLAHACTQ 303
mouse           AQDLLAGDVIVQARPIRARASKYFIPRAVYGLPVYPAYAGGGGFVLSGATLRRLADACSQ 305
human           AQDLLAGDVIVHARPIRTRASKYYIPEAVYGLPAYPAYAGGGGFVLSGATLHRLAGACAQ 283
Chick           AEDLFVGDIIYNARPIRTRKSKYYIPETMYGLSIYPAYAGGGGFLLSSCTMRKLSRACGE 311
S0085016        DKDLFVGDIIHHAHPIRRRSSKYFVPEVVYCQTMYPSYAGGGGFVMSGHTARRLSEACQQ 333
S0188I22        DKDLFVGDIIHHARPIRRRSSKYFVPEFVYGQTMYPSYAGGGGFVMSGHTARRLSEACQQ 335
                 :**:.**:* :*:*** * ***::*. :*  . **:*******::*. * ::*: ** :

Rat             VELFPIDDVFLGMCLQRLRLTPEPHPAFRTFGISQPSAAPHLRTFDPCFYRELVVVHGLS 363
mouse           VELFPIDDVFLGMCLQRLRLTPEPHPAFRTFGISQPSAAPHLRTFDPCFYRELVVVHGLS 365
human           VELFPIDDVFLGMCLQRLRLTPEPHPAFRTFGIPQPSAAPHLSTFDPCFYRELVVVHGLS 343
Chick           VELFPIDDVFLGMCLQRISLKPILHEGFKTFGIVKPSAAPHLQTFDPCFYKDLMVVHSLK 371
S0085016        VELFPIDDVFLGMCLQRIGVKPSHHEGFRTFGIVRPSAAPHLQVFDPCFYRELMVVHSLT 393
S0188I22        VELFPIDDVFLGMCLKRIGVKPSRHEGFRTFGIVRPSAAPHLQVFDPCFYRELMVVHSLT 395
                ***************:*: :.*  * .*:**** :******* .******::*:***.*.

Rat             AADIWLMWRLLHGPQGPVCAHPQPVATGPFQWNS-------------------------- 397
mouse           AADIWLMWRLLHGPOGPVCAHPOPVATGPFOWNS-------------------------- 399
human           AADIWLMWRLLHGPHGPACAHPQPVAAGPFQWDS-------------------------- 377
Chick           VAEIWLMWNLLHSPR-LSCTQKRQVKK-PFQWKRRAQIATQTSPLS-------------- 415
S0085016        VPQIWLMWNLLHDPQ-LSCHSKLAPTLWHFKSRGKVLGTTGQKDSEITVEQDYDVKVFVK 452
S0188I22        VPQIWLMWNLLHDPQ-LSCHSNLAPTPWPFKWRGKVLGTTGQTDSETTVEQDYDVKVFVK 454
                ..:*****.***.*:   *          *.

Rat
mouse
human
Chick
S0085016        H 453
S0188I22        H 455
```

3.7.2 Incomplete Duplicate Pairs

Incomplete duplicate pairs are classified as such due to the absence of exons in the Atlantic salmon BAC sequence data that are found in the orthologous sequences in human, chicken, or zebrafish. An exon could be missing for several reasons: (1) its location falls into one of the gaps in the BAC sequence data, (2) it is positioned outside the boundaries of the sequence data available, or (3) it simply does not exist in Atlantic salmon. It is possible to account for the first two situations, while the third situation is irresolvable without mRNA data. Four of the ten genes pairs were classified as incomplete: CBFB, GNAO1, Herp, and Cetp.

3.7.2.1 CBFB

The CBFB gene in both human and zebrafish has 6 exons, separated by large introns, with the largest intron occurring between exons 3 and 4 (29kb in human, 55 kb in zebrafish). Within the Atlantic salmon S0085O16 sequence, exons corresponding to human and zebrafish exons 1, 2, 3, and 6 were found. Notably, after exon 1, the splice donor site was a GC instead of a GT. It is probable that exons 4 and 5 are absent from the S0085O16 sequence because there are gaps in the sequence where the exons are likely to be located. The expected location of these two exons is based on their corresponding location in the S0188I22 sequence, where exons 1 to 5 could be identified, but exon 6 was not found. Exon 6 is most likely missing due to sequence gaps in the S0188I22 BAC sequence data. As observed in human and zebrafish, there is a very large intron (53kb) between exons 3 and 4 in the S0188I22 sequence, in which the blue minisatellite is found in S0188I22. An amino acid multiple sequence alignment of the CBFB orthologs is shown in Figure 25.

Figure 25 CBFB amino acid multiple sequence alignment

```
Chick            -------------------------------------------------------AFVAT 5
S0188I22         -------------------------------------------------------AFVAT 5
Human            MPRVVPDQRSKFENEEFFRKLSRECEIKYTGFRDRPHEERQARFQNACRDGRSEIAFVAT 60
Rat              MPRVVPDQRSKFENEEFFRKLSRECEIKYTGFRDRPHEERQTRFQNACRDGRSEIAFVAT 60
Mouse            MPRVVPDQRSKFENEEFFRKLSRECEIKYTGFRDRPHEERQTRFQNACRDGRSEIAFVAT 60
Zebrafish        MPRVVPDQRSKFENEEFFRKLSRECEIKYTGFRDRPHEERQARFQNACRDGRSEIAFVAT 60
S0085016         MPRVVPDQRSKFENEEFFRKLSRECEIKYTGFRDRPHEERQARFHTACRDGRSEIAFVAT 60
                                                                        ●●●●●

Chick            GTNLSLQFFPASWQGEQRQTPTREYVDFEREGGKVYLKAPMILNGVCVIWKGWIDLQRLD 65
S0188I22         GTNLSLQFFPANLHGEQRQTPTRDYVDFDRETGKVYLKAPMILNGVCVIWKGCIDLQRLD 65
Human            GTNLSLQFFPASWQGEQRQTPSREYVDLEREAGKVYLKAPMILNGVCVIWKGWIDLQRLD 120
Rat              GTNLSLQFFPASWQGEQRQTPSREYVDLEREAGKVYLKAPMILNGVCVIWKGWIDLHRLD 120
Mouse            GTNLSLQFFPASWQGEQRQTPSREYVDLEREAGKVYLKAPMILNGVCVIWKGWIDLHRLD 120
Zebrafish        GTNLSLQFFPANLHGDQRQAPTREYVDFERETGKVYLKAPMILNGVCVIWRGWLDLHRLD 120
S0085016         GTNLSLQFFPANLHGEQRQTPTREYVDFDRETGK-------------------------- 95
                 ***********. :*:***:*:*:***::** **

Chick            GMGCLEFDEERAQQEDALAQQAFEEARRRTREFEDRDRSHREEMEVRVSQLLSVTG---- 121
S0188I22         GMGCLEFDEERAQHEDALAQASFEESRRRTRDFEDRDRSHREDLE--------------- 110
Human            GMGCLEFDEERAQQEDALAQQAFEEARRRTREFEDRDRSHREEMEVRVSQLLAVTGKKTT 180
Rat              GMGCLEFDEERAQQEDALAQQAFEEARRRTREFEDRDRSHREEMEVRVSQ---------- 170
Mouse            GMGCLEFDEERAQQEDALAQQAFEEARRRTREFEDRDRSHREEMEARRQQDPSPGSNLGG 180
Zebrafish        GMGCLEYDDERAQHEDALAQAAFEEARRRTRDFEDRDRSHREDLEPRRQQDPSPGSNMGN 180
S0085016         ---------------------------------------------PRRQQDPSPGTNMGN 110

Chick            --------
S0188I22         --------
Human            RP------ 182
Rat              --------
Mouse            -GDDLKLR 187
Zebrafish        -TDDHKMR 187
S0085016         NADDHKMR 118
```

3.7.2.2 GNAO1

Human GNAO1 has two known splice variants composed of 8 exons each, where exons 1 to 6 are common to both variants. After exons 1 to 6, the first splice variant is then comprised of two more exons that are located immediately downstream from exons 1 to 6, and will be referred to as exons 7 and 8. The second splice variant is comprised of two exons that will be designated exons 7' and 8', due to their sequence similarity to exons 7 and 8. Exons 7' and 8' are located further downstream from exons 1 to 6 and exons 7 and 8. This gene structure is also seen in chick, mouse, rat, and in each Atlantic salmon BAC sequence. The observation that the Atlantic salmon genome also contains the GNAO1 splice variants suggests that the duplication event that produced this gene structure is very ancient, and pre-dates the divergence of fish and tetrapods. In accordance with this hypothesis, the orthologous coding sequence for all of the splice variants that contain exons 7 and 8 are more similar to each other than to any of the splice variants containing exons 7' and 8'. Novel Ensembl gene predictions have been made for GNAO1 homologs in zebrafish and *Takifugu*, but the splice variants are not predicted.

Compared to the tetrapod sequences, in the Atlantic salmon BAC sequences exons 1 to 3 were not identified, presumably because they lie outside the boundaries of the sequence data available; the GNAO1 duplicates lie at the end of each BAC sequence. For this reason, a multiple sequence alignment was created that excluded exons 1 to 3 for all orthologs, such the coding sequence for exons 4 to 6 were aligned, followed by the coding sequence for exons 7 and 8, and finally exons 7' and 8' (Figure 26). A remarkably high degree of sequence conservation is observed between the GNAO1 orthologs.

Figure 26 GNAO splice variant amino acid multiple sequence alignments. Exon boundaries are indicated by underlines and are shown in the following order: 4, 5, 6, 7, 8, 7', and 8'.

```
Mouse          TDSKMVCDVVSRMEDTEPFSAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLD 60
Rat            ADSKMVCDVVSRMEDTEPFSAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLD 60
Human          ADAKMVCDVVSRMEDTEPFSAELLSAMMRLWGDSGIQECFNRSREYQLNDSAKYYLDSLD 60
Chick          ADAKMVCDVVSRMEDTEPFSPELLSAMMRLWADSGIQECFNRSREYQLNDSAQYYLDSLD 60
S0085016       ADAKLVCDVVSRMEDTEPYSPELLGAMIHLWSDSGIQECFSRAREYQLNDSAQYYLDSLD 60
S0188I22       ADAKMVCDVVSRMEDTEPYSPELLGAMIRLWSDSGIQECFSRAREYQLNDSAQYYLDSLD 60
               :*:*:*************:*.***.**::**.********.*:*********:*******

Mouse          RIGAGDYQPTEQDILRTRVKTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
Rat            RIGAADYQPTEQDILRTRVKTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
Human          RIGAADYQPTEQDILRTRVKTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
Chick          RIGAADYQPTEQDILRTRVKTTGIVETHFTFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
S0085016       RIGAPDYQPTEQDILRTRVKTTGIVETHFVFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
S0188I22       RIGAPDYQPTEQDILRTRVKTTGIVETHFVFKNLHFRLFDVGGQRSERKKWIHCFEDVTA 120
               **** **************************** **************************

Mouse          IIFCVALSGYDQVLHEDETTNRMHESLKLFDSICNNKWFTDTSIILFLNKKDIFEEKIKK 180
Rat            IIFCVALSGYDQVLHEDETTNRMHESLKLFDSICNNKWFTDTSIILFLNKKDIFEEKIKK 180
Human          IIFCVALSGYDQVLHEDETTNRMHESLKLFDSICNNKWFTDTSIILFLNKKDIFEEKIKK 180
Chick          IIFCVALSGYDQVLHEDETTNRMHESLKLFDSICNNKWFTDTSIILFLNKKDIFEEKIKK 180
S0085016       IIFCVALSGYDQVLHEDETTNRMHESMKLFDSICNNKWFTDTSIILFLNKKDIFEQKIKK 180
S0188I22       IIFCVALSGYDQVLHEDETTNRMHESMKLFDSICNNKWFTDTSIILFLNKKDIFEQKIKK 180
               **************************.****************************.****

Mouse          SPLTICFPEYTGPSAFTEAVAHIQGQYESKNKSAHKEVYSHVTCATDTNNIQFVFDAVTD 240
Rat            SPLTICFPEYTGPSAFTEAVAHIQGQYESKNKSAHKEVYSHVTCATDTNNIQFVFDAVTD 240
Human          SPLTICFPEYTGPSAFTEAVAYIQAQYESKNKSAHKEIYTHVTCATDTNNIQFVFDAVTD 240
Chick          SPLTICFPEYTGPSSFTEAVAYIQAQYESKNKSPNKEIYTHITCATDTNNIQFVFDAVTD 240
S0085016       SPLSICFSEYTGADTFTEAVAHIQSQYESRNKSLHKEVYAHVTCATDTNNIQFVFDAVTD 240
S0188I22       SPLSICFPEYTGTDTFMEAVGHIQSQYESRNKSLQKEVYAHVTCATDTNNIQFVFDAVTD 240
               ***:***.****..:* ***.:**.****:*** :**:*:*:******************

Mouse          VIIAKNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICF 300
Rat            VIIAKNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICF 300
Human          VIIAKNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFGEKIKKSPLTICF 300
Chick          VIIANNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFAEKIKKSPLTICF 300
S0085016       VIIANNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFAEKIKKSALSICF 300
S0188I22       VIISNNLRGCGLYNRMHESLMLFDSICNNKFFIDTSIILFLNKKDLFAEKIKKSALSICF 300
               ***::******************************************.******.*:***

Mouse          PEYPGSNTYEDAAAYIQTQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNL 360
Rat            PEYPGSNTYEDAAAYIQTQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNL 360
Human          PEYTGPNTYEDAAAYIQAQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNL 360
Chick          PEYAGPNTYEDAAAYIQAQFESKNRSPNKEIYCHMTCATDTNNIQVVFDAVTDIIIANNL 360
S0085016       PEYTGPNTYDDAAAYIQAQFESKNRSPNKEIYCHLTCATDTGNIQVVFDAVTDIIIANNL 360
S0188I22       PEYTGSNTYDDAAAYIQAQFESKNRSPNKEIYCHLTCATDTGNIQVVFDAVTDIIIANNL 360
               ***.*.***:*******:****************:******.******************

Mouse          RGCGLY 366
Rat            RGCGLY 366
Human          RGCGLY 366
Chick          RGCGLY 366
S0085016       RGCGLY 366
S0188I22       RGCGLY 366
               ******
```

3.7.2.3 Herp

The Herp gene is 8 exons long in human, and it is predicted to contain the same number of exons in chick and zebrafish. Based on this information, 8 Herp exons were identified in the S0085O16. In the S0188I22 sequence, on the other hand, only exons 1 to 3 could be found, with the third exon lying only 800 bp from one end of the BAC sequence data; it is likely that the other five exons, if they exist, are located outside the sequence data available. A rainbow trout EST was available to confirm the intron/exon boundaries for the first 6 exons.

The first three exons of the Atlantic salmon Herp gene duplicates are 90% identical at both the nucleotide and amino acid level. Considering only the amino acid sequence coded for by the first three exons to orthologous sequences from human, mouse, rat, chick, and zebrafish (Figure 27) reveals that the salmon duplicates are 67% similar to zebrafish, and range from 31 to 43 % identical to orthologous tetrapod proteins.

The orthologous proteins are more conserved in the N-terminal region, likely because the ubiquitin domain lies within the first 100 amino acid residues. Beyond this point, chick, zebrafish, and S0085O16 exhibit variation in exon length relative to human, mouse, and rat, although for the most part, sequence conservation remains high.

Figure 27 Herp amino acid multiple sequence alignment

```
Mouse           MEPEPQ----PEPVTLLVKSPNQRHRDLELSGDRS-WSVSRLKAHLSRVYPERPRPEDQR 55
Rat             MEPEPQ----PEPVTLLVKSPNQRHRDLELSGDRG-WSVSRLKAHLSRVYPERPRPEDQR 55
Human           MESETE----PEPVTLLVKSPNQRHRDLELSGDRG-WSVGHLKAHLSRVYPERPRPEDQR 55
Chick           ---MAE----DLSLSLLVRSPARRHPDLRLRAAPA-WSVRRLKAELRRRAPGAPAEEDQK 52
S0085016        MDNTGF--LRQKTIKLVIKTPNQAHGDQTIEGVDMDWTVKELKTHLSRMYPNNPAESDQR 58
S0188I22        MDNSGFPNVRQKTITLVIKTPNQAHGDQTIEGVDTDWTVKELKTHLSRVYPNNPAESDQR 60
Zebrafish       MENSTV--FDKETISLVIKTPNQFHGDQLIEGVRADWTVKDLKCHLSKVYPNNPAEKDQR 58
                            .:.*::::* : * *  : .    *:*  ** .* :  *  *  .**:

Mouse           LIYSGKLLLDHQCLQDLLPKQ--EKRHVLHLVCNVKN-PSKMPETSTKGAESTEQPDNSN 112
Rat             LIYSGKLLLDHQCLQDLLPKQ--EKRHVLHLVCNVRS-PSKKPEASTKGAESTEQPDNTS 112
Human           LIYSGKLLLDHQCLRDLLPKQ--EKRHVLHLVCNVKS-PSKMPEINAKVAESTEEPAGSN 112
Chick           LIYCGKLLLDHQLLREFLPGQ--EELHALHLVYNMRT-PTDVPESSTEVLTLSEGTAFSA 109
S0085016        LIYSGKLLPDHLHVREIFRKT--DLTPTVHLVCAVRTQPIGPLGARPKVRESEQQEAQTS 116
S0188I22        LIYSGKLLPDQLHVRDIFRKT--DLTPTVHLVCAVRTQPRGPLGARPK------------ 106
Zebrafish       LIYSGKLLLDNLLIRDVFSKVPSDTKPTLHLVCAVRPQPASQLGARPKSSQPSPLTPSQS 118
                ***.**** *:  :::.:     :   .:***  ::  *       .:

Mouse           QTQ----------HPGDSSSDGLRQREVLRNLSPSGWENISRPEAVQQTFQGLG-PGFSG 161
Rat             QAQ----------YPGDSSSDGLREREVLRNLPPSGWENVSRPEAVQQTFQGLG-PGFSG 161
Human           RGQ----------YPEDSSSDGLRQREVLRNLSSPGWENISRPEAAQQAFQGLG-PGFSG 161
Chick           GVQ----------EPAAASSGGGRLRCSPGGQAAA--EADARPETPRHPFQAVP-PGFSV 156
S0085016        AMPTGQNPEGASPAPSVPSEPELRQRRPPAPSHTP--PAAAWPGTTTLVAAEMTNPTFPT 174
S0188I22        ------------------------------------------------------------
Zebrafish       SGP---------SVTSLHSTDGLRQRGHATLPDTS--ANTSAP-----VTAAMNHPAFPT 162

Mouse           YTTYGWLQLSWFQQIYARQYYMQYLAATAASGTFVPTPSAQEIPVVSTPAPAPIHNQFPA 221
Rat             YTTYGWLQLSWFQQIYARQYYMQYLAATAASGAFGPTPSAQEIPVVSTPAPAPIHNQFPA 221
Human           YTPYGWLQLSWFQQIYARQYYMQYLAATAASGAFVPPPSAQEIPVVSAPAPAPIHNQFPA 221
Chick           YTTYSMLQMSWLQQIYARQYYMQYLASTAASADPSSTRRSSEIPVT---PPAPLPDPFPA 213
S0085016        YSLYSPQQLLWLQHMYARQYYMQYHAAYAAAAASVPFAPAAGPSLPVAPHQAAIPAALP- 233
S0188I22        ------------------------------------------------------------
Zebrafish       YSLYSPQQLLWLQQMYARQYYMQYQAAMAAAASAPMTTPAAASSLPVGPHQAAVPAALPN 222

Mouse           E----NQPANQNAAAQAVVNP-GANQNLRMNAQGGPLVEEDDE-INRDWLDWTYSAATFS 275
Rat             E----NQPANQNAAAQAVVNP-GANQNLRMNAQGGPLVEEDDE-INRDWLDWTYSAATFS 275
Human           E----NQPANQNAAPQVVVNP-GANQNLRMNAQGGPIVEEDDE-INRDWLDWTYSAATFS 275
Chick           Q----NQPGDQNAAPQANV----ANQNLRMNAQGGPLMGEEEEGGNRDWLDWLYSATLFY 265
S0085016        --------ANQNAQDAAFINPGAANQNLRMNAQGGPVMEDEED-IDRDWLDWVYTAARLG 284
S0188I22        ------------------------------------------------------------
Zebrafish       QGPINDLPANQNAPGPAFINPEGANQNLRMNAQGGPVVEDEED-MNRDWLDWMYTASRLG 281

Mouse           VFLSILYFYSSLSRFLMVMGATVVMYLHHVGWFPFRQRPVQNFPDDGGPRDAANQDPNNN 335
Rat             VFLSILYFYSSLSRFLMVMGATVVMYLHHVGWFPFRQRPVQNFPDDGPPQEAANQDPNNN 335
Human           VFLSILYFYSSLSRFLMVMGATVVMYLHHVGWFPFRPRPVQNFPNDGPPPDVVNQDPNNN 335
Chick           VFVNIVYFYSSVSRFLLVMGGTVLMYLHHVGWFPFRQRRAQPFPDNVPPQAAVNQDQNNN 325
S0085016        VFLSIVYFYSSLSRFVLVMSSLLLMYLHTAGWFPFRRRPLVRGPNNQVPDVIQNH-QDRN 343
S0188I22        ------------------------------------------------------------
Zebrafish       VFLSIVYFYSSMSRFILVMSSLVIMYLHTAGWFPFRQRPQARPQNEPAPEVNQNQ-QNQN 340

Mouse           LQGGMDPEMEDPNRLPPDREVLDPE-HTSPSFMSTAWLVFKTFFASLLPEGPPALAN 391
Rat             LQGGLDPEMEDPNRLPVGREVLDPE-HTSPSFMSTAWLVFKTFFASLLPEGPPALAN 391
Human           LQEGTDPETEDPNHLPPDRDVLDGE-QTSPSFMSTAWLVFKTFFASLLPEGPPAIAN 391
Chick           LQGGNAGRAEEPEALPDAGQVLPELPQVNPSLMSTAWLFFKTFFASLLPEGPRLTRN 382
S0085016        PVPPAELDEPHP------LTAVLVP-PHRVSIVWTAWIFFKAFFASLVPEG-LAVAN 392
S0188I22        ---------------------------------------------------------
Zebrafish       EDQHPEPVMEDVGVVDPAMTAVPVP-QVRAPILWTAWVFFKAFFASLIPEPPQGIAN 396
```

3.7.2.4 Cetp

Cetp is only found on the minus strand of the S0085O16 sequence (19 to 32kb); the S0188I22 BAC sequence does not extend far enough to include any exons of this gene. The human Cetp ortholog is 16 exons in length, and all 16 of these exons were located in the S0085O16 Atlantic salmon sequence. One consensus splice site, however, is absent in intron fourteen.

In general, the Cetp gene seems to be quite diverged among Atlantic salmon, zebrafish, *Takifugu*, human, and chick (40-50% sequence identity) (Figure 28). The full-length mRNA for Cetp must be isolated and sequenced from Atlantic salmon in order to verify the intron and exon boundaries that have been tentatively annotated for this gene. Of note, a rainbow trout EST was available to help discern the 5' end of this gene, although the EST sequence only covered the boundaries of the first three exons.

Figure 28 Cetp amino acid multiple sequence alignment

```
Takifugu         MPCEVLRLPLHCLLLLSVIGLSQACLEDPASAYRFTGAVCRLTYPAAVVLNEKTTKVIEA 60
S0085016         MDKCLMTLPV-LLLFLGLVGISRSCLN-PVSAYRFTGAVCRLTYSAAMVLNEKTTKVIQA 58
Zebrafish        ------------------------------------GAVCRLTYPAAVVLNEKTTEVIQA 24
Human            -------MLAATVLTLALLGNAHACSKGTSHEA---GIVCRITKPALLVLNHETAKVIQT 50
Chick            ------------------------------------GIVFRMTKPAALLLNQETARLIQA 24
                                                     * * *:* .* ::**.:*:.:*::

Takifugu         AFQHARYPNLKGEKSLSFVGTIRYGLENLEVHNLSIGASEFELHPNEGISMEISNVSAVF 120
S0085016         AFQHAKYPSINSEKSILFVGKVKYGLTNLEIHNLSIGRSEFELKPVKGIEIAISNVSAVF 118
Zebrafish        AFQHAKYPSVQGERSIGFG-TVKYGFHNLEIHNLSIGKSEFELKENVGIGIAISNVYAVF 83
Human            AFQRASYPDITGEKAMMLLGQVKYGLHNIQISHLSIASSQVELVEAKSIDVSIQNVSVVF 110
Chick            AFKNAKFPNITGERSMRFLGTVAYTLANIQVSDLSIEQSEVELKENDAIDIAIKNVTAFF 84
                 **:.* :*.: .*::: :   : * : *::: .***  *:.**     * : *.** ..*

Takifugu         RGTIQYGYG-SWLVSVANSIDFEIQSQIDLGINPKLHCKEGKVAADTSDCYLRFDKLLLH 179
S0085016         KGTIQYEYG-SWLVNVGHSVDFEIESHIDLGINPKLYCGKEKVAADTSVCYLTFHKLNLL 177
Zebrafish        KGTINYGYG-SWLLSLNQSMDFEVESQIDLVINPKLYCGKGKVAADTSDCYLTFHKLKLL 142
Human            KGTLKYGYTTAWWLGIDQSIDFEIDSAIDLQINTQLTCDSGRVRTDAPDCYLSFHKLLLH 170
Chick            RGTLTYGYAGAWFLQLFHSVDFEIQSSIDLQINIKLLCQEEQVAADASDCYLSFHKLMLH 144
                 :**: * *  :* : : :*:***::* *** ** :* * . :* :*:. *** * ** *

Takifugu         LQGDKEPNWLKKLFTDFISFTVKMAIKGQ-ICKEINKVANILADFIQDTAEHFLSDGDIS 238
S0085016         LQGDKEPGWLKRLFTDFITFTGKLVIKSQPICKEINNVANILADFIQERAEQFLSDGDIS 237
Zebrafish        LQGDREPGWMKKMFTDIVTFTVKLVVKSQ-ICKEINSVANILADFIQEQAEQFLSEGGIG 201
Human            LQGEREPGWIKQLFTNFISFTLKLVLKGQ-ICKEINVISNIMADFVQTRAASILSDGDIG 229
Chick            LQGDKEPGWLKQLFTDFISFTLKFVLKRE-LCKEINLLAQVMANFVHNVAD-------AS 196
                 ***::**.*:*::**::::** *:.:* : :***** :::::*:*::           ..

Takifugu         MDIGVTAAPIITANYIESYHKGLTKYNNGSAVINDSVFHPKQLTENRMLYFWLS------ 292
S0085016         MNIGVTSAPVITSHYIESYHKGLAKYNNVTAVINASVFHPSQLTEDRMLYFWISDEVFNT 297
Zebrafish        VDISVTSSPVIKSNYIESYHKGLVTFNNETSDLSNSVFHPSQLSEGRMLYFWFSDGLLDP 261
Human            VDISLTGDPVITASYLESHHKGHFIYKNVSEDLPLPTFSPTLLGDSRMLYFWFSERVFHS 289
Chick            VDIFFCLPAVCPCORSRPLFKGLVLYKNYSDVLSDSVFSPSLLSESRMLYFWISEHILNS 256
                 ::* .   .:  .   .. .**   ::* :  :  ..* *. * :.******:*

Takifugu         -------DGRFQANFSGTEFT--------------------------------------- 306
S0085016         LITASHQDGRFVLNISGLELTYLSSEMSEFLSQLLSSEE-------PFLRVWSVSVPQMW 350
Zebrafish        LLGAIHNDGRFVRSISGTELTVRYYNISIRCGGDISN----------------VMYCKLC 305
Human            LAKVAFQDGRLMLSLMGDEFKAVLETWGFNTNQEIFQEVVGGFPSQAQVTVHCLKMPKIS 349
Chick            LASAAFLDGRLVLAIRGEKLQVTCR--GGSSVLHVCLIFQGNSYNDSVAKVWSLALPEIS 314
                        ***:   : * ::

Takifugu         ---------------------------------NVVVDVTASYADKKLSLHGKSSEIYGL 333
S0085016         TTPLGTSVRALAAVKLTS---GAEDIPGLYFETEVEVVVRTSYAEKKLILNATVPQIS-I 406
Zebrafish        LYPSVSLHVGVINVLTVE---VFVRFFVFCHQHSISITKVSSSEGAIEVRVILIRLIFYI 362
Human            CQNKGVVVNSSVMVKFLFPRPDQQHSVAYTFEEDIVTTVQASYSKKKLFLSLLDFQITPK 409
Chick            LQPEGTVVKSLVAVEISIFPPGEEPLTALYMEEEITVTIQAAYVEKKLILRPVDSQIEFK 374
                                                  .:     ::              *

Takifugu         QAELPLQEKPLLDEAQLEFLRETVKKIGVSKVLSVLEIALTRLLDKQGVYLFDIFNPEVL 393
S0085016         MKESLSKDTLQMVEAHIEYLRGSEKNR--CPKSNKLEPGLTALMDKQGANLFDIIKPQVV 464
Zebrafish        LYHAHYVISSKMEESLRAYLQEAVEKMGIPKVVSYLEVELTSLMDEHGLNLFDIFNPEVV 422
Human            TVSNLT-ESS--SESVQSFLQSMITAVGIPEVMSRLEVVFTALMNSKGVSLFDIINPEII 466
Chick            VFNCTA-DPSGNDQSVRNFLQKMISAVGIPEVISKIEPALTSLMNSKGLHLFEIKNPEII 433
                              ::   :*:            . :*  :* *::.:*  **:* :*:::

Takifugu         PQD------------------------- 396
S0085016         SQDGYVIIQTDFGFPHHLLVEFLRKTLE 492
Zebrafish        PQDGYVMVQLDFGFPQHLLVEFLK---- 446
Human            TRDGFLLLQMDFGFPEHLLVDFLQSLS- 493
Chick            TRKRYLIVQLDFSFPNHLLLDFLE---- 457
                 .:.
```

3.7.3 Complete Gene/Pseudogene Duplicate Pairs

In two cases, BBS2 and SLC12A3, one gene in the duplicate pair was observed to be complete, while the accumulation of deleterious mutations in the second copy has led to pseudogenization.

3.7.3.1 BBS2

In the S0085O16 sequence, the BBS2 gene exists as a 17 exon gene on the plus strand, spanning 5 kb (98-103 kb), and encodes a 717 amino acid protein (2154 nucleotides). An Atlantic salmon EST was available to confirm the intron/exon boundaries of the last 6 exons. In comparison, both human and zebrafish also have known genes with 17 exons, while in chick the gene is a novel Ensembl prediction, for which the first two exons from human and zebrafish are not predicted (Figure 29).

In comparison to the S0085O16 BBS duplicate, the S0188I22 BBS2 locus is a pseudogene and contains highly recognizable remnants of the BBS2 gene, but has accumulated deleterious mutations, leading to six nonsense mutations, scattered in exonic regions throughout the gene. Additionally, several acceptor and donor splice sites have degenerated, and there does not appear to be a recognized start codon in exon one.

Figure 29 BBS2 amino acid multiple sequence alignment

```
Mouse           MLLPVFTLKLRHKISPRMVAIGRYDGTHPCLAAATQAGKVFIHNPHTRSQHFSASRVFQS 60
Rat             MLLPVFTLKLRHKISPRMVAIGRYDGTHPCLAAATQAGKVFIHNPHMRSQHFSTSRVFQS 60
Human           MLLPVFTLKLRHKISPRMVAIGRYDGTHPCLAAATQTGKVFIHNPHTRNQHVSASRVFQS 60
Chick           ------------------------------------------------------------
S0085016        MLVPIFTLKLNHKINPRMVTMGKFDGIHPCLTAATQAGKVFIHNPHTRGQRQAAHRLSQS 60
Zebrafish       MLVPIFTLKLNHKINPRMVAIGKYDGIHPCLTAATQAGKVFIHNPHTRAQRPTAHRLSQS 60


Mouse           PLESDVSLLNINQTVSCLGSGVLNPELGYDTLLVGTQTSLLAYDIYNNSDLFYREA---N 117
Rat             PLESDVSLLNINQTVSCLGAGVLNPELGYDTLLVGTQTSLLAYDIYNNSDLFYREASVAD 120
Human           PLESDVSLLSINQAVSCLTAGVLNPELGYDALLVGTQTNLLAYDVYNNSDLFYREAT--D 118
Chick           --------------VS-------------------------------------------D 3
S0085016        AQDSDISLLNINQAVSCLTAGTLGPNTTGDTLLVGTQTNLLAYDVHDNADIFYREVT--D 118
Zebrafish       TQDSDISLLNINQSVSCLTGGTLGPKSTGDTLLVGSQTNLLAYDVHDNTDVFYKEVT--D 118
                              **                                           :

Mouse           GANAIVLGTLGDIAPPLAIIGGNCALQGFDHEGNDLFWTVTGDNVHSLALCDFDGDGKTE 177
Rat             GANALVLGTLGDIAPPLAIIGGNCALQGFDHEGNDLFWTVTGDNVHSLALCDFDGDGKSE 180
Human           GANAIVLGTLGDISSPLAIIGGNCALQGFNHEGSDLFWTVTGDNVNSLALCDFDGDGKKE 178
Chick           GANAIVLGKLGDIPAPLAIIGGNCALQGFDYEGNDLFWTVTGDNVRSLALCDFDGDGKTE 63
S0085016        GANAIVLGKLGNIESPVAIIGGNCALQGFDYEGSDQFWTVTGDNVRSLVLCDFTGDGKNE 178
Zebrafish       GANAIVLGKLGDIQSPLAIIGGNCALQGFDYEGNDLFWTVTGDNVRSLVLCDFTADGKNE 178
                ****:***.**:* .*:************::**.* *********.**.**** .***.*

Mouse           LLVGSEDFDIRVFKEDEIVAEMTETEIVTSLCPMYGSRFGYALSNGTVGVYDKTARYWRI 237
Rat             LLVGSEDFDIRVFKEDEIVAEMTETEIVTSLCPMYGSRFGYALSNGTVGVYDKTARYWRI 240
Human           LLVGSEDFDIRVFKEDEIVAEMTETEIVTSLCPMYGSRFGYALSNGTVGVYDKTSRYWRI 238
Chick           LLVGSEDFDIRVFKEDEMVAEMSETETVTALSPMYGSRFGYALSNGTVGVYDRTSRYWRI 123
S0085016        LLVGSEDFDIRVFKEDELVAEIAENETVTSLCHMHGSRFGYALANGTVGVYDRTARYWRI 238
Zebrafish       LLVGSEDFDIRVFKEDELVTEMAENETVTSLCHMHGSRFGYALANGTVGVYDRTARYWRI 238
                *****************:*:*::*.* **:*. *:********:********:*:*****

Mouse           KSKNHAMSIHAFDINSDGVCELITGWSNGKVDARSDRTGEVIFKDNFSSAVAGVVEGDYR 297
Rat             KSKNHAMSIHAFDINSDGVCELITGWSNGKVDARSDRTGEVIFKDNFSSAVAGVVEGDYR 300
Human           KSKNHAMSIHAFDLNSDGVNELITGWSNGKVDARSDRTGEVIFKDNFSSAIAGVVEGDYR 298
Chick           KSKNHAMSIHAFDLNSDGVCELITGWSNGKVDARSDRTGEVIFKDNFASSIAGVVEGDYR 183
S0085016        KSKNHAMSIHAFDLNADGVVELITGWSNGKIDARSDRTGEVIFKDNFSSSVAGVVEGDYR 298
Zebrafish       KSKNHAMSIHAFDLNADGVVELITGWSNGKIDARSDRTGEVIFKDNFSSSVAGVVEGDYR 298
                *************:*:*** **********:****************:*::*********

Mouse           MDGHVQLICCSVDGEIRGYLPGTAEMKGNLLDTSVEQDLIRELSQKKQNLLLELRNYEES 357
Rat             MDGHVQLICCSVDGEIRGYLPGTAEMKGNLLDTSVEQGLIRELSQKKQNLLLELRNYEEN 360
Human           MDGHIQLICCSVDGEIRGYLPGTAEMRGNLMDTSAEQDLIRELSQKKQNLLLELRNYEEN 358
Chick           MDGSTQLICCSVDGEVRGYLPGGEEMKGNLMDTCAEQDMIRELSQKKQNLLLELRNYEEN 243
S0085016        MDGQIQLICTSVEGEVRGYLPASKEMKGNLMDSSVEQDLIRELSQRRQNLLLELRNYEEN 358
Zebrafish       MDGQIQLICTSVEGEVRGYLPASKELKGNLMDSSIEQDLIRELSQRKQNLMLELRNYEEN 358
                ***  **** **:**:*****.  *::***:*:. **.:******::***:********

Mouse           TKAELSSPLNEADGQKGIIPANTRLHTALSVNMGNDLQDAHAELGISTSNDTIIRAVLIF 417
Rat             TKAELSSPLNEADGQKGIIPANTKLHTALSVNLGNDAQDAHAELRISTSNDTIIRAVLIF 420
Human           AKAELASPLNEADGHRGIIPANTRLHTTLSVSLGNETQTAHTELRISTSNDTIIRAVLIF 418
Chick           AKAELSPQLKEADGQRGVIPANTQLQTSLSVNLGSDSQSAHVELCISTTNDTIIRAVLIF 303
S0085016        AKQAVPG-ASERDTQMGVIPANTQLQTVLSVREATESQRSHIELSISTPNETIIRAVLIF 417
Zebrafish       AK-ALPG-LSEGESKMGVIPANTQLQTALSVRRASESQKAHIELNISTPNETIIRAVLIF 416
                :*  :.   .* : : *:*****:*:* ***  ..: * :* ** ***.*:*********

Mouse           AEGIFVGESHVVHPSIHNLSSSLRVPITPPKDVPVDLHLKTFVGYRSSTQFHVFELTRQL 477
Rat             AEGIFAGESHVVHPSTHNLSSSIRVPITPPKDVPVDLHLKTFVGYRSSTQFHVFELIRQL 480
Human           AEGIFTGESHVVHPSIHNLSSSICIPIVPPKDVPVDLHLKAFVGYRSSTQFHVFESTRQL 478
Chick           AEGIFEGESHVVHPDLQNLSGCIRVPLTPPKDVPVDLHIKAFVGYRNSTQFYVFELTRQL 363
S0085016        AEGIFEGESHVVHPSAQNLCGSIRVPIIPPKDIPVDLHIKAFVGGKSSSQFHVFEITRQL 477
Zebrafish       AEGIFEGESHVVHPSAQNLSGCVRVPIIPPKDIPVDLHIKAFVGGKTSTQFHVFEITRQL 476
                ***** ********. :**...: :*: ****:*****:*:*** :.*:**:***  ***


Mouse           PRFTMYALTSPDAASEPVSYVNFSVAERTQRMVTWLNQNFLLPEDSNVQNSPFHVCFTSL 537
Rat             PRFTMYALTSPDAASEPVSFVNFIVVERAQRMVTWLNQNFLLPEDSNIQNAPFHVCFTSL 540
Human           PRFSMYALTSLDPASEPISYVNFTIAERAQRVVVWLGQNFLLPEDTHIQNAPFQVCFTSL 538
Chick           PRFSMYALSSPDSATEPLSFVSCATNERPQRIVMWLNQSFLLPEDAEFQSAPFQVCFTSL 423
S0085016        PRFSMYDLNVEPEAPQPTGKVTFTINDRPQRVVMWLNQNFLLLEGIDTPD----VTFTSL 533
Zebrafish       PRFSMYDLNVDPSAPEPTGKVTFCINDRPQRVVMWLNQNLLLPEGIDSPD----VTFSAL 532
                ***:** *.    *.:* . *.    :*.**:* **.*.:** *. .  .    * *::*
```

```
Mouse           RNGGQLYIK-MKQSGEITVNTDDIDLAGDIIQSIASFFAIEDLQVEADFPVYFEELRKVL 596
Rat             RNGGQLYIK-MKPSGEITVNTDDIDLAGDIIQSMASFFAIEDLQVEADFPVYFEELRKVL 599
Human           RNGGHLHIK-IKLSGEITINTDDIDLAGDIIQSMASFFAIEDLQVEADFPVYFEELRKVL 597
Chick           RNAGQLLIK-IKPGGEISISTDDIDLAGDIIQSMASFLAIEDLQVEADFPAYFEELRKVL 482
S0085016        RGGGLLTISMLSTSGEITLNTDDIDLAGDLVQSLASFLAIEDLQAEADFPTYFGELRTTL 593
Zebrafish       RGGGLLRIS-LQTSGEITLRTDDIDLAGDLVQSLASFLAIEDLQAESDFPVYFKELRATL 591
                *..* * *. :. .***:: *********::**:***:******.*:***.** *** .*

Mouse           VKVDEYHSVHQKLSADMADNSNLIRSLLVRAEDARLMRDMKTMKSRYMELYDLNKDLLNG 656
Rat             LKVDEYHSVHQKLSANMADNSNLIRSLLVRAEDARLMRDMKTMKTRYMELYDLNKDLLNG 659
Human           VKVDEYHSVHQKLSADMADHSNLIRSLLVGAEDARLMRDMKTMKSRYMELYDLNRDLLNG 657
Chick           VKVDEHHSVNQRLTADMADHSNLIRVMLVQAEDARLLGDMKNMKARYVELYDLNRDLINQ 542
S0085016        TEVDDYHSVHQKLTAAIADHSNHIRNMLVQAEDARLMGDIRNMKKRYIELYDLNRDLINE 653
Zebrafish       TE---FHSVHQKLTAAMADHSPYIRDMLVPAEDARLLGDC--------DVVVCGGHLTKI 640
                    :.***:*:*:* :**:*  ** .** ******. *         ::   . .* :

Mouse           YKIRCNNHTELLGNLKAVNQAIQRAGRLRVGKPKNQVISACRDAIRSNNINTLFRIMRVG 716
Rat             YKIRCNNHTELLGNLKAVNQAIQRAGRLRVGKPKNQVISACRDAIRSNNINTLFRIMRVG 719
Human           YKIRCNNHTELLGNLKAVNQAIQRAGRLRVGKPKNQVITACRDAIRSNNINTLFKIMRVG 717
Chick           YKIRCNNHTELLNNLKAVNQAIQRAGRLRVGKPKAQVIAACRDAIRSNNFNTLFRIMRSG 602
S0085016        YKIRSNNHNALLACLKSVNQAIQRAGRLRVGKPKNQVITACRDAIKNNNVNVLFKIMKAG 713
Zebrafish       YKLLT---LPLFIGFNCYPLLIHLKSCFSVGKPKSQVITACRDAIKNNNINALFKIMRAG 697
                **:       *:  ::.    *:  . : ***** ***:******:.**.*.**:**: *

Mouse           TAPS 720
Rat             TAPS 723
Human           TASS 721
Chick           VASS 606
S0085016        TASS 717
Zebrafish       TTSS 701
                .:.*
```

3.7.3.2 SLC12A3

In the case of SLC12A3, the Atlantic salmon complete gene is found in the S0188I22 sequence (3.5 to 11.5kb), whereas the pseudogene is located in S0085O16. In human, the SLC12A3 gene is known to have 26 exons and in chick the predicted gene also has 26 exons. In zebrafish the predicted gene has 27 exons, where exon 9 is only 2 nucleotides long; the first exon in zebrafish is also shorter than expected and the current prediction does not contain a start codon. The Atlantic salmon sequence is similar to zebrafish, in that the coding sequence will result in an in-frame translation only if a 2 nucleotide ninth exon is included, and therefore the gene should be considered to have 27 exons in total. Rainbow trout and Atlantic salmon ESTs were available to determine the proper intron/exon boundaries for exons 13 to 27. Presently, the first exon in salmon is also shorter than in human and is missing a start codon, similar to zebrafish.

The SLC12A3 homologs are long proteins, averaging 1000 amino acid residues, where the sequence similarity among tetrapods (chicken and human) and among fish (zebrafish and Atlantic salmon) is about 75%, but between tetrapods and fish is only 65% (Figure 30). With respect to the Atlantic salmon pseudogene located in S0085O16, what remains of the coding sequence for this gene has degenerated quite a bit more than the BBS2 pseudogene seen in S0188I22. In this case, it is not possible to identify the location of all the exons, although remaining identifiable exons lie on the minus strand between 45-53kb in the S0085O16 sequence data, between the Herp and dead eye genes.

Figure 30 SLC12A3 amino acid multiple sequence alignment

```
Takifugu        DGIHLAAYKPRHSPSLPEGNGSAGGYSDSDYYHRYGDGSNLASS----GSDALT---GYE 53
Zebrafish       ------------------------------------DTDSVASR----SSGVFS---GYD 17
S0188I22        ------------------------------------DGTSVASQN---STQILT---GYD 18
Rat             -MAELPVTEMPGDA-LCSGRFTISTLLGGDEP-PPAACDNSQPSHLTHGSTLYLRTFGYN 57
Mouse           -MAELPVTELPGDA-LCSGRFTISTLMGGDEP-PPAACDSSQPSHLTHGSTLYMRTFGYN 57
Human           -MAELPTTETPGDATLCSGRFTISTLLSSDEPSPPAAYDSSHPSHLTHSSTFCMRTFGYN 59
Chick           -MAELPCAEP---LPRCSGRFTISTLLGAEEA---AGCEGTQLS----GSSLCTRTFGYN 49
                                                      .        .:        **:

Takifugu        TLDSPPNYDFYANTEVWGRQRHFRPSLFQLYAQPEDD---TRPPMYEETTGEQGISGDSS 110
Zebrafish       TLDAPPSYDFYTNTEVFGRAKKSRPSLFELHSNPQDDP--SPPPLYEESSVDK-----QS 70
S0188I22        TLDVGPNYDFYANTNAQGRVRRSRPSLFQLHSNIEEDS--CPPPLYEETNTGR----APG 72
Rat             TIDVVPAYEHYANSALPGEPRKVRPTLADLHSFLKQEGSHLHALAFDGRPGHELTDGLVE 117
Mouse           TIDVVPAYEHYANSALPGEPRKVRPTLADLHSFLKQEGSHLHALAFDGRQGRELTDGLVE 117
Human           TIDVVPTYEHYANSTQPGEPRKVRPTLADLHSFLKQEGRHLHALAFDSRPSHEMTDGLVE 119
Chick           TVDVVPAYEHYANSRGVGEAGRGRPSLADLHSILKPDPGHLRVPLPDPQRSNGLPD--AE 107
                *:*  * *:.*:*:   *.  : **:* :*::  : :         :

Takifugu        CDDEEEQKEPPPEPTRFGWIQGVMIRCMLNIWGVILYLRLPWITAQAGIGMTWVIILLSS 170
Zebrafish       PEDEDEPTEPPPEPARFGWAQGVMIRCMLNIWGVILYLRLPWITAQAGIGLTWIIILVSS 130
S0188I22        DSSEEDVEEPQSEPTRFGWVKGVMVRCMLNIWGVILYLRLPWITAQAGIGLTWVIILLSS 132
Rat             DETGANSEKSPGEPVRFGWVKGVMIRCMLNIWGVILYLRLPWITAQAGIVLTWLIILLSV 177
Mouse           DETGTNSEKSPGEPVRFGWVKGVMIRCMLNIWGVILYLRLPWITAQAGIVLTWLIILLSV 177
Human           GEAGTSSEKNPEEPVRFGWVKGVMIRCMLNIWGVILYLRLPWITAQAGIVLTWIIILLSV 179
Chick           EGAGEPSSAPAAEPVRFGWVKGVMIRSMLNIWGVILYMRLPWITAQAGIALTWLIILMSV 167
                            **.**** :***:*.**********:*********** :**:***:*

Takifugu        CITGITGLSTSAIATNGKVKGGGTYFLISRSLGPELGGSIGLIFAFANAVAVAMHTVGFA 230
Zebrafish       SITGITGLSTSAIATNGKVKGGGTYFLISRSLGPELGGSIGLIFAFANAVAVAMHTVGFA 190
S0188I22        CITGITGLSTSAIATNGRVKGGGTYFLISRSLGPKLGGSIGLIFAFANAVAVAMHTVGFA 192
Rat             MVTSITGLSISAISTNGKVKSGGTYFLISRSLGPELGGSIGLIFAFANAVGVAMHTVGFA 237
Mouse           MVTSITGLSISAISTNGKVKSGGTYFLISRSLGPELGGSIGLIFAFANAVGVAMHTVGFA 237
Human           TVTSITGLSISAISTNGKVKSGGTYFLISRSLGPELGGSIGLIFAFANAVGVAMHTVGFA 239
Chick           TVTTITGLSISAISTNGKVKSGGTYFLISRSLGPELGGSIGLIFAFANAVAVAMHTVGFA 227
                 :* ***** ***:***:**.*************:***************.*********

Takifugu        ETVTDLMRENGVSMVDRTNDIRIIGIITVTCLLGISMAGMAWESKAQVLFFLVIIVSFAS 290
Zebrafish       ETVQVLMQETEVSMVDKLNDIRIIGVITVTCLLAISMAGMEWESKAQVLFFFVIMISFAS 250
S0188I22        ETVQALMEESGASIVDPINDIRIIGVITVTCLLAISLAGMEWEAKAQVLFFFVILVSFAN 252
Rat             ETVRDLLQENGTPIVDPINDIRIIGVVTVTVLLAISLAGMEWESKAQVLFFLVIMVSFAN 297
Mouse           ETVRDLLQEYGTPIVDPINDIRIIGVVTVTVLLAISLAGMEWESKAQVLFFLVIMVSFAN 297
Human           ETVRDLLQEYGAPIVDPINDIRIIAVVSVTVLLAISLAGMEWESKAQVLFFLVIMVSFAN 299
Chick           ETVRDLLQEHNSLIVDPTNDIRIIGVLTVTVLLGISLAGMEWEAKAQILFFLVILVSFIN 287
                ***  *:.*    :**  ******.:::** **.**:*** **:***:***:**::** .

Takifugu        YIVGTAIPATPQKQAKGFFSYKADIFATNFVPNWRGEQGSFFGMFSIFFPSATGILAGAN 350
Zebrafish       YIIGTIIPATPQKQARGFFSYRADIFATNFVPGWRGPEGSFFGMFSIFFPSATGILAGAN 310
S0188I22        YIVGTIIPATPQKQAKGFFSYQADIFAENFVPGWRGPEGNFFGMFSIFFPSATGILAGAN 312
Rat             YLVGTLIPASKDKASKGFYSYHGDIFVQNLVPDWRGIDGSFFGMFSIFFPSATGILAGAN 357
Mouse           YLVGTLIPASEDKASKGFYSYHGDIFVQNLVPDWRGIDGSFFGMFSIFFPSATGILAGAN 357
Human           YLVGTLIPPSEDKASKGFFSYRADIFVQNLVPDWRGPDGTFFGMFSIFFPSATGILAGAN 359
Chick           YLVGTVIPASAEKQAKGFFSYRADIFAQNFVPDWRGPEGSFFGLFSIFFPSATGILAGAN 347
                *::** **.: :* ::**:**:.***. *:**.*** :*.***:****************

Takifugu        ISGDLKN--PAVAIPRGTLLAIFCTTVSYIIISATIGACVVRDASGLLNDSLSVTASPES 408
Zebrafish       ISGDLKAR-PNVAIPRGTMLAIFWTTVSYLIISATIGSCVVRDASGDVNDTISSLTG--E 367
S0188I22        ISGDLKAGDPTVAIPRGTLMAIFWTTLSYLIIAATIGACMVRDASGIMNDTLVMSSAD-S 371
Rat             ISGDLKD--PAVAIPKGTLMAIFWTTISYLAISATIGSCVVRDASGDVNDTVTPGPGL-- 413
Mouse           ISGDLKD--PAVAIPKGTLMAIFWTTISYLAISATIGSCVVRDASGDVNDTMTPGPGP   413
Human           ISGDLKD--PAIAIPKGTLMAIFWTTISYLAISATIGSCVVRDASGVLNDTVTPGWGA   415
Chick           ISGDLKD--PALAIPKGTLMAIFWTTVSYLVLSATIGACVLRDASGSLNDSVAVGSPG-- 403
                ******   * :***:**::*** **:**: ::****:*::***** :**::


Takifugu        CTGFACHYGWDFSECTNNKSCTYGISNYYQSMGLVSAFAPLITAGIFGATLSSALACLVS 468
Zebrafish       CLGVGCNYGWNFTDCMTNNTCTYGLSNYYQSMSMVSAVAPLITAGIFGATLSSALACLVS 427
S0188I22        CQGLACQYGWDFTNCITNRTCTYGLSNQYQSMSLVSGFAPLITAGIFGATLSSALACLVS 431
Rat             CEGLACGYGWNFTECSQQHSCRYGLINYYQTMSMVSAFAPLITAGIFGATLSSALACLVS 473
Mouse           CEGLACGYGWNFTECSQQRSCRYGLINYYQTMSMVSAFAPLITAGIFGATLSSALACLVS 473
Human           CEGLACSYGWNFTECTQQHSCHYGLINYYQTMSMVSGFAPLITAGIFGATLSSALACLVS 475
Chick           CEGLGCSYGWNFTDCAQRQSCRYGLSNYYQSMSMVSGFGPLITAGIFGATLSSALACLVS 463
                * *..* ***:*::*  ..:* **: * **:*.:**...*********************
```

```
Takifugu    APKVFQCLCKDKLYPFIGFFGKGYGRNNEPIRSYILAYFIAACFILIAELNTIAPIISNF 528
Zebrafish   APKVFQCLCKDKLYPGIGFFGKGYGKNNEPLRSYLLAYIIAICFILIAELNTIAPIISNF 487
S0188I22    APKVFQCLCKDNLYPVIGFFGKGNGKNDEPIRGYVLAYIIAVCFVLIAELNTIAPIISNF 491
Rat         AAKVFQCLCEDQLYPLIGFFGKGYGKNKEPVRGYLLAYAIAVAFIIIAELNTIAPIISNF 533
Mouse       AAKVFQCLCEDQLYPLIGFFGKGYGKNREPVRGYLLAYAIAVAFIIIAELNTIAPIISNF 533
Human       AAKVFQCLCEDQLYPLIGFFGKGYGKNKEPVRGYLLAYAIAVAFIIIAELNTIAPIISNF 535
Chick       APKVFQCLCKDQLYPLIGFFGKGYGKNSEPIRGYMLTYAIAIGFILIAELNAIAPIISNF 523
            *.*******:*:*** ******* *:* **:*.*:*:* **  *::*****:********

Takifugu    FLCSYSLINFSCFHASITNSPGWRPSFKLYNKWLSLLGAVCCVVIMFLLTWWAALIAFGV 588
Zebrafish   FLCSYALINFSCFHASITNSPGWRPTFRFYSKWLSLLGAVVSVIIMFLLTWWAALIAIGI 547
S0188I22    FLCSYALINFSCFHASITNSPGWRPSFRFYSKWMSLLGAVVSVIIMFLLTWWAALIAIGI 551
Rat         FLCSYALINFSCFHASITNSPGWRPSFRYYSKWAALFGAVISVVIMFLLTWWAALIAIGV 593
Mouse       FLCSYALINFSCFHASITNSPGWRPSFRYYSKWAALFGAVISVVIMFLLTWWAALIAIGV 593
Human       FLCSYALINFSCFHASITNSPGWRPSFQYYNKWAALFGAIISVVIMFLLTWWAALIAIGV 595
Chick       FLCSYALINFSCFHASITNSPGWRPSFRYYSKWAALFGATISVVIMFLLTWWAALIALGI 583
            *****:*******************:*: •.** :*:**  .*:*************:*:

Takifugu    VLILLSYTLYKKPDVNWGSSVQAGSYNIALNQCVALNHVEDHVKNYRPQCLVLTGAPGSR 648
Zebrafish   VIFLLGYVLYKKPEVNWGSSMQASSYNMALSQCVGLNQVEDHIKNYRPQCLVLSGPPCAR 607
S0188I22    VIFLLGYVLYKRPTVNWGSSVQAGSYNMALSYCVSLNQVDDHIKNYRPQCLVLTGPPSCR 611
Rat         VLFLLLYVIYKKPEVNWGSSVQAGSYNLALSYSVGLNEVEDHIKNYRPQCLVLTGPPNFR 653
Mouse       VLFLLLYVIYKKPEVNWGSSVQAGSYNLALSYSVGLNEVEDHIKNYRPQCLVLTGPPNFR 653
Human       VLFLLLYVIYKKPEVNWGSSVQAGSYNLALSYSVGLNEVEDHIKNYRPQCLVLTGPPNFR 655
Chick       VIFLLGYVLYKKPDVNWGSSMQASSYNLALSYSVGLSEVDEHIKNYRPQCLVLTGPPNFR 643
            *::** *.:**:* ******:**.***:**. .*.*..*::*:**********:*.*  *

Takifugu    PALVDLAACFTKYLSLMMCGNVITE-EPSPSAVEKASGKTHVTWLNQRKVKSFYRGVVAP 707
Zebrafish   PSLVDFIGAFTKNQSLMICANVLAS-GPSPGTADSMS-STHLKWLNNRKIKSFYHTVVAD 665
S0188I22    PALVDFVGTFTKNLSLMICGNVVTG-GPSPATLDTASSNSHVTWLNKRQIKSFYHGVVAN 670
Rat         PALVDFVSTFTQNLSLMICGHVLIAPGKQRVPELRLIASGHTKWLNKRKIKAFYSDVIAE 713
Mouse       PALVDFVSTFTQNLSLMICGHVLIGPGKQRVPELRLIASGHTKWLNKRKIKAFYSDVIAE 713
Human       PALVDFVGTFTRNLSLMICGHVLIGPHKQRMPELQLIANGHTKWLNKRKIKAFYSDVIAE 715
Chick       PALVDFVGTFTKNLSLMICGNVLIGPSKQKVLEARQASDGHTRWLLKRKIKAFYTNVLAE 703
            *:***: . **:  ***:*.:*:     .         . *  ** :*::*:**  *:*

Takifugu    ELRSGVNVLLQGAGLGRLKPNVLLMGFKSDWRSDSPCAAHSYIGILQDAFDLQYGVCILR 767
Zebrafish   DLRTGVQMLLQSTGLGRMKPNVLVMGYKKNWRKVQPGIIENYVGILHDAFDLQYGVCVLR 725
S0188I22    DLRTGVKMLLQGAGLGRIRPNVLLTGFKRDWRKDKPSCIDNYIGILHDAFDLQYGVCVLR 730
Rat         DLRSGVQILMQASGLGRMKPNILVVGFKKNWQSAHPATLEDYIGILHDAFDFNYGVCIMR 773
Mouse       DLRSGVQILMQASGLGRMKPNILVVGFKRNWQSAHPATVEDYIGVLHDAFDFNYGVCVMR 773
Human       DLRRGVQILMQAAGLGRMKPNILVVGFKKNWQSAHPATVEDYIGILHDAFDFNYGVCVMR 775
Chick       DLRSGVQMLLQAAGLGKMRPNIVTLGYKRDWQAAAPQSLEDYVGILHDAFDFKHGVCLLR 763
            :** **::*:*.:***:::**::  *:* :*:   *   ..*:*:*:****:::***::*

Takifugu    TKEGLDVSRPSQSHS---------ESFLLDFSAPVS---FFSDNEDTTVLAPQP--ITVF 813
Zebrafish   MKEGLDITRTIQAQVNLGFETSTEQGLDTNSTAPTSPTIEASLDPETLMALTQP--STLF 783
S0188I22    MREGLDMFRQAQTHVNPGFESS--PERGVNTCVPPAPPANFSLDPDSMVTEPQPQPSTVF 788
Rat         MREGLNVSEALQTHT-----------------------------APEALVQ-EEQTSTIF 803
Mouse       MREGLNVSEALQTHT-----------------------------TPEALIQ-EEQASTIF 803
Human       MREGLNVSKMMQAHINPVFDPA---EDGKEASARGARPSVSGALDPKALVK-EEQATTIF 831
Chick       LREGLNVSRVPQAHINPAFGAA---EHPDGNGAGG-------RAAPSTSIDPEQQASTIF 813
             :***:: .  *::                                      :    *:*

Takifugu    QKKQGKKTIDVYWLSDDGGLTLLLPYLLTRRKRWARCKVRVFVGGEVEKKETRKEEVVAL 873
Zebrafish   QTRQGKKTIDVYWLSDDGGLTLLIPYLLTRKKRWGRCKVRVFVGGEAQQIEEQKKELKGL 843
S0188I22    QSQQGKKTIDVYWLFDDGGLTLLLPYLLTRKKRWARCKVRVFVGGDIQRKEEQRKEVMDL 848
Rat         QSEQGKKTIDIYWLFDDGGLTLLIPYLLHRKKRWGKCKIRVFVGGQINRMDEERKAIISL 863
Mouse       QSEQGKKTIDIYWLFDDGGLTLLIPYLLHRKKRWGKCKIRVFVGGQINRMDEERKAIISL 863
Human       QSEQGKKTIDIYWLFDDGGLTLLIPYLLGRKRRWSKCKIRVFVGGQINRMDQERKAIISL 891
Chick       QSQQGKKTIDIYWLFDDGGLTLLIPYLLGRKKRWGKCKVRVFVGGQINRMDEERKAIVSL 873
            *..*******:*** ********:**** *::**.:**:******: :: : .:: :  *

Takifugu    IKKFRLGFRDVEVLPDVYQSPQPANIQRFENMLSDFRIVTNPKQDTEAELPRQQEEPWMI 933
Zebrafish   ISRFRLGFKDIQVLPDINGAPQSEHIRKFEDFIAPYRVSSVQKDGQEADEAT-KEFSWMV 902
S0188I22    ISKFRLGFHDVEVLPDINARPQPEHVRRFEDLIGPYRLNTAQKDGHVTAEQLNQDCPWMV 908
Rat         LSKFRLGFHEVHVLPDINQKPQVEHTKRFEDMIAPFRLNDGFKDEATVAEMR-RDYPWKI 922
Mouse       LSKFRLGFHEVHVLPDINQKPQAEHTKRFEDMIAPFRLNDGFKDEATVTEMR-RDCPWKI 922
Human       LSKFRLGFHEVHILPDINQNPRAEHTKRFEDMIAPFRLNDGFKDEATVNEMR-RDCPWKI 950
Chick       LSKFRLGFHEVHILPDINQQPRPEHIRRFDELIAPFRLNDGFKDEAAVNELR-HGCPWKI 932
            :.:*****:::.:***:   *:  : ::*::::. :*:    *:   .     :  .* :

Takifugu    NDQDLEKNKSKSLRQIRLNEVLHDYSRDAALIIITMPVGRRGVCPSTLYLAWLDFLSHDL 993
Zebrafish   SDEEMETFKAKSLRQIRLNEVIQDYSRDAALIVVTMPVGRRGSCPSPLYMAWLEIVSRDL 962
S0188I22    SDEEIETNKPKTLRQIRLNEVLQDYSRDAAIIFVTMPVGRRGQCPSALYMAWLETLSRDL 968
Rat         SDEEINKNRIKSLRQVRLNEILLDYSRDAALIILTLPIGRKGKCPSSLYMAWLETLSQDL 982
Mouse       SDEEINKNRIKSLRQVRLSEILLDYSRDAALIILTLPIGRKGKCPSSLYMAWLETLSQDL 982
Human       SDEEITKNRVKSLRQVRLNEIVLDYSRDAALIVITLPIGRKGKCPSSLYMAWLETLSQDL 1010
Chick       SDEEVHRHRAKSLRQVRLNEILLDYSRDAALIAITLPIGRKERCPSSLYMAWLETLSQDL 992
            .*:::   : *:***:**.*:: *******:* :*:*:**:  ***.**:***: :*:**
```

```
Takifugu        RPPVLLVRGNQENVLTFYCQ 1013
Zebrafish       RPPVLLVRGNQENVLTQYCQ 982
S0188I22        RPPVLLVRGNQENVLTFYCQ 988
Rat             SPPVLLIRGNQENVLTFYCQ 1002
Mouse           RPPVLLIRGNQENVLTFYCQ 1002
Human           RPPVILIRGNQENVLTFYCQ 1030
Chick           RPPVILIRGNQENVLTFYCQ 1012
                 ***:*:********* ***
```

3.8 Phylogenetic analysis

Dead eye, Lin10, and GNAO1 coding sequences from the Atlantic salmon duplicate genes as well as orthologous gene sequences from other species (zebrafish, pufferfish, human, mouse, rat, and chick when available) were used to estimate the divergence time of the Atlantic salmon duplicate genes, and therefore to approximate the time of the salmonid genome duplication event. The pairwise number of nucleotide and amino acid changes was computed between each pair of orthologs. The divergence times of these organisms are quite well known from fossil records and previous phylogenetic analyses and are assumed to be as follows: tetrapods and fish - 450mya; birds and mammals - 310mya; rodent and humans – 110mya; mouse and rat – 40mya; zebrafish, pufferfish and salmon – 200mya (Nei and Glazko, 2002; Kumar and Hedges, 1998; Van de Peer, 2004). It is therefore possible to plot the number of nucleotide or amino acid changes between orthologous sequences versus the time since divergence. The straight line connecting the data points allows for an estimate of the divergence time of the Atlantic salmon duplicate genes based on the number of changes that are observed between the duplicate sequences (Figure 31 and Figure 32).

In this way, the divergence time of the Atlantic salmon dead eye duplicates was estimated to be 91 and 89 mya based on the number of nucleotide and amino acid changes, respectively. Similarly, for Lin10, the divergence time was estimated to be 87 and 115mya. Finally, the GNAO1 nucleotide and amino acid data placed the duplication time at 76mya and 118mya, respectively. Taken together, these results suggest the divergence of these duplicate gene pairs occurred between 80 and 120 mya approximately. This estimation is consistent with the more ancient extremity of the predicted range of the tetraploidization event (20-120mya) (Allendorf and Thorgaard, 1984).

Figure 31 Number of nucleotide changes versus divergence time. Black arrows indicate known divergence times (Mouse/Rat 40mya; Human/Rodent 110mya; Fish/Fish (zebrafish, *Takifugu*, Atlantic salmon) 200mya; Bird/Mammal 310mya; Fish/Tetrapod 450mya). Red arrows represent the estimated divergence times of the three Atlantic salmon duplicate genes.

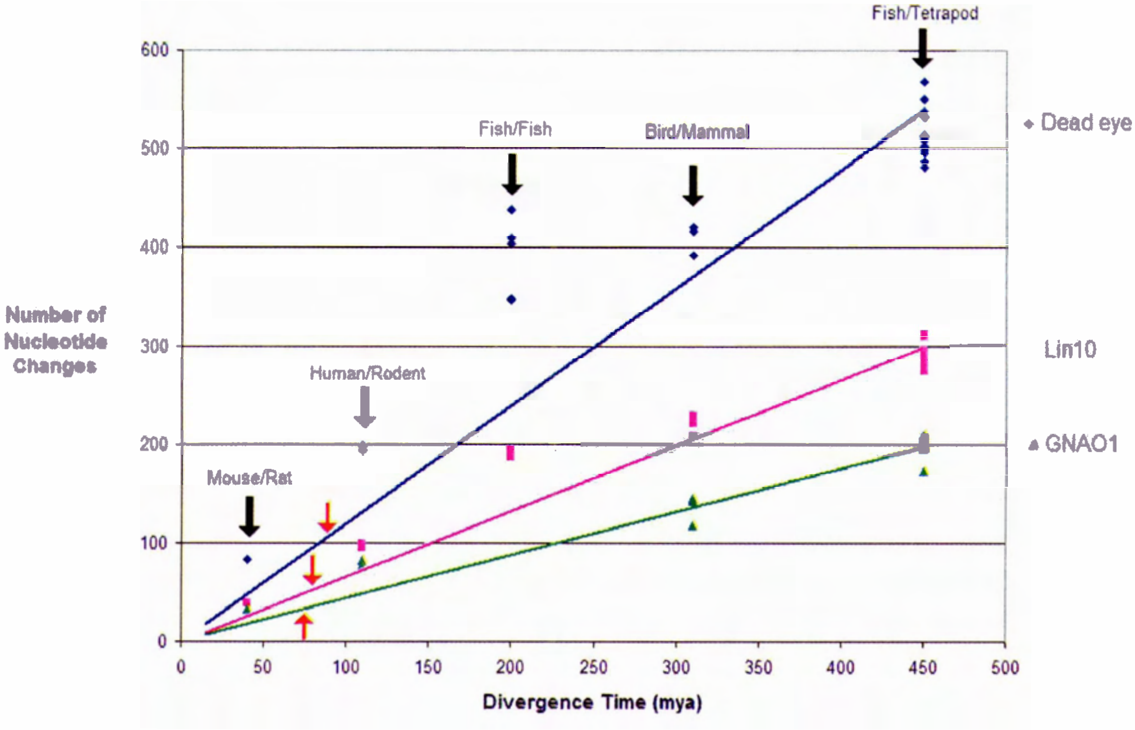
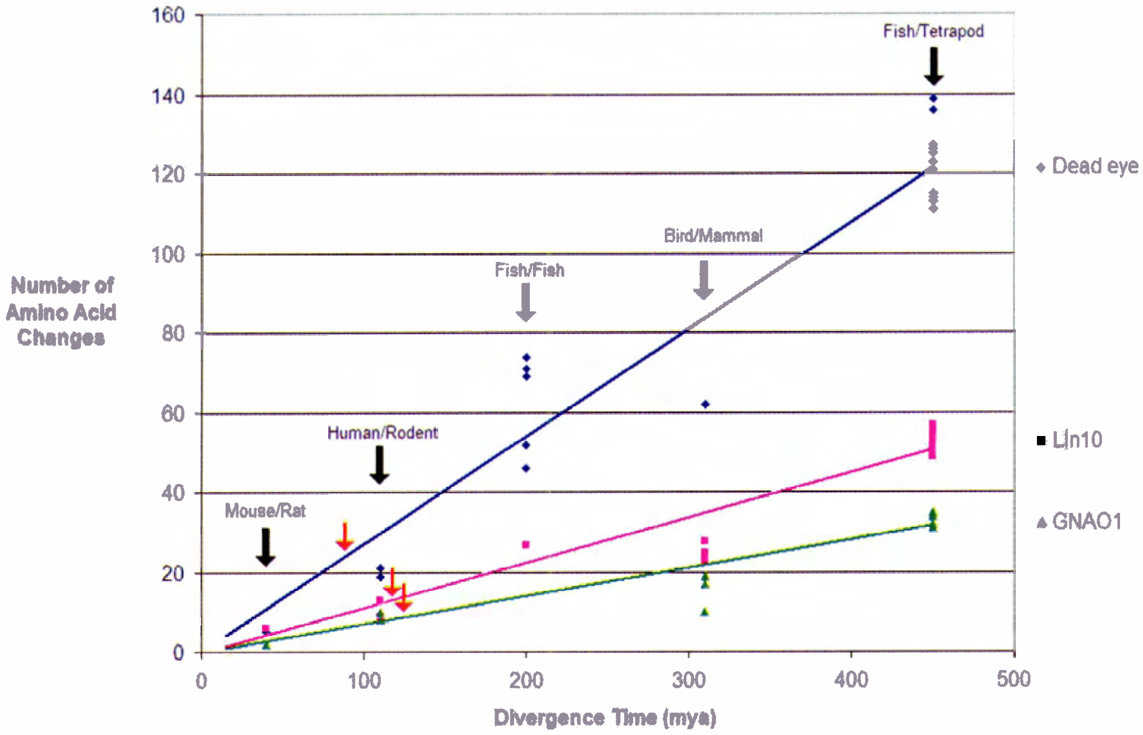


Figure 32 Number of amino acid changes versus divergence time. Black arrows indicate known divergence times (Mouse/Rat 40mya; Human/Rodent 110mya; Fish/Fish (zebrafish, *Takifugu*, Atlantic salmon) 200mya; Bird/Mammal 310mya; Fish/Tetrapod 450mya). Red arrows represent the estimated divergence times of the three Atlantic salmon duplicate genes.



CHAPTER 4 DISCUSSION

This study was designed to compare homeologous regions of the genome of Atlantic salmon to increase our knowledge of how a genome reorganizes itself to cope with duplicated segments of DNA following a tetraploidization event, and furthermore to study the importance of gene duplications for evolution and adaptation. Homeologous regions of the Atlantic salmon genome were isolated by identifying clones within a BAC library that contained the metallothionein gene duplicates (MetA and MetB), that are believed to have arisen via the genome wide duplication event in the ancestor common to all salmonids. A representative MetA and MetB BAC clone were then shotgun subcloned and sequenced to provide 170kb of overlapping, homeologous sequence data.

4.1 Metallothionein and the Salmonid Tetraploidization Event

Based on the sequences that appear in GenBank, all salmonids studied to date, including Artic charr, rainbow trout, and Atlantic salmon, possess at least two copies of the metallothionein gene that can be distinguished from one another based on the length of the coding sequence (MetA is 3 nucleotides longer than MetB in exon two) as well as the length of intron two (MetB has a much longer second intron relative to MetA). Since these features are common to the MetA and MetB sequences for several salmonids, it suggests that the duplication and subsequent divergence of MetA and MetB occurred before the salmonid radiation. Furthermore, only a single metallothionein locus is found in the genomes of many other teleost fish, including the Northern pike (*Esox lucius*), the closest known relative to the salmonids (Ishiguro et al., 2003).

Because the common ancestor of all salmonids is known to have undergone a tetraploidization event (Allendorf and Thorgaard, 1984), it is likely that the duplicate metallothionein genes in salmonids arose through this genome duplication event. To test this hypothesis, MetA and MetB sequences were obtained for several additional salmonid species as well as the Northern Pike. The sequence data included intron 1, exon 2, and intron 3 of the metallothionein genes and therefore highlighted the major sequence differences between the duplicates. Phylogenetic analysis reveals two salmonid metallothionein gene clusters, one including all of the MetA sequences, and a second that includes all of the MetB sequences (Figure 3). The single Northern pike metallothionein sequence appears as an outgroup to these two salmonid groups. This topology supports an early duplication event for the metallothionein genes, consistent with and suggestive of a whole genome duplication event.

Notably, the genome of one salmonid species, the Arctic grayling (*Thymallus arcticus*), was found to possess three copies of the metallothionein gene – one MetA gene, and two MetB genes (Figure 3). The most parsimonious explanation is that the duplicate MetB genes arose via a grayling-specific duplication event. It would be informative to obtain genomic DNA from another member of the *Thymallidae* family, European grayling (*Thymallus thymallus*) for instance, and determine whether the MetB gene duplication is common to the entire *Thymallidae* family or specific to the Arctic grayling lineage. It is suspected that the MetB duplicate gene arose via a tandem duplication.

Metallothionein tandem duplication events are fairly common in vertebrate genomes. At the human metallothionein locus on chromosome 16 for example, tandem duplications have resulted in the accumulation of fourteen adjacent metallothionein genes. Similarly, mouse and rat each possess four tandemly arrayed copies of metallothionein genes. Additionally, metallothionein tandem duplications have occurred in the genomes of teleost fish besides the Arctic grayling, as seen in the completely sequenced genomes of the pufferfish *Takifugu rubripes* and *Tetraodon nigroviridus,* where two metallothionein genes lie adjacent to one another in both genomes. Furthermore, the genomes of eight species of Antarctic fish belonging to the order *Notothenioidei* have also been found to contain duplicate metallothionein genes, where phylogenetic analysis reveals the gene copies to be the result of a duplication event distinct from the salmonid tetraploidization event (Bargelloni et al., 1999).

In teleost fish, the most parsimonious explanation for the evolution of the metallothionein locus is as follows: (1) following the teleost specific genome-wide duplication event that occurred at the base of the teleost lineage (Jaillon et al., 2004), one of the two duplicate metallothionein loci was lost prior to the teleost radiation; (2) the salmonid common ancestor underwent the salmonid-specific tetraploidization event and during the process of re-diploidization, the MetA and MetB sequences diverged prior to the salmonid radiation; (3) independent metallothionein tandem duplication events have occurred along various lineages, possibly providing a selective advantage to the species in which the duplications exist.

A duplication event that is followed by retention of both duplicate copies can be indicative of a selective advantage associated with the increase in gene dosage (Kondrashov et al., 2002). In this situation, purifying selection can work to maintain both gene sequences, and hence allow for the production of more protein (Kondrashov et al., 2002). In the case of metallothionein, duplications of the metallothionein gene have been observed in strains of *Drosophila* and yeast that have been grown in the presence of high concentration of heavy metals (Maroni et al., 1987; Tohoyama et al., 1996). The critical role of metallothionein proteins in metal detoxification provides a logical explanation as to why duplicate metallothionein genes might have been selected for and retained in organisms grown in the presence of heavy metals. Other organisms such as fish that may be exposed to high levels of dissolved heavy metals would also appear to benefit from the presence of multiple copies of the metallothionein gene.

Following a tandem gene duplication, the resulting duplicate copy is more likely to initially retain its function if the entire regulatory region associated with the gene is also duplicated. If the duplication does not happen to include the proper promoter elements the gene will be non-functional, unless the new location allows the duplicate gene to fall under the control of alternate promoter elements. The regulatory region of metallothionein has been well characterized in many taxa, and is dominated by multiple, highly conserved cis-acting DNA elements, termed metal responsive elements (MREs). In response to the presence of various heavy metals, mitogens, cytokines and/or oxidative stress, various transcription factors bind to the MRE sequences leading to the induction of metallothionein gene expression (Coyle et al., 2002). In the human MT-IIA gene, for instance, there are seven MRE elements oriented in either direction, located between -300 bp and the transcriptional start site (Ogra et al., 2001).

In fish, it has been suggested that metallothionein can act as a biomarker for metal contamination in aquatic environments (Lin et al., 2004), and therefore the gene and associated regulatory region have been extensively characterized in numerous species. In general, the promoter region is composed of two clusters of MRE elements that are located proximal and distal to the transcriptional start site. In zebrafish, for instance, the metallothionein gene has four MREs, with the first three located within 250 bp of the transcriptional start site, and the fourth further upstream at -835 bp (Yan and Chan, 2002). Similarly, carp has four MREs that exhibit the same spatial arrangement within the promoter region as zebrafish, but the most distal MRE is only 530 bp from the transcriptional start site (Chan et al., 2004). In both of these studies, the promoter regions were shown to be functional in metal-induced transcriptional regulation, and most strongly upregulate gene expression by induction with zinc (Yan and Chan, 2002; Chan et al., 2004). The ayu or sweetfish (*Plecoglossus altivelus*), on the other hand, has eight MREs located as far upstream as -1185bp (Lin et al., 2004).

Of the three studies that characterized the metallothionein gene promoter region, the work by Lin et al. (2004) in the ayu is likely the most robust, as 2500bp of DNA upstream of the transcriptional start site was sequenced and subsequently examined for MRE consensus sequences, whereas in zebrafish and carp, less than 1000bp were studied. In fact, by utilizing the zebrafish genomic sequence data, it is possible to look further upstream in of the metallothionein gene for additional MREs. Examining 1500 bp upstream of the transcriptional start site does indeed reveal the presence of additional MRE sequences, including two more in the distal region (around -850bp) and two further MREs around -1375 bp. Functional studies need to be performed to determine whether these MRE consensus sequences actually function in gene regulation.

With respect to salmonid fish, only in rainbow trout have the regulatory regions of the metallothionein gene duplicates been characterized. Rainbow trout MetA has had its promoter region characterized up to -1052 bp, resulting in the identification of six MREs that are organized into proximal and distal clusters, with the proximal region containing two MREs located close to the TATA box, and the distal region consisting of four MREs lying 576 to 760 bp upstream from the start of transcription (Olsson, 1995). The MetB gene has four MREs, with two located proximally (-100 and -60bp) and the other two are located distally (-680 and -570bp) (Samson et al., 2002). Functional assays have identified the importance of the MetA and MetB MREs in metal-induced transcriptional activation of genes (Olsson et al., 1995; Samson et al., 2002).

From the BAC seqeunces, the promoter regions of the Atlantic salmon MetA and MetB genes were determined for the first time. Because the full BAC sequences were available, there was no lack of sequence data available upstream of the metallothionein transcriptional start sites; based on the studies mentioned above, 2000 bp of DNA before the start site of transcription was analyzed for the presence of MREs. In this way, five MRE consensus sequences (MREa-e) were identified in each Atlantic salmon metallothionein gene (Figure 21), once again displaying proximal (MREa and MREb) and distal (MREc-e) groupings similar to other teleosts. Between the Atlantic salmon duplicate loci, MREa to MREd align quite closely and are therefore predicted to be homologous. The putative MREe identified in the distal region of MetB at -1467 bp may be a bit far away to actually be functional, although the most distal MRE in the ayu is located at -1185bp and appears to be part of gene regulation based on deletion analysis of the promoter region (Lin et al., 2004). It would be informative to perform promoter deletion studies of the Atlantic salmon MetA and MetB regulatory regions to understand the importance of each MRE.

4.2 Annotation of Homeologous Regions of the Atlantic salmon Genome

In addition to providing the complete MetA and MetB promoter sequences in Atlantic salmon, this BAC sequencing project allows for an exploration into the evolutionary genomic re-organization steps that have occurred in these homeologous regions since the tetraploidization event. Various sequence features in addition to gene identification and analysis were examined, including G+C content and the presence of repetitive DNA.

4.2.1 G+C Content

A striking feature of mammalian genomes is that they exhibit large-scale variation in base composition such that there are regions of high and low G+C content. The isochore structures exist over segments of hundreds of kilobases to megabases and reflect large regions of the genome that contain local similarities in base content (Eyre-Walker and Hurst, 2001). Gene density, SINE density, and recombination frequency are all higher in G+C rich regions of mammalian genomes; conversely, low G+C regions are gene poor and associated with increased LINE density in mammals (Bernardi, 2000).

The genomes of cold-blooded vertebrates contrast their warm-blooded counterparts in that the isochore structure is not as striking, although gene rich regions still exist in the G+C richest parts of their genomes (Bernardi, 2000). In the recently completed *Tetraodon nigroviridus* genome sequence, for instance, although the overall G+C content is quite high (49.3%), gene-rich and gene-poor regions exhibit the same associations with high and low G+C regions, respectively, as seen in mammals (Jaillon et al., 2004). Another interesting finding within the *Tetraodon* sequence is that although repetitive elements are not numerous, LINES preferentially occupy G+C rich regions of the genome and SINES are found in A+T rich regions, precisely the opposite of the trend observed in mammals (Jaillon et al., 2004). The explanation for this is unclear.

Although the *Tetraodon* genome exhibits an isochore-like structure, the *Takifugu rubripes*, a closely related pufferfish, displays a fairly uniform G+C content throughout its genome (Aparicio et al., 2002), an observation that Jabbari and Bernardi (2004a) explain with the thermodynamic stability hypothesis: warm-blooded animals (or *Tetraodon*, which inhabits tropical waters) have evolved a genome core (G+C rich region where most genes reside) that exists in an open state throughout interphase. This region is selected to be G+C rich in order to provide stability within the warm-blooded animal cells. On the other hand, the gene poor region is packed into chromatin during interphase, and a high G+C content is unnecessary for stability. Therefore, cold water-dwelling poikilotherms such as *Takifugu* have not had the same G+C mosaicism evolve within their genomes.

Following this logic, it would be expected that Atlantic salmon, living in cold water, would have a genome characterized by lower intermolecular compositional heterogeneities, similar to *Takifugu*. The length of sequence data provided by this study, however, is not sufficient to make firm conclusions about the overall heterogeneity of G+C content in the Atlantic salmon genome. The homeologous regions in this study exhibit approximately the same gene density (~1 gene/20kb) as the "gene-rich" regions of the human genome, but the overall G+C content (~43%) is only equal to the overall G+C content in the human genome and does not reflect the G+C content of gene-rich regions in human. Whether this means that the Atlantic salmon genome has gene rich regions that do not exhibit overall higher G+C content values is unclear. On the other hand, the average G+C content of zebrafish is only 36% (Jabbari and Bernardi, 2004b), and likely better reflects the expected average G+C content across the entire Atlantic salmon genome. If this is the case, it would suggest that these Atlantic salmon BACs represent regions of the genome that are greatly enriched in G+C content, and therefore suggest the presence of isochore-like structures within the Atlantic salmon genome. Additional BAC sequence data must be examined to test for a correlation within the genome of the Atlantic salmon between gene rich and gene poor regions existing in segments of DNA with high and low G+C content, respectively.

In terms of the G+C content within genes, and more specifically the G+C content of introns relative to exons, a correlation has been found between higher G+C content within exons, and lower G+C content within introns (Bernardi, 2000). This difference is especially striking in teleost fish, where Winnard et al. (2002) found that most teleost introns exhibit a higher A+T content (up to 30%) than that of the flanking exons. In the Atlantic salmon BAC sequence data obtained during this study, this correlation was also found to be true (Table 5). On average, the exonic portions of the genes that were identified in the BACs were enriched in G+C content by approximately 12% relative to the introns.

The presence of a pseudogene, such as the BBS2 locus in the S0188I22 sequence, provides an opportunity to examine the evolution of genic G+C content following a pseudogenization event. In comparison to the SLC12A3 pseudogene in the S0085O16 sequence data, where all of the introns and exons are no longer even distinguishable, the remnants of these boundaries are still clear in the BBS2 pseudogene, and therefore the average exon and intron G+C content was calculated. It is expected that after the initial deleterious mutation, the DNA at this locus began to evolve neutrally, as no constraints were being imposed by natural selection. In as much, the G+C content of the "exons" in the pseudogene is slightly lower when compared to the exons at the functional locus (50.4% compared to 51.5%). On the other hand, the G+C content of the "introns" at the pseudogene locus has increased to 43.1% from 38.6% at the functional BBS2 locus. These observations suggest that the DNA in the BBS2 pseudogene locus is evolving in such a way as to return to the average for the overall G+C content within this region.

4.2.2 Repetitive DNA

Eukaryotic genomes contain substantially differing amounts of repetitive DNA due to the differential propagation and deletion of selfish genetic elements (Eichler and Sankoff, 2003). Human, for instance, contains 46% repetitive DNA, mouse 38%, *Takifugu* less than 10% (Eichler and Sankoff, 2003). Some repetitive elements are distinguished by their mode of propagation; long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE), and retrovirus-like elements with long terminal repeats (LTR), propagate by reverse-transcription of an RNA intermediate, whereas DNA transposons move by a direct “cut-and-paste” mechanism. Additional repetitive elements, including microsatellites (tandemly repeated stretches of 2 to 6 nucleotides) and minisatellites (tandemly repeated stretches of 20-50 nucleotides) may also play a major role in genome evolution due to expansion and contraction of the repeating unit. Lineage specific repetitive DNA is commonly observed, such that even closely related species have varying amounts and types of repetitive elements.

Based on hybridization experiments in our laboratory, it was expected that the Atlantic salmon genome would contain a large fraction of repetitive elements, and furthermore, that many of those elements would be specific to either the salmonids or Atlantic salmon alone. In approximately 30% of hybridization experiments, probes designed within the flanking regions of microsatellite sequences hybridize non-specifically to a large portion of the clones in the BAC library (data not shown), which is indicative of the probe sequence containing some form of repetitive element. These repetitive elements are likely to be specific to Atlantic salmon or possibly even all salmonids because the probe sequences are not significantly similar to any repetitive elements that have been identified from whole genome sequences of other organisms. Based on the work presented in this thesis, it is expected that the Atlantic salmon genome will be full of repetitive DNA.

This study suggests a baseline number for the fraction of the Atlantic salmon genome that might contain repetitive DNA (Table 3). Within these two BAC sequences, 13 to 14% of the sequence was composed of known repetitive DNA, including retroelements such as LINEs, SINEs, and LTRs, transposable elements, simple sequence repeats such as microsatellites and minisatellites, and low complexity DNA. In general, LINEs were more prevalent than SINEs in both BACs and the S0188I22 BAC sequence contained more than twice as many LINEs as S0085O16. On the other hand, transposons were more plentiful in S0085O16. Using a salmonid specific repeat database developed in our lab, a further 13% of each BAC sequence was found to be similar to salmonid specific repeats. Finally, further characterization of the BAC sequences identified a large portion of minisatellite DNA (3.8% and 9.2%), yielding an overall estimate of 30-36% of repetitive DNA within the BAC sequences. This fraction is probably an underestimate, and could increase as the salmonid specific repeat database becomes more robust with respect to the number of identified Atlantic salmon and salmonid specific repetitive elements.

The proportion of minisatellites within the BAC sequences is intriguing considering the fact the fluorescent *in situ* hybridization (FISH) experiments performed with these two BACs resulted in the identification of unique chromosomes for both BACs. If these particular minisatellite sequences, especially the blue minisatellite considering its ample size, were distributed throughout the Atlantic salmon genome, it would be expected that a BAC clone containing this sequence would hybridize to multiple chromosomes. As this was not observed, it is possible that either these minisatellites are unique to their respective locations in the genome, or their size is insufficient to allow for hybridization during FISH. It would be informative to screen the BAC library with probes designed based on the minisatellite sequences and determine whether or not many contigs would be identified.

Comparing the relative location of repetitive elements between the BAC sequences allows for an estimation of how active these units have been in the Atlantic salmon genome since the genome duplication event. In fact, with the exception of the blue minisatellite, which falls within an intron in the CBFB paralogs, it appears that none of the annotated repeat locations is conserved between the BACs, suggesting a high level of repeat activity since the duplication event. For instance, several LINEs have infiltrated intron 18 in the S0188I22 deadeye paralog, creating an intron almost three times as long as the corresponding intron in the S0085O16 sequence where no LINEs exist. Intergenic distance has also been affected by the insertion of repetitive elements. For example, the distance between the Lin10 and CBFB paralogs in S0188I22 is 16kb in length, in which a number of transposable elements are present, which contrasts the corresponding intergenic distance in S0085O16 that contains no transposable elements and is only 4.5kb in length. Finally, the presence and location of the shorter minisatellites are not conserved between the BAC sequences with the exception of the blue minisatellite.

4.2.3 Duplicate Genes

Theory suggests three possible outcomes for the evolutionary trajectory of gene duplicates: (1) non-functionalization, where one copy is silenced by degenerative mutations leading to pseudogenization, (2) neo-functionalization, where one copy acquires a selectively advantageous function such that it is preserved by natural selection, and (3) subfunctionalization, where both copies are partially compromised by deleterious, but complementary mutations to a point where their cumulative function is equal to that of the original gene (Force et al., 1999).

Both the explosion of sequence data and the examination of fully sequenced genomes have provided ample material for studying the evolutionary fates of duplicate genes. In this study, nine duplicate genes were identified within the overlapping BAC sequence data and an additional gene was found in one of the BAC sequences outside the boundaries of the overlapping sequence data (Table 2). In order to make predictions about which evolutionary path each gene in a duplicate pair is taking, it is important to correctly annotate each coding sequence. Of these genes, only the metallothionein gene duplicates have been characterized in salmonids. Therefore, the remaining genes were annotated using a comparative genomic approach by making use of orthologous sequences previously annotated in human, chicken, and zebrafish. Additionally, when possible, ESTs from Atlantic salmon and rainbow trout were used in the annotation process. Overall, the exon lengths among homologs of the genes examined in this study were quite conserved, thereby allowing strong predictions to be made on the correct coding sequence in Atlantic salmon genome in the absence of EST data. Consensus splice sites at the intron/exon boundaries provided further evidence for correct exon predictions.

Of the three evolutionary fates of duplicate genes suggested by the model, the most readily identified is that of non-functionalization. A pseudogene is quite easily recognized by the accumulation of deleterious mutations, provided that the initial pseudogenization event was recent enough such that remnants of the coding sequence are still recognizable. In this study, two complete gene/pseudogene duplicate pairs were identified: BBS2 and SLC12A3. Pseudogenization is thought to be the most probable fate of duplicate genes considering that the majority of mutations are likely to be deleterious (Ohno, 1970a; Nadeau and Sankoff, 1997).

In contrast to this, in several instances the coding sequences for both genes of a duplicate pair are intact. Such is the case for the Lin10, dead eye, beta-1,3-galactosyltransferase, and of course the metallothionein gene duplicates. It is more difficult in these instances to predict what is happening to these duplicate genes with respect to the fate of these gene duplicates. It is possible that deleterious, but complementary mutations in the regulatory regions have led to subfunction partitioning.

This is likely the case with the MetA and MetB genes, considering the differences in the MRE locations, and also that in rainbow trout, the genes are differentially regulated in response to various metals, as well as during different developmental stages (Olsson et al., 1990). Alternatively, it is possible that an increase in dosage associated with two copies of the gene has provided a selective advantage such that both duplicates have been preserved. In fact, mounting evidence indicates that most duplicated genes are not redundant from the start because of selection for increased gene dosage (Force et al., 1999; Graur and Li, 1999; Kondrashov et al., 2002).

In the cases of the duplicate genes where the complete coding sequence could not be identified, predictions about the fate of these duplicates become even more tenuous. For instance, because of sequence gaps in the BAC data, complete coding sequence could not be predicted for either of the CBFB gene duplicates. Within the available sequence data, however, one recognized splice site is missing in the S0085O16 copy of the gene, which could indicate non-functionalization at this locus; no other indications of pseudogenization exist, but without the complete coding sequence it is difficult to make a firm statement about the condition of this gene pair. For the GNAO1 gene duplicates, the sequence of the first three exons of both duplicates is presumed to lie beyond the available sequence data, and examining the known sequence data reveals no evidence of non-functionalization. The portion of the GNAO1 Atlantic salmon genes that is known is highly conserved with respect to orthologous sequences, suggesting that natural selection is working to maintain the sequence of both Atlantic salmon copies. This could be suggestive of subfunctionalization. Finally, in the Herp gene duplicates, the complete coding sequence was identified in S0085O16, but only the first three exons of the gene in S0188I22 due to the gene residing at the end of the sequence data. Predictions on the evolutionary fate of this duplicate pair are also difficult to make.

4.3 Conservation of Synteny

The gene order and transcriptional orientation of the duplicate genes identified in the homeologous BAC sequences have been conserved in Atlantic salmon since the genome duplication event (Figure 18). Remarkably, the 10 genes identified in this study also exhibit an astounding degree of conservation of synteny across a diverse range of vertebrate taxa (Figure 19). With respect to human and chick, the 10 genes reside on chromosomes 16 and 11, respectively, although gene order and orientation is variable with respect to Atlantic salmon. Interestingly, in zebrafish, a chromosomal translocation event has disrupted the synteny and three of the genes, including metallothionein, lie on chromosome 14, while the rest, with the exception of the beta-1,3-galactosyltransferase whose ortholog could not be identified, are on chromosome 18. The most parsimonious explanation is that all ten genes were originally found on a single ancestral chromosome, and a zebrafish specific translocation event has led to the disruption of synteny. Of note, three sets of genes have remained linked in Atlantic salmon, zebrafish, human, and chick: Cetp and Herp; deadeye and SLC12A3; metallothionein and BBS2.

4.4 Dating the Tetraploidization Event

Plenty of evidence suggests that the common ancestor of all salmonids underwent a whole genome duplication; the duplicate genes examined in this study are likely the result of this event, as discussed above. The duplication had to occur after the salmonids diverged from the nearest ancestor, but before the three subfamilies diverged. A scanty fossil record has not provided firm evidence as to the date of the duplication event and currently, it is thought to have occurred between 20 and 120 million years ago (Allendorf and Thorgaard, 1984).

This study makes available, for the first time, a set of duplicate genes from regions of the Atlantic salmon genome that arose via the tetraploidization event. By estimating the divergence time of each pair of gene duplicates, it is possible to provide a more defined estimate of the likely time of the genome duplication, which, by definition, had to pre-date the divergence of these genes. In performing this analysis, it is important to consider whether the gene duplicates are evolving at similar rates. Unlike orthologous gene sequences, where it is generally accepted that evolutionary rates for a particular gene are highly correlated in diverse lineages (Brohman and Penny, 2003), the evolutionary rate of duplicate genes within a single organism may not be the same because of the effect of relaxation of natural selection on the two copies.

For this analysis, the gene duplicates for dead eye, lin10, and GNAO1 were considered. To determine whether the duplicate genes are evolving asymmetrically, the number of nucleotide and amino acid changes between each duplicate gene and several orthologous sequences was examined. In all cases, there were no major difference in the number of sequence changes between each of the S0085O16 and S0188I22 sequences and the orthologous sequence and therefore, with respect to these three sets of duplicate genes, the evolutionary rate between duplicates is assumed to be equal. Furthermore, each of these three gene sequences is quite long, which nullifies the effect of any local selection occurring at specific sites along the length of the sequence.

Divergence time estimates were based on pairwise distances between orthologs, and calibrated based on accepted speciation times. For comparative purposes, both nucleotide and amino acid sequence data were included in the pairwise distance analyses, although at deep divergence points, saturation of four fold degenerate sites may lead to an underestimate of divergence times with respect to the nucleotide analyses.

Based on the three pairs of gene duplicates, the divergence time is estimated at 80 to 120 million years ago, which lies at the more ancient end of the predicted time of the salmonid tetraploidization event (Figure 31 and Figure 32). This time point provides a minimal estimate of the timing of the salmonid tetraploidization event. Following a whole genome duplication, disomic inheritance must be re-established before independent evolution of the homeologous chromosomes can begin, and therefore if divergence of the genes occurred about 100mya, then the duplication event had to have preceded that time.

Other duplicate gene pairs were excluded from the study based on short sequence length and missing exon sequence (Herp and CBFB), the existence of paralogs in other species confounding the identification of the true ortholog of the Atlantic salmon duplicates (metallothionein), and poor sequence conservation among orthologs (beta-1,3-galactosyltransferase).

4.5 Future Directions

This study characterized the evolutionary events that have occurred within homeologous regions of the Atlantic salmon genome since the time of the tetraploidization event and provides insight into the organization of the whole genome of the Atlantic salmon in terms of G+C content and repetitive DNA.

In order to further characterize the genes identified in this study, full length coding sequences need to be determined to verify the annotations, and furthermore expression studies should be carried out to test for subfunctionalization; variable temporal and spatial expression patterns of duplicate genes are indicative of this.

Furthermore, it would extremely interesting from an evolutionary perspective to examine the evolutionary trajectory of the genomic region surrounding the metallothionein loci in another salmonid species such as rainbow trout, where a genomic BAC library has also been developed.

Finally, comparative sequence analysis to identify conserved non-coding regions may prove to be informative in identifying regulatory elements associated with the genes.

REFERENCE LIST

Allendorf FW and GH Thorgaard. 1984. *Tetraploidy and the evolution of salmonid fishes*, pp. 1-53 in Evolutionary Genetics of Fishes, edited by BJ Turner. Plenum Press, New York.

Altschul SF, W Gish, W Miller, EW Myers, and DJ Lipman. 1990. Basic local alignment search tool. J*ournal of Molecular Biology*. 215: 403-410.

Amores A, A Force, Y-L Yan, L Joly, C Amemiya, A Fritz, RK Ho, J Langeland, V Prince, Y-L Wang, M Westerfield, M Ekker, and JH Postlethwait. 1998. Zebrafish *Hox* Clusters and Vertebrate Genome Evolution. *Science*. 282: 1711-1714.

Aparicio et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*. 297: 1301-1310.

Artieri C, IE Bosdet, R Chiu, RG Danzmann, WS Davidson, MM Ferguson, CD Fjell, B Hoyheim, SJM Jones, MA Marra, L Mitchell, C Mathewson, SHS Ng, K Osoegawa, SE Parisotto, RB Phillips, ML Rise, KR von Schalburg, JE Schein, H Shin, A Siddiqui, J Thorsen, and N Wye. 2004. Towards a physical map of the genome of the Atlantic salmon, *Salmo salar*. *Genome Research*. Submitted.

Bargelloni L, R Scudiero, E Parisi, V Carginale, C Capasso, and T Patarnello. 1999. Metallothioneins in Antarctic Fish: Evidence for Independent Duplication and Gene conversion. *Molecular Biology and Evolution*. 16: 885-897.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene*. 241: 3-17.

Bromham L and D Penny. 2003. The modern molecular clock. *Nature Reviews, Genetics*. 4: 216-224.

Burge C and S Karlin. 1997. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology. 268: 78-94.

Chan PC, CKM Shiu, FWY Wong, JKY Wong, WL Lam, and KM Chan. 2004. Common carp metallothionein-1 gene: cDNA cloning, gene structure and expression studies. *Biochimica et Biophysica Acta*. 1676: 162-171.

Cousins RJ. 1985. Absorption, Transport, and Hepatic Metabolism of Copper and Zinc: Special Reference to Metallothionein and Ceruloplasmin. *Physical Reviews*. 65: 238-209.

Coyle P, JC Philcox, LC Carey and AM Rofe. 2002. Metallothionein: the multipurpose protein. *Cellular and Molecular Life Sciences*. 59: 627-647.

Crespi BJ, and R Teo. 2002. Comparative Phylogenetic Analysis of the Evolution of Semelparity and Life History in Salmonid Fishes. *Evolution*. 56: 1008-1020.

Eichler EE and D Sankoff. 2003. Structural Dynamics of Eukaryotic Chromosome Evolution. *Science*. 301: 793-797

Elemento O, O Gascuel, and M-P Lefranc. 2002. Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution*. 19: 278-288.

Eyre-Walker A and LD Hurst. 2001. The Evolution of Isochores. *Nature Reviews, Genetics*. 2: 549-555.

Force A, M Lynch, FB Pickett, A Amores, Y-I Yan, and J Postlethwait. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*. 151: 1531-1545.

Graur D and W-H Li. 2000. *Fundamentals of Molecular Evolution, 2nd Edition*. Sinauer Associates, Sunderlan, Massachusetts.

Galtier N, M Gouy, and C Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*. 12: 543-548.

Gu X, Y Wang, J Gu. 2002. Age Distribution of Human Gene Families Shows Significant Roles of both Large- and Small-scale Duplications in Vertebrate Evolution. *Nature Genetics*. 31: 205-209.

Heierhosrt J, K Lederis, and D Richter. 1992. Presence of a member of the Tc1-like transposon family from nematodes and Drosophila within the vasotocin gene of a primitive vertebrate, the Pacific hagfish *Eptatretus stouti*. *Proceedings of the National Academy of Science, USA*. 89: 6798-6802.

Hong Y and M Schartl. 1992. Structure of the rainbow trout metallothionein gene. *Gene*. 120: 277-279.

Ishiguro NB, M Masaki, and M Nishida. 2003. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protocanthopterygii". *Molecular Phyogenetics and Evolution*. 27: 476-488.

Jabbari K and G Bernardi. 2004a. Body temperature and evolutionary genomics of vertebrates: a lesson from the genomes of *Takifugu* rubripes and *Tetraodon* nigoviridis. *Gene*. 333: 179-181.

Jabbari K and G Bernardi. 2004b. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*. 333: 143-149.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*. 431: 946-957.

Kashkush K, M Feldman, and AA Levy. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Nature Genetics*. 31: 1651-1659.

Kellis M, BW Birren, and ES Lander. Proof and Evolutionary Analysis of Ancient Genome Duplication in the yeast *Saccharomyces cerevisiae*. *Nature*. 428: 617-624.

Kille P, J Kay, and GE Sweeney. 1993. Analysis of regulatory elements flanking metallothionein genes in Cd-toleran fish (pike and stone loach). *Biochim et Biophysics Acta*. 1216: 55-64.

Kim CB, D Weiss, and F Ruddle. 2003. Survey of *Hox* genes in the skate, *Raja egalanteria*, pp. 133-138 in *Aquatic Genomics: Steps towards a Great Future*, edited by N Shiimizu, T Aoki, I Hirono, and F Takashima. Springer, New York.

Kirkness EF, V Bafna, AL Halpern, S Levy, K Remington, DB Rusch, AL Delcher, M Pop, W Wang, CM Fraser, and JC Venter. 2003. The Dog Genome: Survey Sequencing and Comparative Analysis. *Science*. 301: 1898-1903.

Kumar, S and B Hedges. 1998. A molecular timescale for vertebrate evolution. *Nature*. 392: 917-920.

Kumar S, K Tamura, and M Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Analysis and sequence alignment. *Briefings in Bioinformatics*. 5: 150-163.

Li W-H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderlan, Massachusetts.

Lin C-H, JAC John, LW Ou, J-C Chen, C-H Lin, and C-Y Chang. 2004. Cloning and characterization of metallothionein gene in ayu *Plecoglossus altivelus*. *Aquatic Toxicology*. 66:111-124.

Lynch M and AG Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. *American Naturalist*. 156: 590-605.

Lynch M and JS Conery. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*. 290: 1151-1155.

Maroni G, J Wise, JE Young, and E Otto. 1987. Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*. *Genetics*. 117: 739-744.

Massaro EJ. 1972. Isozyme patterns of coregonus fishes: Evidence for multiple cistrons for lactate and malate dehydrogenases and achromatic band in the tissue of *Prosopium cyclindraceum* (Pallas) and *P. coulteri*. *Journal of Experimental Zoology*. 179: 247-262.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*. 264: 421-423.

Meyer A and M Schartl. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current Opinion in Cell Biology*. 11: 699-704.

Mouse Genome Sequencing Consortium. 2002. Initial Sequencing and Comparative Analysis of the Mouse Genome. *Nature*. 420: 520-562.

Murphy MF, J Collier, P Koutz, and B Howard. 1990. Nucleotide sequence of the trout metallothionein A gene 5' regulatory region. *Nucleic Acids Research*. 18: 4622-4622.

Nadeau JH, and D Sankoff. 1997. Comparable rates of gene loss and functional divergence after genome duplications in early vertebrates. *Genetics*. 147: 1259-1266.

Nei M, and GV Glazko. 2002. Estimation of Divergence Times for a Few Mammalian and Several Primate Species. *The Journal of Heredity*. 93: 157-164.

Nelson, JS. 1994. *Fishes of the World, 3rd Edition*. Wiley and Sons, New York.

Oakley, TH and RB Phillips. 1999. Phylogeny of salmoninae fish based growth hormone introns: Atlantic (*Salmo*) and Pacific (*Oncorhynchus*) salmon are not sister taxa. *Molecular Phylogenetics and Evolution*. 11: 381-393.

Ogra Y, K Suzuki, P Gong, F Otsuka, and S Koizumi. 2001. Negative regulatory role of Sp1 in Metal Responsive Element-mediated Transcriptional Activation. *The Journal of Biological Chemistry*. 276: 16534-16539.

Ohno S, U Wolf, and NB Atkin. 1968. Evolution from fish to mammals by gene duplication. *Hereditas*. 59: 169-187.

Ohno S. 1970a. *Evolution by Gene Duplication*. Springer, Heidelberg.

Ohno S. 1970b. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*. 99:120-130.

Ohno S, J Muramoto, J Klein, and NB Atkin. 1969. Diploid-tetraploid relationship in clupeoid and salmonid fishes of the Pacific, pp. 139-147 in *Chromosomes Today, Volume 2*, edited by CD Darlington and KR Lewis. Plenum Press, New York.

Olsson P-E, SJ Hyllner, M Zafarullah, T Andersson, and L Gedamu. 1990. Differences in Metallothionein Gene Expression in Primary Cultures of Rainbow trout Hepatocytes and the RTH-149 Cell Line. *Biochimica et Biophysica Acta*. 1049: 78-82.

Olsson P-E, P Kling, LJ Erkell, and P Kille. 1995. Structural and functional analysis of the rainbow trout (Oncorhynchus mykiss) metallothionein-A gene. *European Journal of Biochemistry*. 230: 344-349.

Perez J, P Moran, and E Garcia-Vasquez. 20002. Isolation, characterization, and chromosomal location of the tRNAMet genes in Atlantic salmon (*Salmo salar*) and brown trout (*Salmo trutta*). *Genome*. 43: 185-190.

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

Prince, V. 2002. The *Hox* Paradox: More Complex(es) than Imagined. *Developmental Biology*. 249: 1-15.

Pruitt KD, T Tatusova, and DR Maglott. 2003. NCBI Reference Sequence Project: update and current status. Nucleic Acids Research. 31: 34-37.

Posthletwait J, A Amores, W Cresko, A Singer, YL Yan. Subfunction Partitioning, the Teleost Radiation and the Annotation of the Human Genome. *Trends in Genetics*. 20: 481-490.

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

Roberts FL. 1970. Atlantic salmon (Salmo salar) Chromosomes and Speciation. *Transactions of the American Fisheries Society*. 1: 105-111.

Rooney AP, H Piontkivska, and M Nei. 2002. Molecular Evolution of the Nontandemly Repeated Genes of the Histone 3 Multigene Family. *Molecular Biology and Evolution*. 19: 68-75.

Samonte RV, and EE Eichler. 2002. Segmental duplications and the evolution of the primate genome. *Nature Reviews, Genetics*. 3: 65-72.

Samson SLA, WJ Paramchuk and L Gedamu. 2002. The rainbow trout metallothionein-B gene promoter: contributions of distal promoter elements to metal and oxidant regulation. *Biochimica et Biophysica Acta*. 1517: 202-211.

Seoighe C. 2003. Turning Back the Clock on Ancient Genome Duplication. *Current Opinion in Genetics & Development*. 13: 636-643.

Spring, J. 1997. Vertebrate Evolution by interspecific hybridization – are we polyploid? *FEBS Letters*. 400: 2-8.

Schwartz S, Z Zhang, KA Frazer, A Smit, C Riemer, J Bouck, R Gibbs, R Hardison, and W Miller. 2000. PipMaker – a web server for aligning two genomic DNA sequences. *Genome Research*. 10: 577-586.

Sonhammer EL and R Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA andn protein sequence analysis. *Gene*. 167: GC1-10.

Stalker J, B Gibbons, P Meidl, J Smith, W Spooner, HR Hotz, and AV Cox. 2004. the Ensembl Web site: mechanics of a genome browser. Genome Research. 14: 951-955.

Taylor JS, I Braasch, T Frickey, A Meyer, and Y Van de Peer. 2003. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. *Genome Research*. 13: 382-390.

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409: 860-921.

Thompson JD, DG Higgins, and TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 22: 473-4680.

Thorsen J, B Zhu, E Frengen, K Osoegawa, PJ de Jong, BF Koop, WS Davidson, B Hoyheim. 2004. A highly redundant BAC library of Atlantic salmon (*Salmo salar*): an important tool for salmon projects. Submitted.

Tohoyama H, E Shiraishi, S Amano, M Inouhe, M Joho, and T Murayama. 1996. Amplification of a gene for metallothionein by tandem repeat in a strain of cadmium-resistant yeast cells. *FEMS Microbiology Letters*. 136: 269-273.

Van de Peer Y, JS Taylor, J Joseph, and A Meyer. 2002. Wanda: a database of duplicated fish genes. *Nucleic Acids Research*. 30: 109-112.

Van de Peer Y, JS Taylor, and A Meyer. 2003. Are all fishes ancient polyploids? pp. 65-73 in *Genome Evolution*, edited by A Meyer and Y Van de Peer. Kluwer Academics, Netherlands.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nature Reviews, Genetics*. 5: 752-763.

Vandepoele K, W De Vos, JS Taylor, A Meyer, and Y Van de Peer. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences, USA*. 101: 1638-1643.

Venkatesh, B. 2003. Evolution and diversity of fish genomes. *Current Opinion in Genetics and Development*. 13: 588-592.

Winnard P, B Sidell, and M Vayda. 2002. Teleost introns are characterized by a high A+T content. *Comparative Biochemistry and Physiology Part B: Biochemistry and Biology*. 133: 155-161.

Wagner GP, C Amemiya, and F Ruddle. 2003. *Hox* cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences, USA*. 100: 14603-14606.

Wolfe KH. 2001. Yesterday's Polyploids and the Mystery of Diploidization. *Nature* Reviews Genetics. 2: 333-341.

Woram RA, K Gharbi, T Sakamoto, B Hoyheim, L-E Holm, K Naish, C McGowan, MM Ferguson, RB Phillips, J Stein, R Guyomard, M Cairney, JB Taggart, R Powell, W Davidson, RG Danzmann. 2003. Comparative Genome Analysis of the Primary Sex-Determining Locus in Salmonid Fishes. *Genome Research*. 13: 272-280.

Wright JE, JR Heckman, and LM Atherton. 1975. Genetic and developmental analyses of LDH isozymes in trout, pp. 375-399 in *Isozymes III: Developmental Biology*, edited by CL Markert. Academic Press, New York.

Yan C H-M and K M Chan. 2002. Characterization of zebrafish metallothionein gene promoter in a zebrafish caudal fin cell-line, SJD.1. *Marine Environmental Research*. 54: 335-339.