ANALYSIS OF THE ANTIFREEZE GLYCOPROTEIN CONTAINING GENOMIC LOCUS IN THE ANTARCTIC NOTOTHENIOID FISH *DISSOSTICHUS MAWSONI*

BY

JESSIE DEE NICODEMUS JOHNSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Molecular and Integrative Physiology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Art DeVries, Chair
Professor C.-H.C. Cheng, Director of Research
Professor Byron Kemper
Professor Gary Olsen
Professor Kurt Kwast

ABSTRACT

Development of the Antarctic Circumpolar Current (ACC) circa 25 mya resulted in cooling of the high latitude waters of the Southern Ocean to a chilly -1.86 °C (near the freezing point of seawater) and extinction of most of the late Eocene temperate fish fauna. A notothenioid ancestral stock survived and went through an adaptive radiation that gave rise to a variety of ecotypes that filled the empty niches. The notothenioid fishes now account for 95% of the fish biomass that inhabits the continental shelf of Antarctica and islands of the Scotia Arc. The adaptive radiation was linked to the evolution of antifreeze glycoproteins (AFGPs). High blood levels of AFGPs (25 to 35 mg/ml) lower their freezing point a few tenths of a degree below that of seawater (-1.86°C) and are a vital part of their freeze avoidance strategy. The AFGP gene evolved from a trypsinogen-like protease (TLP) gene, and presumably through an ancestral intermediate, a chimeric AFGP/TLP gene. All three types of genes (TLP, AFGP, and chimeric AFGP/TLP) are found in Antarctic notothenioid genomes, but it is not known whether the chimeric gene is transcribed and translated into a protein that would provide both AFGP and TLP molecules.

The AFGP/TLP genomic locus of an Antarctic notothenioid, *Dissostichus mawsoni* was characterized in order to determine the mechanism of gene family expansion that would provide the high blood AFGP concentrations. The AFGP/TLP locus was isolated by screening a bacterial artificial chromosome (BAC) library for AFGP/TLP positive clones. Seven BAC clones representing two haplotypes encompassed the AFGP/TLP locus. Assembly of the AFGP/TLP locus was complicated by its highly repetitive nature. Thus, an assembly protocol was developed which entailed construction of subclone libraries of two insert size ranges (1-5 kbp and 5-30 kbp) for some of the positive BAC clones. BAC clone shotgun subclone libraries were then sequenced and subjected to automated and manual sequence reconstruction. Matching of paired-end sequences of some of the 1-5 kbp and all of the 5-30 kbp shotgun subclones to the locus sequence assembly was carried out to establish the linear order of genes. The AFGP/TLP locus assembly and analysis showed a high AFGP gene dosage (14 AFGP polyprotein genes in haplotype 1 and 8 AFGP polyprotein genes in haplotype 2) that very likely resulted from segmental duplications of the AFGP gene and its flanking regions, as seen in the >95 % sequence identity between AFGP gene modules. Thus it is clear that extensive AFGP gene duplication resulting in high gene dosage is the molecular basis for the high serum AFGP concentrations observed in the Antarctic notothenioids.

Besides the AFGP genes, the locus contains three AFGP/TLP chimeric genes and two TLP genes. Bayesian and Maximum Likelihood phylogenetic reconstructions of the AFGP, TLP, and AFGP/TLP chimeric coding regions indicated that the AFGP gene family arose from an ancestral chimeric gene related to a specific chimeric gene in the locus. Analysis of this extant paralog of the chimeric gene ancestor of AFGP gene revealed that the first stand-alone AFGP gene was most likely formed by slipped misalignment on the template strand during DNA replication in the chimeric ancestor, resulting in the removal of the bulk of the TLP coding regions. We hypothesize that extensive AFGP gene duplication may have been propagated by a recombination hotspot located downstream of all AFGP genes. Double stranded DNA breakage at this recombination hotspot may have resulted in AFGP gene duplication via non-homologous segmental duplication by unequal crossing-over. Increased AFGP gene dosage conceivably was selected for upon advent of icy Antarctic marine conditions, increasing survival fitness in Antarctic notothenioids in the form of increased serum AFGP concentrations.

Examination of the AFGP polypeptide coding regions of AFGP and AFGP/TLP chimeric genes (exon 2) showed that the AFGP genes encode predominantly the smaller AFGP molecules, consistent with their high abundance observed in the serum. The larger AFGP molecules are predominantly encoded in the AFGP/TLP chimeric genes. The chimeric genes are transcribed, and the tissue distribution of the chimeric gene transcript expression is similar to that of AFGP genes, suggesting that the chimeric gene may be functional in present day Antarctic notothenioids as an AFGP, providing the larger serum AFGP molecules.

Three types of trypsinogen genes are also associated with the AFGP/TLP genomic locus. Bayesian and Maximum Likelihood phylogenetic reconstructions using spliced *D. mawsoni* trypsinogen coding sequences, and a large sampling of vertebrate trypsinogen sequences from the NCBI EST and nucleotide databases, indicate *D. mawsoni* trypsinogens belonged to two of the three previously classified teleost trypsinogen gene types, group I (digestive) trypsinogen and group III (cold active) trypsinogen. Group I and group III trypsinogens are located within clade I and clade III respectively identified in our analysis. Phylogenetic analysis and intron-exon structure mapping revealed that clade III trypsinogens consists of two distinct subclades (clade IIIA and IIIB), encompassing the previously classified cold-active group III trypsinogens. BLAST searches of NCBI ESTs from different teleosts revealed clade III trypsinogens to be present in more basal warm-water teleosts (catfish, Siluriformes), suggesting they evolved in a time of warm climate, and thus were not of cold adaptive origin thus the cold active capability of some extant paralogs may be a subsequent evolutionary acquisition. Antarctic notothenioid clade I and III trypsinogen transcript abundance were determined by relative qPCR. Both clade III trypsinogens transcripts were higher than that of clade I in the Antarctic notothenioid *D. mawsoni*. Tissue expression distributions of Antarctic notothenioid clade III trypsinogens, determined by PCR, were also broader than those of temperate *O. mykiss* clade III paralogs. Absolute qPCR quantification of transcript expression showed that in a warm-acclimated Antarctic notothenioid fish, clade III trypsinogen transcripts decreased while clade I trypsinogen transcripts increased. These analyses suggest clade III trypsins, which are expressed at very low levels in warm water teleosts, were recruited in Antarctic notothenioids for potential cold-active capabilities, and evolved into the primarily expressed trypsinogen type in these fishes. The clade III trypsinogens that persist at low transcript levels in warm-water teleosts may perform other proteolytic functions unrelated to digestion or their potential cold-active capabilities.

To my husband who helped me maintain my sanity.

ACKNOWLEDGMENTS

This project would not have been possible without the help and support of many people. I would like to thank my advisors Dr. Chris Cheng and Dr. Art DeVries for their support throughout my time here. I would also like to thank my thesis committee Dr. Gary Olsen, Dr. Kurt Kwast, and Dr. Byron Kemper for their additional guidance and the University of Illinois Urbana Champaign for the Dissertation Completion Fellowship which helped me finish in a timely manner. I would also like to thank my husband Dr. Keven Johnson whose love and support helped me immensely. Last but not least I would like to thank my friend Vanessa Noboa who made my time in the basement more bearable and definitely more interesting.

TABLE OF CONTENTS

OVERVIEW

The fragmentation of Gondwana and continental drift led to the complete isolation of Antarctica approximately 38 Mya. Southern Ocean temperatures decreased as a result of the development of the Antarctic Circumpolar Current (ACC) 25-22 Mya, reaching present day icy conditions about 10-15 Mya (Eastman 1993). Today the ACC is a clockwise and northeastward moving current that varies in width between 200 and 1000 km depending upon location and reaches all the way to the ocean floor (Foster 1984; Eastman 1993). The ACC is both a thermal barrier, separating the cold Antarctic waters from temperate subtropical waters, and a physical barrier, preventing the migration of fishes between the two environments (Kennett 1982; Anderson 1999).

As the Southern Ocean cooled much of the late Eocene fish fauna failed to cold adapt and became extinct. The ancestral stock of the perciform suborder Notothenioidei adapted to the cooling waters and went through an adaptive radiation generating ecotypes that occupied the empty niches vacated by the extinct fauna. The present day Antarctic notothenioids comprise five families: Nototheniidae, Artedidraconidae, Harpagiferidae Channichthyidae, and Bathydraconidae (Eastman 2005). Notothenioid adaptive radiation was linked to the evolution of a biological antifreeze in the ancestral stock that ensured survival in freezing seawater. AFGPs occur as a family of approximately 16 size isoforms that circulate in the blood at concentrations from 5 to 35 mg/ml in the various members of the notothenioid families (Jin and DeVries, 2006). The AFGPs along with the salts in the blood lower the notothenioid freezing point to a few tenths of a degree below that of seawater (-1.86°C) (O'Grady, Ellory, and DeVries 1982). AFGPs depress the freezing point by a non-colligative mechanism, adsorption-inhibition where AFGP adsorption to ice crystals results in inhibition of crystal growth, by making it thermodynamically unfavorable for water molecules to join the ice crystal lattice at a given temperature (DeVries 1971).

AFGPs consist of a varying number of glycotripeptide repeat (Ala-Ala-Thr)$_n$, with each threonine linked to a disaccharide of galactose-N-acetylgalactosamine (DeVries, Komatsu, and Feeney 1970; DeVries, Vandenheede, and Feeney 1971). These proteins are encoded and translated as a polyprotein precursor of multiple AFGP isoforms separated by conserved three-residue spacers. It is thought that the polyprotein is cleaved at the spacers by a chymotrypsinogen-like protease into individual AFGP molecules, containing from 4 to 86 repeats (Hsiao et al. 1990; Chen, DeVries, and Cheng 1997b). The AFGPs, originally thought to be secreted by the liver, are synthesized by the pancreas and secreted into the anterior end of the small intestine. AFGP deposition into the hypoosmotic intestinal fluid prevents innoculative freezing of the gut by ice crystals that enter during feeding and drinking seawater. Thus far it is unknown how the AFGPs enter the circulation, but it is possible that they may be transported from the gut to the blood (Cheng et al. 2006). However, thus far there is no conclusive evidence for translocation and this is an area that requires investigation. The appearance and evolution of the AFGP lineage was estimated to have taken place about 15 mya which was when icy freezing conditions were established in the high latitude Antarctic waters. This correlation supports the hypothesis that AFGP evolution was driven by environmental selective pressures (Chen, DeVries, and Cheng 1997b). Studies of temperate water notothenioids that apparently escaped the freezing Antarctic waters show that the blood

AFGP level is physiologically insignificant (0.4 μg/mL) (Cheng et al. 2003) and AFGP serum concentrations have also been shown to be low in Antarctic notothenioids inhabiting the deep warm water and less ice-laden surface waters of the Antarctic Peninsula (Jin and DeVries 2006; Bilyk and DeVries 2009), further linking serum AFGP concentration to environmental modulation. Endemic high latitude Antarctic species show many strongly hybridizing bands in genomic southern blots indicative of a large AFGP multigene family (Cheng et al. 2003).

The AFGP gene arose from a functionally unrelated trypsinogen-like protease (TLP)(Chen, DeVries, and Cheng 1997b). Phylogenetic analyses of available cDNA sequences of various types of vertebrate trypsinogens suggested that TLP or a TLP-like gene arose from a digestive trypsinogen, however the specific TLP evolutionary precursor has not been determined (Roach et al. 1997; Roach 2002). Genomic screens, via Southern blots, showed that the basal notothenioid species genomic DNA contained the TLP gene, but not the AFGP gene (Cheng et al. 2003). The TLP-like gene and protein product are found in species of the order Gadiformes (Gudmundsdottir and Palsdottir 2005), a relatively old teleost lineage predating the notothenioids (Perciformes). This suggests that a TLP-like gene might have evolved early in the Teleostei lineage however, a more definitive time frame of TLP emergence remains to be determined.

The AFGP coding region evolved from a single 9-nt Thr-Ala-Ala coding element that spans the intron 1/exon 2 junction of the TLP gene (Chen, DeVries, and Cheng 1997b) (Figure 1). Expansion of this 9-nt coding element may have been the result of unequal crossing over or slippage duplication driven by a repetitive (gt_n) sequence always found immediately upstream of the Thr-Ala-Ala coding element at the TLP intron 1/exon

2 junction, forming an ancestral intermediate termed the AFGP/TLP chimeric gene (Cheng and Chen 1999). Development of the AFGP gene involved recruitment of the 5' and 3' ends of the TLP gene, as well as deletion of most of the TLP coding region by an unknown mechanism (Chen, DeVries, and Cheng 1997b). All three of these genes persist within extant notothenioid genomes. The TLP-like protease has been shown to be a psychrophilic (cold active) trypsin in *G. morhua* (a cold temperate water teleost) and a similar function is suggested for the Antarctic notothenioid because they share significant sequence identity. The function of the chimeric gene in Antarctic notothenioids is unknown.

It is clear that the driving force for the abundance of AFGP genes and protein within Antarctic notothenioids is the frigidity of the environment they inhabit. As the environment became colder, and ice more prevalent, duplications of the AFGP gene and expansion of the AFGP genomic locus, and/or alteration of the promoter region may have occurred augmenting increased protein production within the Antarctic notothenioids. What remains to be understood is how this occurred at the molecular level. The overall goal of my research was to sequence and characterize the AFGP/TLP genomic locus of the Antarctic nototheniid *D. mawsoni.* AFGP locus sequencing would allow insight into the adaptation of *D. mawsoni* AFGP gene containing regions in response to extreme cold (i.e. environmental selective pressures). The research involved first isolating, sequencing and analyzing the AFGP genomic locus. The locus in *D. mawsoni* is highly complex because the AFGP gene family is large. Characterizations of the *D. mawsoni* AFGP/TLP genomic locus included the following:

1) Sequencing and reconstruction of the AFGP/TLP locus in *D. mawsoni*. Alignment of the *D. mawsoni* AFGP/TLP gene containing region was predicted to be 400 kbp in size and consist of 3 individual, tightly grouped loci by Finger Printing Contig analysis of the AFGP/TLP positive BAC clones from a BAC library constructed from *D. mawsoni* genomic DNA. The AFGP gene coding region is known to be highly repetitive. Sequence assembly of the highly repetitive genomic regions (both Ala-Ala-Thr repeat and tandem AFGP genes) presented many technical problems. I developed and outlined in Chapter 1 an efficient way to reconstruct the *D. mawsoni* AFGP/TLP locus.

2) AFGP/TLP locus sequence analysis. The AFGP genes were presumed, and confirmed by our *D. mawsoni* AFGP/TLP genomic locus sequence reconstruction, to be located within the same chromosomal region as the AFGP/TLP chimeric gene. The detection of all three types of genes within the same locus further establishes AFGPs TLP ancestry through a chimeric ancestral intermediate. Further analysis of locus sequence allowed determination of the molecular and genomic mechanisms responsible for the AFGP gene family expansion that provides a high gene dosage producing the high concentration of serum AFGP observed in most of the Antarctic notothenioids (Chapter 2).

3) TLP has been shown to be related to other teleost digestive trypsinogen genes by sequence identity (Roach 2002). There are multiple types of trypsinogen genes and subsequent trypsin products detectable in teleosts (Roach et al. 1997). The *D. mawsoni* AFGP/TLP locus sequence reconstruction revealed multiple trypsinogen genes to be associated with the locus. Phylogenetic analysis of these *D. mawsoni* trypsinogens with

respect to AFGP and TLP provided needed insight into TLP evolution, as well as vertebrate trypsin evolution in general (Chapter 3).

CHAPTER 1

ASSEMBLY OF THE ANTIFREEZE GLYCOPROTEIN/TRYPSINOGEN-LIKE PROTEASE GENOMIC LOCUS

ABSTRACT

Antarctic notothenioids possess many adaptations to extreme cold not found in their temperate ancestors. The cold adaptive genomic architecture has yet to be studied for many Antarctic notothenioid adaptive phenotypes. Construction of a bacterial artificial chromosome (BAC) library (10X genome coverage) for the Antarctic notothenioid, *D. mawsoni,* and corresponding sequence data for BAC clones of interest would allow for the targeted examination of underlying genomic architecture associated with a given phenotype. The antifreeze glycoprotein (AFGP) gene locus, being one of the most prominent notothenioid cold adaptations, was examined first. AFGP/trypsinogen-like protease (TLP) positive *D. mawsoni* BAC clones (86) *Hind*III restriction digest alignments determined by Finger Printing Contig indicated there to be three separate AFGP/TLP genomic loci. Chromosomal FISH showed all three loci were confined to one chromosomal region. Shotgun subclone sequence reconstruction of BAC clones within the minimal tiling path (MTP) indicated the locus to be highly repetitive, complicating alignment methods. Manual verification and manipulation of shotgun subclone alignments (1-5 kbp and 5-30 kbp) allowed assembly of most of the AFGP/TLP locus. Shotgun subclone sequence reconstruction indicates there to be one AFGP/TLP locus and two AFGP haplotypes within the genome of our *D. mawsoni* specimen. Modifications to standard sequence alignment procedures and extensive measures to

check alignment accuracy provided genomic sequence data for two AFGP/TLP locus haplotypes in *D. mawsoni*.

INTRODUCTION

Antarctic fishes represent a cold stenothermal class of fishes, isolated within the Southern Ocean by the Antarctic Circumpolar Current (ACC). Development of the ACC approximately 25 mya and cooling of Antarctic waters (Kennett 1982) resulted in adaptations to the cold environment, unique to extant Antarctic fish families. Proteins that show cold adaptation by residue modifications that increase activity at low Antarctic temperatures are: lactate dehydrogenase-A (Somero 2003), lens crystallins (Kiss et al. 2004), and digestive trypsins (Genicot et al. 1996). One of the most notable notothenioid adaptations to cold is the development of the Antifreeze glycoprotein (AFGP) gene family (Chen, DeVries, and Cheng 1997b; Cheng and Chen 1999).

Antifreeze glycoprotein gene genesis is the single most important factor allowing Antarctic notothenioid survival in the icy waters of the Southern Ocean. Antarctic notothenioid fishes, like all saltwater teleosts, have an internal osmolality hypoosmotic to that of seawater (DeVries 1971; O'Grady, Ellory, and DeVries 1982), and thus they are in danger of freezing in icy Antarctic sea water. Hypoosmotic serum fluid, having a lower salt concentration than seawater has a higher freezing point, and would readily freeze upon contact with ice. Antifreeze glycoproteins, present in the serum of these fishes, lower the freezing point of serum below that of sea water by binding to ice crystals that enter the fish and preventing their growth, and thus freezing (DeVries 1971). AFGPs conferral of freeze avoidance is necessary for Antarctic notothenioid survival and thus allows for subsequent adaptations to cold.

The AFGP gene evolved from a trypsinogen-like protease (TLP) through an ancestral AFGP/TLP chimeric intermediate (Cheng and Chen 1999). The AFGP protein is synthesized as a polyprotein consisting primarily of Ala-Ala-Thr tandem repeats separated by a three residue linker sequence that is removed by a protease cleavage yielding multiple mature proteins. AFGPs are present in the serum of Antarctic notothenioids at concentrations ranging from 5-35 mg/mL (Jin and DeVries 2006). Although the AFGP gene origin has been established, many aspects of its gene evolution are unknown. Characterization of the AFGP/TLP genomic locus may shed light on the genomic architecture behind the observed AFGP heterogeneity and the genomic basis of the high levels of serum AFGP in Antarctic notothenioids. AFGP linker sequence origin and an explanation (if any) for ancestral AFGP/TLP chimeric genomic retention are all important additional questions. These questions can be answered by sequencing the AFGP/TLP genomic locus of an Antarctic notothenioid fish. Answers to these questions and others in a genomic context will allow us to address the question of genomic adaptation (AFGP gene genesis and gene family expansion) to extreme environmental selective pressures of the freezing Southern Ocean.

BAC library construction is a standard method used for targeted genomic sequencing. Protocols for BAC library construction, identification of clones containing the gene of interest, construction of a minimal tiling path (MTP) (smallest number of overlapping clones spanning an entire region of interest) have become more common (Amemiya et al. 1999; Osoegawa et al. 2001; Szinay et al. 2008; Yasui et al. 2008; Jacobs et al. 2009). The BAC technique of targeted genomic sequencing coupled with the Sanger method of sequencing is being replaced by more efficient pyrosequencing methods that are able to generate sequences for entire genomes or BAC clones in a fraction of the time (Ronaghi 2001; Metzker 2005; Goldberg et al. 2006). However this technique is not optimal for our purposes as highly repetitive coding regions, similar to that of the AFGP coding sequence, are not easily aligned using the pyrosequencing method because of the short reads. BAC sequences generated sequencing of their shotgun subclone libraries using the Sanger method, affords longer reads and paired end matching of subcloned BAC fragments, both of which have been shown to be more effective in the sequencing of repetitive DNA elements (Wicker et al. 2006). We report sequence reconstruction of the highly repetitive AFGP/TLP genomic locus by identification and sequencing of pertinent clones within a *D. mawsoni* BAC library.

MATERIALS AND METHODS

Construction and Screening of a D. mawsoni Genome BAC library

Dissostichus mawsoni specimens were caught in McMurdo Sound, Antarctica. To obtain more suitably sized segments of the AFGP gene containing region(s) for characterization, a large insert bacterial artificial chromosome (BAC) library was constructed (Figure 2). A BAC library was necessary due to the predicted large size of the AFGP gene containing region(s). *D. mawsoni* genomic DNA was isolated from red blood cells. Genomic DNA handling and BAC library construction procedures were performed following methods previously described (Amemiya, Ota, and Litman 1996). The library was constructed in the pCC1BAC vector (Epicentre) using *Eco*RI partial genomic digests. The library was macroarrayed and screened for clones containing AFGP and TLP genes by Southern blot of library filters with probes specific for each. (BAC library construction and Southern blot screening was carried out by Dr. C.-H.C. Cheng).

Construction of a Minimal Tiling Path

To maintain insert stability, cells are grown at low plasmid copy numbers and must be induced to produce high insert plasmid copy numbers necessary for obtaining large amounts of DNA necessary for subsequent manipulation. Stationary cultures of AFGP/TLP positive clones in the copy control pCC1BAC vector system/EPI300 cells (Epicentre) were grown overnight at 37 °C in LB/chloramphenicol. Cultures were induced to produce high numbers of copies per cell following the manufacturer's suggestions (Epicentre). Induced cultures were isolated by an alkaline lysis miniprep method, digested with *Hind*III restriction enzyme, and digest fragments separated on a 0.8% agarose gel and run for 18 h at 4ºC. The gel was stained with ethidium bromide, and its digital image was refined in the program Image, to establish BAC clone digest patterns unique to each clone (data not shown). The program Finger Printing Contig (FPC) identified BAC clone overlapping regions based on their shared digest patterns (data not shown). This is based on the assumption that shared digest patterns are equivalent to shared sequence identity. FPC uses an algorithm that calculates the pair-wise probability of coincidence that two clones will contain matching fragments by random chance as opposed to by shared sequence (Soderlund, Longden, and Mott 1997; Soderlund et al. 2000).

Chromosomal In Situ Hybridization

Chromosomal preparation and fluorescent i*n situ* chromosomal hybridization (FISH) were carried out by C.-H.C. Cheng and collaborators according to the protocol previously described by (Mazzei et al. 2007). Purified plasmid extracts of DmBAC 64, 78, and 35 were used as probes.

Shotgun Library Construction, Sequencing and Assembly

MTP BAC clone DNA was isolated using an alkaline lysis procedure after high copy number induction of cultures as indicated above in the MTP construction section. Individual shotgun libraries (8X minimum coverage) were constructed and archived for each of the BAC clones within the MTP, using the pCR 4Blunt-TOPO vector in the TOPO shotgun subcloning kit (Invitrogen) according to manufacturer's instructions. Additional 5-30 kbp shotgun libraries were made and archived using the Big Easy subcloning system according to the manufacturer's instructions (Lucigen). Library glycerol stocks were stored at -80°C until needed. Plasmid preps from glycerol stocks were stored at -20°C until needed.

Plasmid inserts were sequenced. One to five kilobase pair library inserts were sequenced with T3 and T7 primers (provided by the manufacturer) which correspond to universal primer sites on the pCR 4Blunt-TOPO vector (Invitrogen). The pJAZZ-OK Blunt Big easy vector (Lucigen) 5-30 kbp insert libraries were sequenced using SL1 and NZ1 primers provided by the manufacturer. One to five and five to thirty kilobase pair reads were sequenced from both ends and matched up later (pair end matched) during the alignment process. Sequencing reactions were carried out using the ABI PRISM Big Dye v3 Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems) and run on the ABI3730*xl* sequence analyzer (Applied Biosystems) at the High throughput Sequencing unit at the Keck Center for Comparative and Functional Genomics (UIUC).

Sequence reads were generally of high quality and had about a 4 % reaction failure rate. Shotgun sequence files were edited, aligned, and analyzed independently in Sequencher 4.5 (Gene Codes). At least 8X sequence coverage was generated for each MTP BAC clone sequence assembly.

As a result of highly repetitive sequence and hairpin structures within the locus a contiguous sequence of the entire *D. mawsoni* AFGP/TLP genomic locus was not obtained. Gap sizes and contig orientation relative to flanking clones were determined by paired end matching of shotgun subclone end sequences (1-5 and/or 5-30 kbp) that spanned these gaps. Paired end matching is the manual verification that each pair of shotgun subclone end sequences align near each other and are oriented correctly (with respect to each other) within each assembled MTP BAC shotgun subclone sequence alignment. Large insert subclone sizes were determined by pulsed field gel electrophoresis of *Not*I restriction digested 5-30 kbp clones. Small insert subclone sizes were determined by gel electrophoresis and comparison to a 1 kbp standard (Promega) of *Eco*RI restriction digested clones.

BAC Clone Size Estimation by PFG Electrophoresis

AFGP/TLP positive BAC clone insert sizes were estimated from a *Not*I restriction digestion of purified *D. mawsoni* BAC clones. Digests were run on a 1% Agarose III (Amresco, Solon, OH) gel using a pulse field gel apparatus (CHEF mapper XA System, BioRad). PFG settings were as follows: Initial switch time = 0.1 seconds, Final switch time = 40 seconds, Field angle = 120°, Field Strength= 6v/cm; Run time = 10 hours, buffer temperature = 14 ° C. *Not*I digests were electrophoresed with low range pulsed field gel standards (New England Biolabs) for photo documentation.

BAC Clone Size Estimation by Shotgun Subclone MTP Consensus Sequence Assembly

The end sequences for AFGP/TLP positive BAC clones determined by FPC to overlap with MTP BAC clones (termed MTP overlapping BAC clones-MOBC) were obtained as mentioned in the preceding section and aligned within each of the MTP BAC clones 1-5 and 5-30 kbp shotgun subclone sequence assemblies. MOBC predicted sequence size was determined to be the distance between the BAC clone end sequences (paired ends) within their corresponding MTP BAC clone 1-5 and 5-30 kbp shotgun subclone assemblies. Determination of correct BAC clone placement required that BAC end sequences share ≤ 99% identity with the assembly region and were oriented correctly with respect to each other within their respective MTP BAC clone shotgun subclone sequence assembly. The predicted MOBC sequence size must also be comparable to the *Not*I enzymatic digest size (preceding section).

Assembly of the AFGP/TLP Genomic Locus Consensus Sequence from MTP BAC Clone Shotgun Subclone Library Sequence Assemblies

The consensus sequence of each MTP BAC clone 1-5 and 5-30 kbp shotgun subclone assemblies were analyzed for shared sequence identity between other MTP BAC clones. High amounts of shared sequence identity between BAC clones are indicative of a shared overlap. AFGP paralogs contain single nucleotide variations in intron 1 and exon 3 regions that are unique to each paralogous gene. Further confidence was placed in the BAC overlap prediction if the AFGP genes shared 100% nucleotide sequence identity. The alignment of two BAC clone overlaps, identified by high sequence identity, were verified by identification and paired end matching of MOBCs that span the predicted overlap. These MOBC paired end sequences would then be located within separate DmBAC clone shotgun subclone assemblies.

Although determination of high molecular weight insert size by *Not*I enzymatic digest and PFG electrophoresis can never be exact, the MOBC insert size determined by the physical distance between the paired end sequences within the DmBAC shotgun subclone assembly was verified to be approximately the same as MOBC insert size determined by *Not*I enzymatic digest. If MOBC insert sizes were significantly different (variability > 10-30 bp) or paired end sequences were not oriented correctly with respect to each other the predicted overlap was not supported by MBOC paired end matching. It should be noted that an existing variation between individual BAC clone shotgun subclone assembly consensus sequence size and the same BAC clone *Not*I enzymatic digest size would be carried artificially through all MOBCs spanning that region. Artificial variations in between MBOC paired end sequence size and *Not*I enzymatic digest were taken into account.

Identification and Analysis of Gene Presence and Abundance within the AFGP/TLP Gene Containing Region(s)

All MTP BAC clone shotgun subclone assembly consensus sequences were analyzed using BLASTn and tBLASTx to identify gene presence and orientation within each BAC clone. As NCBI BLAST functions limit query size, each BAC clone was screened for the presence of protein coding genes in 10 kbp segments. The protein coding sequence and intron-exon junctions were determined manually for AFGP genes using the Protein Translator function of JustBio (www.justbio.com). The AFGP gene exon 2 encodes for the AFGP polyprotein. AFGP genes within each BAC clone assembly were named according to the length (number of Ala-Ala-Thr repeats) within the last peptide encoded. For example if the last protein were 8 Ala-Ala-Thr repeats in length the protein was named A8. BAC clones containing multiple genes encoding the same terminal mature peptide length were numbered consecutively (i.e. A8-1, A8-2).

Phylogenetic Analysis

Phylogenetic relatedness of AFGP genes were determined using the AFGP signal peptide (exon 1) and intron 1. Sequence was collected from all BAC clones. The AFGP exon 2 sequence is highly repetitive. Highly repetitive sequence cannot be aligned with high confidence, thus exon 2 sequence was excluded from our phylogenetic analysis. Sequences were aligned using the Clustal function of MEGA v4 (Tamura et al. 2007). Phylogenetic analysis was performed using the Neighbor-joining tree function portion of MEGA v4. Analysis was run under pairwise deletion and maximum composite likelihood parameters. The tree represents one run of 1000 bootstrap replicates displayed as a consensus tree with branch support > 50%.

RESULTS AND DISCUSSION

Analysis of Dissostichus mawsoni Genomic BAC Library Shows AFGP Gene Presence Is Restricted to One Chromosomal Location.

Construction of a *Dissostichus mawsoni* genomic BAC library allows genomic examination of adaptations to extreme environmental temperatures in an organism known to have undergone large amounts of gene family expansion (Chen et al. 2008) and adaptation (DeVries 1971; Kiss et al. 2004; Jin and DeVries 2006; Cheng and Detrich 2007) in response to extreme cold. C.-H.C. Cheng determined the *D. mawsoni* genome to be 0.97 pg (roughly 970 Mbp (Dolezel et al. 2003) in size). The *D. mawsoni* BAC library consisted of 67,584 BAC clones. The average BAC library insert size was 110 kbp. This BAC library provided approximately 7.6 X genome coverage (C.-H.C. Cheng, unpublished data). Eighty six clones positive for AFGP and TLP were identified by a Southern Blot screen of *D. mawsoni* BAC library filters with probes specific for AFGP and TLP genes (C.-H.C. Cheng, unpublished data). These 86 clones represent AFGP/TLP genomic presence in *D. mawsoni.* FPC predicted these clones to span three distinct regions of the genome (collectively encompassing approximately 500 kbp) (Figure 3) (Table 1). Contig one (55 clones) was covered by a MTP of four BAC clones (DmBAC 39, 64, 10 and 42). Southern blot screens indicated contig one to consist of AFGP and TLP genes. Contig 2 (8 clones) consisted of only AFGP genes. Contig 2 has a MTP of one clone (DmBAC 78). The third locus consisted of a singleton (DmBAC 35). Southern blot screens indicated DmBAC 35 contained only TLP genes. Chromosomal *in situ* (Figure 4) indicates that all three AFGP/TLP positive contigs are located within the same region of the chromosome (C.-H.C. Cheng *et al.*, unpublished data).

Twenty three additional AFGP/TLP BAC clones were not placed within the three loci (contigs) listed above. Five of these clones (DmBAC 74, 75, 79, 80, and 85) determined by Southern blot screens to be positive for AFGP were aligned within the AFGP/TLP locus. Clone placement was determined by the alignment of their end sequences within the AFGP/TLP BAC clone MTP sequence reconstruction. FPC aligned eighteen additional clones, with trace AFGP/TLP signal into one contig (data not shown). End sequence data obtained from a few of these eighteen BAC clones (20, 17, 16, and 47) could not be aligned within the AFGP/TLP locus assembly due to low sequence identity. This suggests that these clones do not fall within the AFGP/TLP locus. These clones are most likely false positives. TLP, used for Southern Blot screening of BAC library filters, belongs to the trypsin family of proteases. Trypsin is a member of a large and diverse class of serine/threonine proteases (Powers et al. 1993; Barrett and Rawlings 1995). Probe recognition of alternate serine/threonine proteases (not trypsins) present within the *D. mawsoni* genome most likely account for the false positives observed here.

AFGP/TLP MTP BAC Clone 1-5 kbp Shotgun Subclone Sequence Reconstruction.

DmBAC clone 1-5 kbp shotgun sequence reconstruction (Figure 5) (Figure 6) (average insert size of 900 bp, 8-12X coverage for each BAC clone) shows AFGP/TLP gene presence corresponds with the gene content predictions of Southern blot screens of BAC library filters (data not shown). BAC clone 1-5 kbp shotgun subclone sequence reconstruction indicated this individual *D. mawsoni* genome contained a large number of AFGP genes**.** Preliminary sequence data also indicated there were many genes sharing high similarity to other previously identified gene types: chimeric genes (5), TLPs (4), and trypsins (13). The detection of a large number of genes identifiable as AFGP, introduces a plethora of highly repetitive coding sequence which complicated sequence alignment. Assembly of all AFGP containing regions were aided by 1-5 kbp shotgun subclone paired end sequence matching (approximately3.5 X coverage). A contiguous sequence alignment for each BAC clone shotgun subclone assembly was not obtained due to the presence of secondary structures that could not be sequenced through. BAC clone alignments are composed of a series of contigs, or regions of shotgun subclone sequence alignment, (ranging in size from 300bp -95 kbp) separated by gaps of varying sizes (300bp – 5 kbp).

The one to five kilobase pair shotgun subclone assembly of DmBAC 10 consisted of 10 contigs and 9 gaps which spanned 154.6 kbp (Figure 5). The DmBAC 10 shotgun subclone assembly predicted sequence size was similar to the *Not*I enzymatic digest size of 150 kbp. The correlation between both DmBAC 10 size prediction methods indicated there are few if any sequence alignment errors within DmBAC10 shotgun subclone sequence assembly. DmBAC 10 contained two AFGP/TLP chimeric, one trypsinogen III, eight trypsinogen III pseudogenes, five AFGP (10_A8, A9, A11, A7, A8-2) genes, and one 5' truncated AFGP pseudogene (10_A6) (Figure 5). *D. mawsoni* trypsinogen III genes are named as such because of the high sequence identity shared with *Paralicthys olivaceus* trypsinogen III discovered during NCBI BLAST and tBLASTx screens of the database nucleotide and protein sequences. Trypsinogen III pseudogenes correspond to exon 4, intron 4, and exon 5 of the intact *D. mawsoni* trypsinogen III genes. *D. mawsoni* 5' truncated pseudogenes correspond to nucleotide regions 465-874 of the *P. olivaceus* NCBI transcript (AB029752.2) unless otherwise indicated. DmBAC 10_A6 AFGP pseudogene contained a 5' truncation which resulted in the deletion of the AFGP signal peptide and intron 1 sequence. The 5' truncation corresponds to nucleotide sequences 1-1627 in the NCBI *D. mawsoni* AFGP gene sequence (DMU43149).

The one to five kilobase pair shotgun subclone assembly of DmBAC 64 consisted of 9 contigs and 8 gaps which spanned 145 kbp (Figure 5). The DmBAC 64 shotgun subclone assembly predicted sequence size was similar to the DmBAC 64 *Not*I enzymatic digest size of 150 kbp. The correlation between both DmBAC 64 size prediction methods indicated there were few if any sequence alignment errors within DmBAC64 shotgun subclone sequence assembly. DmBAC 64 contained five AFGP genes (64_A9, A11, A7, A8-1, and A39), two 5' truncated AFGP pseudogenes (Dm64_A6, Dm64_A6-2) and seven trypsinogen III pseudogenes (Figure 5), and one chimeric gene.

The one to five kilobase pair shotgun subclone assembly of DmBAC 39 consisted of 6 contigs and 6 gaps (Figure 6). One AFGP coding sequence contig could not be oriented within the sequence assembly which resulted in the increased gap number. One to five kilobase pair shotgun subclones predicted sequence size of 127.4 kbp were similar to the *Not*I enzymatic digest size (120 kbp). This indicates that although one contig could not be oriented with respect to the assembly the majority of the clone assemblies were most likely correct. DmBAC 39 contained one AFGP/TLP chimeric gene, two AFGP E2 coding fragments, one trypsinogen III pseudogene missing the intron 1/signal peptide encoding exon (Figure 6), and one gene sharing high sequence identity to the translocase of outer mitochondrial membrane 40 (TOMM40) gene of *D. rerio* (BC053295) (data not shown) (BAC shotgun subclone library, sequencing, and sequence editing of DmBAC 39 performed by S. Silic, re-alignment performed by J. Johnson).

DmBAC 35 contained seven trypsinogen I, one trypsinogen III, and one TLP gene (data not shown) (assembly and analysis by S. Silic). Further analysis of the DmBAC 35 one to five kilobase pair assembled consensus sequence (carried out by J. Johnson) identified the presence of one intact gene sharing high sequence identity to the hormone sensitive lipase of the perciform *Acanthopagrus schlegelii* (ACE95864). As DmBAC 39 and 35 contain AFGP/TLP/chimeric genes and then flanking, presumably unrelated genes, these clones most likely represent opposing ends of the AFGP/TLP genomic locus.

AFGP/TLP MTP BAC Clone Sequence Reconstruction in AFGP Gene Rich Regions Required construction and alignment of 5-30 kbp Shotgun Subclone Libraries

DmBAC 42 contained five trypsinogen I genes, three trypsinogen III genes, four trypsinogen III pseudogenes, two AFGP/TLP chimeric genes, and four AFGP (42_A8, A11, A9-1, and A9-2) genes (Figure 7). DmBAC 78 contained 7 AFGP (78_A9-1, A8, A4, A9-2, A10-1, A10-2, and A7) genes and 8 trypsinogen III pseudogenes (Figure 8). One to five kilobase pair shotgun subclone contigs ranged in size from 300bp – 16.8 kbp. Paired end matching of one to five kilobase pair shotgun subclone library assemblies (approximately 8X coverage) were unable to orient many 1-5 kbp contigs. DmBAC 42 and 78 five to thirty shotgun subclone libraries (5-30 kbp) were sequenced (approximately 600 bp reads) to 1.5X coverage and paired end sequences from subclones were aligned within the 1-5 kbp subclone contigs. DmBAC 42 was composed of six contigs and 5 gaps (Figure 7). DmBAC 78 was composed of 14 contigs and 13 gaps (Figure 8). The gap sizes between contigs ranged from 200 bp to 5.5 kbp determined by 1-5 and 5-30 kbp shotgun subclone paired end sequence matching (1 gap of 16 kbp) (Figure 7) (Figure 8). *Not*I digest and shotgun subclone sequence assembly size varied by 24 and 11.4 kbps for DmBAC 42 and 78 respectively. High sequence identity in the AFGP gene 5' and 3' flanking sequence resulted in the formation of multiple gaps within DmBAC 42 and 78 contigs. Shotgun subclones both single and paired end sequences (1-5 kbp) sharing identical sequence with two or more loci were misaligned by the Sequencher alignment algorithm. Five to thirty kilobase pair shotgun subclone paired end sequence matching verified (globally) the constructed contig alignments and oriented 1-5 kbp shotgun subclone contigs to each other.

DmBAC 42 and 78 shotgun subclone assembly predicted sizes are larger than their corresponding *Not*I enzymatic digest sizes. DmBAC 42 variability is larger than noted due to the presence of a gap of unknown size within the assembly. Both shotgun subclone sequence assemblies were realigned from start to finish however similar consensus sequences were obtained for both alignments. The large discrepancy between predicted sequence sizes suggests that these clones most likely contain sequence misalignments that are unable to be resolved by the current alignment methods.

Construction of the AFGP/TLP Genomic Locus by Shared Sequence Identity between Putative Overlapping BAC clones.

Figure 9 shows predicted *D. mawsoni* MTP BAC clone overlapping regions. Overlapping regions were identified by FPC (Figure 3) and/or shared sequence identity

(<90%) between MTP BAC clone consensus sequence alignments (data not shown). The DmBAC 42/35 overlap is consistent with chromosomal FISH data (Figure 4) which indicated the close proximity of this outlier (DmBAC 35) to the major locus. This overlap was not predicted by FPC (Figure 3). The DmBAC 42/35 40.5 kbp overlapping region contains 8 *Hind*III cleavage sites, most producing fragments <5 kbp (data not shown). The small number of bands of suitable size for FPC analysis may have caused the FPC alignment error.

The DmBAC 42/35 overlap also contained nucleotide variation within the TLP gene shared between the two clones in the overlapping region as well as in trypsinogen I and trypsinogen III genes (data not shown). Nucleotide variation is also detectable in DmBAC 64/39 overlap (13.5 kbp) (data not shown). Nucleotide variation is common between allelic loci on homologous chromosomes, i.e. allelic variation (Smith 1998). Our MTP BAC clone consensus sequence assembly while representative of the AFGP/TLP locus may not represent sequence data from the same chromosome thus some variation between allelic sequence on homologous chromosomes is expected.

High sequence identity (>90%) between shared overlapping regions were found for DmBAC42/78 and 10/64. Overlapping regions between DmBAC 42/78 (12.1 kbp) and 10/64 (77.1) did not display allelic variation in gene sequence (100% sequence identity). DmBAC 10/64 overlap was predicted by FPC (Figure 3). Although DmBAC42/78 overlap shares high sequence identity in overlapping regions including one AFGP gene with 100% sequence identity, DmBAC 78 and 42 were not predicted to overlap. FPC predictions are based on AFGP/TLP positive BAC clone *Hind*III digest patterns (Figure 3). The DmBAC 78/76 overlap contained 3 *Hind*III digest sites which produced two shared bands (5.8 and 3.9 kbp). The small number of *Hind*III digest sites within the DmBAC 78/42 overlap most likely affected FPC detection of overlapping regions between these BAC clones. Identical shared AFGP gene sequence between overlapping regions contained on the same chromosome was used to further verify DmBAC clone overlap and indicated BAC clones to be on the same chromosome as all paralogous and allelic AFGP genes within the AFGP/TLP locus contained unique distinguishing nucleotide variations (with the exception of pseudogenes DmBAC 76 A6 and DmBAC 64/10 A6 which are identical).

DmBAC 76 Spans the AFGP/TLP MTP Gap

The alignment of all AFGP/TLP BAC shotgun subclone consensus sequences by shared sequence identity was unable to resolve a gap of unknown size within the AFGP/TLP locus. An overlap between DmBAC 78 and 10 could not be found (Figure 9). DmBAC 78 and 10 do not appear to overlap based on lack of sequence identity and the large amount of variability in AFGP gene copy number between the two consensus alignments. MOBC paired end sequence alignment within the AFGP/TLP locus consensus sequence was performed to identify BAC clones that may span this gap. MOBC paired end sequence alignment within the DmBAC 78 consensus sequences showed these clones partially overlap and additional sequence unique to DmBAC76 fell within the DmBAC 10/78 gap (Figure 10). A DmBAC 76 shotgun subclone library (1-5 kbp) was sequenced to 4.6 X coverage. All shotgun subclones were sequenced from both ends and pair end matched to confirm shotgun subclone assembly. Due to the high amount of predicted overlap between these two clones and the difficulty in aligning

AFGP gene rich regions, DmBAC76 is partially assembled to determine gene content only. DmBAC 76 contains 7 complete AFGP genes, one 5’ truncated AFGP pseudogene (76_A6), and seven trypsinogen III pseudogenes (data not shown). DmBAC 78 and 76 AFGP genes and flanking sequence located in the DmBAC78/76 overlap share 100% sequence identity with each other (data not shown). The DmBAC 76 1-5 kbp shotgun subclone assembly is missing one gene present within DmBAC78 (A9-2) (Figure 10). The discrepancy between DmBAC 76/78 gene content may have been caused by sequence misassembly. The DmBAC 78 *Not*I enzymatic digest and shotgun sequence assembly size varied by approximately 11.4 kbp which is the approximate size of this gene containing region. Although pulsed field gel size predictions of large inserts are not exact, estimates of clone sizes for many clones were predicted with a fair amount of accuracy. Attempts to resolve the DmBAC 78 and 76 AFGP gene content discrepancies by independent reassembly of DmBAC 78 and 76 shotgun subclone library sequences were unsuccessful. Although gene number varies by one between these two overlapping contigs, 100% sequence identity between all other putative AFGP gene homologs and flanking sequence (7 AFGP genes) suggested these regions were homologous. Variation in gene number was most likely a result of shotgun subclone sequence misalignment that is unable to be resolved by the current alignment techniques used.

Our Individual D. mawsoni AFGP/TLP Genomic Locus Is Represented by Two Haplotypes with Variable AFGP Gene Copy Number

The *D. mawsoni* AFGP/TLP locus is represented by two haplotypes, arbitrarily termed haplotype 1 and 2 (Figure 10) (Figure 11). Multiple DmBAC clones were predicted to span both AFGP loci by FPC and the alignment of MOBC paired end sequences within the AFGP/TLP locus consensus sequence (Figure 3) (Figure 10) (Figure 11). This indicates that the variation observed between the two detectable *D. mawsoni* AFGP loci are real and not the product of sequence misassembly as multiple clones are predicted to span both AFGP gene containing regions by two independent methods. None of the MOBC clone paired end sequences alignments within DmBAC shotgun subclone assembly consensus sequences supported a DmBAC 76/10 overlap. DmBAC 35 was linked to both DmBAC 10 and 42. MOBC paired end sequences alignments of DmBAC 5, 7, 25 and 34 suggested an overlap between DmBAC 35 and 42. DmBAC 35 is also suggested to overlap with DmBAC 10 by DmBAC 65 paired end sequence alignments (data not shown). This suggests that sequence sharing similarity to DmBAC 35 is located upstream of both DmBAC10 and 42 in the AFGP/TLP locus alignment.

Similarly DmBAC 76 and 64 were both linked to DmBAC 39 (Figure 11). Paired end sequence alignment of DmBAC 60 and 23 supported DmBAC 64/39 overlap while DmBAC 18 and 38 paired end sequence alignment supported a DmBAC 76/39 overlap. FPC predicted DmBAC 39, 60, 23, 18, and 38 share sequence identity (Figure 3). DmBAC 64 and 76 are distinctly different BAC clones which do not share high sequence identity to each other and most likely do not overlap (data not shown). This suggests that sequence sharing similarity to DmBAC 39 is located downstream of both DmBAC 64 and 76. In light of the MOBC paired end sequence data the AFGP/TLP DmBAC shotgun subclone assembly consensus sequence alignment does not appear to contain a gap. Instead, our locus sequence alignment coupled with MOBC paired end sequence data supported the presence of two AFGP/TLP locus haplotypes containing variable AFGP gene copy numbers flanked by sequence with minor allelic variation.

As stated above allelic variations between the DmBAC 35 and 42 overlap (data not shown) indicates these two clones are on homologous chromosomes. DmBAC 35 shotgun subclone sequence is thus placed in front of DmBAC 10 in haplotype 2 (Figure 11). However as MOBC paired end sequence indicated sequence similar to DmBAC 35 is located upstream of both BAC clones we assume sequence similar to this is located upstream of DmBAC42 (Figure 10). Allelic variation between DmBAC 64/39 overlapping sequence suggests these clones are located on homologous chromosomes as well. DmBAC 39 shotgun subclone consensus sequence is thus assigned to the region downstream of DmBAC 76 in haplotype 1 (Figure 10). However we again assume that downstream sequence is similar based on FPC *Hind*III digest alignments and MBOC paired end matching. Due to the expense of sequencing additional BAC clones, redundant sequence data for these regions was not obtained for both haplotypes. Based on allelic variation detected between BAC clone shotgun subclone assembly consensus sequence overlapping, regions haplotype 1 (H1) consists of DmBAC 42, 78, 76, and 39 (Figure 10) and haplotype 2 (H2) consists of DmBAC 35, 10, and 64 (Figure 11).

Further verification of DmBAC haplotypic presence was obtained by phylogenetic analysis of all the AFGP genes from sequenced BAC clone assemblies. This was accomplished by construction of a Neighbor-Joining tree from the intron 1 and signal peptide sequence of all the AFGP (A) genes from all sequenced BAC clones (Figure 12). All AFGP genes were named according to the length of the last polypeptide encoded in exon 2, i.e. A8 (Figure 11). Overlapping regions sharing identical gene content should share a most recent common ancestor within our phylogenetic analysis. Alleles between haplotypes should also cluster closely together, however not as closely as identical gene sequence. DmBAC 42_A8 and 10_A8 genes are alleles, termed H1_A1 and H2_A1 respectively. DmBAC 10_A11 and 64_A11 are homologous and termed H2_A3. The allele of H2_A3 was found to be H1_A2 (DmBAC 42 A11). DmBAC 76_A8 and 78_A8 genes are homologous and termed H1_A5 in the physical map of haplotype 1 (Figure 10). DmBAC 42 A9 (H1_A3) and 78 A9 (H1_A4) do not have detectable homologs (consistent with DmBAC clone overlap), or alleles however these genes are indicated to share close sequence identity with H1_A5 (Figure 12). DmBAC 78 A4 (H1_A6) is homologous with DmBAC 76 A4; however A4 was excluded from our analysis because of incomplete intron 1 sequence (data not shown). This gene is termed H1_A6 in the physical map of haplotype 1 (Figure 10). DmBAC 78 A9-2 is termed H1_A7. This gene is missing from DmBAC 76 as discussed above. DmBAC 76 and 78 A10 are homologous and termed H1_A8. The H2_A7 is homologous to DmBAC 10 A8-2 and 64 A8-2. DmBAC 78 and 76 10-2 (H1_A9) are homologous. DmBAC 64 and 10 A7 genes (H2_A4) are also homologous; the allelic gene is DmBAC 76 A7 (H1_A10) in the physical map of haplotype 1 (Figure 10). The lack of haplotype 2 AFGP alleles for haplotype 1 AFGP genes H1_A6, H1_A3, H1_A5, H1_A4, and H1_A9 may indicate a recent AFGP gene duplication in haplotype 1 or a deletion event in haplotype 2.

Our AFGP NJ gene tree (Figure 12) again supports the presence of DmBAC clone overlapping regions for DmBAC 42/78, 78/76, and 64/10. AFGP genes located within DmBAC 10 were suggested to be allelic to those in 42. DmBAC 10/64 AFGP alleles are located in DmBAC 76. Homologous AFGP genes represented in multiple DmBAC clones (determined by phylogenetic analysis and shared BAC clone overlap within the same chromosome) are represented in each haplotype physical map by one gene within their shared overlapping region.

AFGP/TLP genomic locus gene content, shared sequence identity in regional BAC clone overlap, and MOBC paired end sequence alignments within the AFGP/TLP consensus haplotype sequences all concur. The AFGP/TLP genomic locus appears to contain two haplotypes with AFGP copy number variability. This variability in AFGP gene copy number appears to have hindered accurate FPC predictions of DmBAC clone overlapping sequence. The correct alignment was determined by shared sequence identity between overlapping DmBAC shotgun subclone consensus sequences. This was paired with MBOC paired end sequence alignment within the AFGP/TLP locus consensus sequence for verification of previously identified and discovery of new DmBAC clone overlapping regions.

CHAPTER 2

THE GENOMIC MECHANISMS OF ANTIFREEZE GLYCOPROTEIN GENE BIRTH AND FIXATION IN AN ANTARCTIC NOTOTHENIOID FISH IN RESPONSE TO ENVIRONMENTAL SELECTIVE PRESSURES.

ABSTRACT

High serum concentrations (5-35 mg/mL) of antifreeze glycoproteins (AFGPs) are the single most important factor in Antarctic notothenioid survival in frigid Antarctic waters. High circulating serum levels of AFGPs bind to ice crystals that enter the fish. AFGPs prevent ice crystal growth, and thus freezing (death) of the fish. We characterized the AFGP genomic locus from the Antarctic notothenioid to determine the genomic architecture underlying this vital cold adaptation. Bacterial artificial chromosome (BAC) library construction and sequence reconstruction of the AFGP/TLP genomic locus in the Antarctic notothenioid, *Dissostichus mawsoni*, shows the AFGP genomic locus spans 400-500 kbp and is represented by two haplotypes that vary in the number of AFGP genes. The AFGP genes are associated with digestive (7) and cold active (4) trypsinogen as well as TLP (2) and AFGP/TLP chimeric (3) genes. Association of AFGP genes with TLP and AFGP/TLP chimeric genes further supports AFGP's TLP ancestry. Phylogenetic and genomic sequence analysis of locus genes shows the AFGP lineage arose from a single AFGP/TLP chimeric gene through slipped misalignment during genome replication. Identification of genomic breakpoint regions indicate, the ancestral AFGP gene appears to have been extensively duplicated as a result of segmental duplication (14 tandem duplicates) by unequal crossing over at a

recombination hotspot associated with all AFGP gene 3' flanking sequences. The increase in AFGP gene copy number may have been selected for due to the increased fitness afforded Antarctic notothenioids in the form of increased AFGP serum concentrations. Characterization of AFGP and AFGP/TLP chimeric coding sequence shows that larger AFGP peptides, observed in the serum of *D. mawsoni*, are only encoded within AFGP/TLP chimeric genes. This suggests the AFGP/TLP chimeric gene may function as an AFGP in present Antarctic notothenioids. These findings are supported by AFGP/TLP chimeric transcript tissue distributions, which are similar to previously identified AFGP transcript tissue distributions.

INTRODUCTION

The formation of the Antarctic Circumpolar Current (ACC) circa 25 MYA, isolated the waters of the Southern Ocean from surrounding temperate waters and sequestered local fish stock within it's boundaries (Eastman 1993). Southern Ocean temperatures decreased as a result of the ACC formation. The colder Antarctic temperatures resulted in massive faunal extinction and an adaptive radiation of the perciform suborder Notothenioidei (Kennett 1982; Anderson 1999). Presently the Notothenioidei suborder accounts for approximately 46% of current Antarctic fish species and greater than 90% of the Antarctic biomass present on the icy high Antarctic shelves (Eastman 2005).

Notothenioid survival in icy Antarctic waters is primarily attributed to the presence of antifreeze glycoproteins (AFGPs). AFGPs bind and inhibit the growth of ice crystals that enter the fish during activities such as feeding, thereby preventing the fish from freezing (DeVries 1971). Primarily synthesized in the stomach and pancreas, AFGPs are deposited into the intestine where they inhibit ice crystal growth within notothenioid hypoosmotic (compared to sea water) intestinal fluid (Cheng, Cziko, and Evans 2006). AFGPs are hypothesized to be transported by an unknown mechanism across the intestinal wall and circulate in the serum at 5-35 mg/mL (Jin and DeVries 2006), where they also inhibit further ice growth.

AFGPs in Antarctic notothenioids are composed of a glycotripeptide repeat (Ala-Ala-Thr)$_N$, with each threonine linked to a disaccharide of galactose-N-acetylgalactosamine (Cheng and DeVries 1991; Eastman 2000). AFGPs are encoded and translated as a polyprotein precursor of multiple AFGP peptides separated by a conserved three-residue spacer. The polyprotein is cleaved at the spacers by an unknown chymotrypsinogen-like protease into individual AFGP proteins, containing from 4 to up to about 86 Ala-Ala-Thr repeats (Figure13) (Hsiao et al. 1990; Chen, DeVries, and Cheng 1997b). Although larger repeats are known to be more potent inhibitors of ice growth, the smaller 4 and 5 repeat proteins predominate in the serum and intestinal fluid of all Antarctic notothenioids (O'Grady, Ellory, and DeVries 1982).

The AFGP coding region evolved from a single 9-nt element (Thr-Ala-Ala) within a functionally unrelated trypsinogen-like serine protease (TLP) gene, a cold-responsive digestive trypsin (Chen, DeVries, and Cheng 1997b). Expansion of this 9-nt non-coding element located in the intron1/exon 2 junction of the TLP gene resulted in expansion of the TLP exon 2 coding sequence to include the AFGP tripeptide repeat forming an evolutionary intermediate termed AFGP/TLP chimeric gene (Chen, DeVries, and Cheng 1997b). The AFGP/TLP chimeric gene is persistent in the genome and transcriptome of the modern notothenioid *Dissostichus mawsoni*; however its function if any is unknown. Evolution of the AFGP gene resulted from recruitment of the 5' and 3' ends of the AFGP/TLP chimeric gene, as well as deletion of most of the TLP coding region by an unknown mechanism (Chen, DeVries, and Cheng 1997b; Cheng and Chen 1999). We have sequenced the AFGP/TLP genomic locus from the Antarctic notothenioid *D. mawsoni.* Our analysis shows the AFGP gene family is quite large (22 genes in one individual). Extensive AFGP gene duplication appears to have been the result of segmental duplication. Duplication was driven by a recombination hotspot located downstream of all AFGP genes. Locus analysis furthers knowledge of the mechanism of AFGP gene genesis from a functionally unrelated TLP gene and identifies the genomic architecture underlying the rapid Antarctic notothenioid AFGP cold adaptation.

MATERIALS AND METHODS

Specimen Collection Purification of D. mawsoni AFGP Serum Proteins

D. mawsoni specimens were captured from McMurdo Sound, Antarctica, and kept in flow through seawater aquarium facilities at the Crary science center at McMurdo Station. Blood was collected from the caudal vein of anesthetized specimens and allowed to clot on ice or at 4°C for approximately 2 hours. The sample was centrifuged and sera removed and stored at -80°C until used. Serum samples from multiple *D. mawsoni* specimens were separately treated with 5% trichloroacetic acid (TCA) and centrifuged. The TCA soluble AFGPs were dialyzed to remove TCA, lyophilized, and resuspended in the original serum volume. A five microliter volume of AFGPs were fluorescently labeled with flurorescamine (Roche) and electrophoresed on a non-denaturing 10-15% gradient polyacrylamide gel as described previously in Chen *et al.* (1997).

D. mawsoni Alignment of Haplotype 1 and 2 Consensus Sequences.

Haplotype one and two consensus sequences from Figure 10 and 11 were aligned. MOBC haplotype overlapping regions, discussed in chapter 1, are included below the haplotype physical map alignment (Figure 14A). MOBC clones are color coded (pink and blue) to indicate their corresponding haplotype. Gene presence and orientation is indicated as a fully colored arrow facing the 3' end of the gene. Pseudogenes are shown as broken arrows. Arrow color corresponds to gene type according to the key. Figure 10 and 11 haplotype consensus sequences are aligned to each other by shared gene identity and gene orientation for allelic trypsinogen, TLP, and AFGP/TLP chimeric gene containing regions. Trypsinogen, TLP, and AFGP/TLP chimeric allelic regions were identified based on MOBC paired end sequence overlapping regions described in Chapter 1. AFGP allelic regions were aligned based on neighbor-joining phylogenetic inference of AFGP signal peptide and intron 1 sequence as described in the proceeding section. Allelic regions represented in both haplotypes by sequence data are indicated in the consensus bar below the physical map of haplotype alignment by a black bar. Regions represented by sequence data from only one haplotype are indicated by a consensus bar corresponding to the haplotype color (blue or pink). Putative gene content in the allelic haplotype region is indicated by an arrow outline corresponding in color to gene identity.

Phylogenetic Analysis

Phylogenies for each analysis were determined using the Maximum Likelihood (ML) and Bayesian algorithms implemented in PhyML and Mr. Bayes (Huelsenbeck and Ronquist 2001; Swofford 2001; Guindon and Gascuel 2003). All trees shown represent 3 independent runs. Each ML tree run was performed with 1000 bootstrap replicates, starting tree=BioNJ tree, and optimized during likelihood analysis. Bayesian analyses represent 4 independent runs of 1 million generations each. The models used were based on Akaike Information Criteria selected by Model Test (Posada 1998). AFGP SP /I1 trees were constructed using ML analysis under the model F81+I. AFGP/chimeric 3' trees were constructed under the model HKY+I. The AFGP, TLP, and AFGP/TLP chimeric phylogeny was generated using ML and Bayesian inference under the model

F81+I+G in PhyML. Bayesian analyses contained four partitions SP- JC; intron 1- F81+G; exon 6- JC+I; 3' flanking sequence –F81+I

RNA Extraction and cDNA Synthesis

Purification of total RNA from brain, spleen, head kidney, caudal kidney, pancreas, gill, intestine, and liver obtained from *D. mawsoni,* was performed using Tri reagent (Molecular Research Company, Cincinnati, OH). Total RNA was obtained for 3 separate organs when available. Tissue samples from separate individuals were available for skin, gill, brain, spleen, liver, pancreas, caudal kidney, and head kidney. Stomach samples consisted of 3 separate regions of a stomach sample from the same organism, and intestinal samples consisted of the anterior and posterior portion from the same individual. First strand synthesis was achieved using 1 μg of total RNA according to the manufacturer's suggested protocol (Superscript, Invitrogen). Each representative cDNA pool was stored at -20ºC until required.

Primer Design, PCR Optimization, and Product Purification

PCR systems were designed to amplify a 200 bp region specific to AFGP/TLP chimeric transcripts**.** Primer sequences were: F 5'-AACAGCTGCAACAGCTGCAGTTC-3′ and R 5'-CTGCATGGAATAGGGATTACTTGTACCAACAG-3′**.** The reverse primer will detect both AFGP/TLP chimeric and TLP transcripts; however the forward primer is specific for AFGP/TLP chimeric genes. Primer sequences for beta actin can be found in Table 3. All PCR products were amplified using 0.5 μL of first strand cDNA template according to the manufacturer's instructions (TaKaRa, Otsu, Shiga, Japan). The following thermal cycling conditions were used: Cycle 1: 95ºC for 15 min (X1); Cycle 2: 95ºC for 15 sec; 60 ºC for 15 sec (X35). All PCR reactions were conducted in 0.2 mL tubes using the MJ research Peltier Thermocycler 2100. PCR products were purified using Qiaquick PCR purification spin columns according to the manufacturer's' instructions (Quiagen, Valencia CA). Product identity and primer specificity were verified by sequencing (Big Dye v3, Invitrogen). Sequence reactions were read on the ABI3730*xl* sequence analyzer (Applied Biosystems) by the Sequencing unit at the Keck Center for Comparative and Functional Genomics (UIUC).

RESULTS AND DISCUSSION

D. mawsoni AFGP/TLP Locus Characterization

We have sequenced the AFGP/TLP containing portion of the Antarctic notothenioid *D. mawsoni* genome. Figure 14A shows a physical map of the AFGP/TLP genomic locus from one *D. mawsoni* individual. The AFGP/TLP genomic locus physical map consensus sequence was determined from the alignment of AFGP/TLP DmBAC shotgun subclone consensus sequences within the two detectable haplotypes (DmBAC clones in bold type). The *D. mawsoni* AFGP/TLP genomic locus within this individual is represented by 2 haplotypes termed haplotype 1 (500 kbp) and 2 (400 kbp) (Figure 14A). The AFGP locus is highly gene rich consisting of 5 gene types: 7 trypsinogen I (clade I), 4 trypsinogen III (clade IIIA), 2 TLP (clade IIIB), 3 AFGP/TLP chimeric, and 14 or 8 AFGP genes per haplotype. All AFGP, TLP, and AFGP/TLP chimeric gene types, not indicated to be pseudogenes, shared high sequence identity in coding regions and conservation of intron/exon structure (Table 2). AFGP and AFGP/TLP chimeric gene exon 2 (E2) $(AAT)_n$

coding repeat regions are unable to be aligned with confidence, therefore percent identity within this region could not be determined. However all E2 regions are free of frame shift mutations. This suggested all are intact genes may be capable of producing viable protein products.

AFGP gene containing regions (and extensive haplotypic variability) were confined to a single locus spanning approximately 300 kbp in haplotype 1 and 200 kbp in haplotype 2. Variability in AFGP locus size between haplotypes was attributed to AFGP gene copy number variation. AFGP gene copy number variability was localized to the central region of both AFGP haplotypes (Figure 14). Haplotypic variability within our individual genomic locus is likely representative of genomic diversity within the *D. mawsoni* species. AFGP serum phenotypic diversity was observed across *D. mawsoni* individuals, and agreed with our hypothesis of genotypic diversity (Figure 13). Many studies have shown that increased genomic diversity is beneficial to adaptation (Reusch et al. 2005; Kalbe et al. 2009). The rapid evolution of the AFGP gene family, driven by environmental selective pressures, may be facilitated by the diversity seen in *D. mawsoni* AFGP haplotypes (Figure 14A).

Although the AFGP gene is known to have arisen from a TLP precursor through an evolutionary intermediate (AFGP/TLP chimeric gene), AFGP gene genesis by a large deletion within an AFGP/TLP chimeric precursor limits the amount of shared sequence for reconstructing the phylogenetic relatedness of the TLP, AFGP/TLP chimeric, and AFGP gene types (Chen, DeVries, and Cheng 1997b; Cheng and Chen 1999). Shared sequence between the three genes was limited to the signal peptide and intron 1 (I1) region which varies from 300bp in TLP to as large as 6 kbp in AFGP/TLP chimeric genes. AFGP/TLP chimeric I1 sequence between the two AFGP/TLP chimeric gene loci varied from 6 kbp (H1/H2-C1 and 2) to 5 kbp (H2 C3). AFGP/TLP chimeric intron 1 variability appeared to be the result of large insertion/deletion events after AFGP gene genesis. Such variability as well as loss of the H2_C3 upstream trypsinogen 3 gene may aid in locus stability by preventing deletion of the entire AFGP gene containing region by interaction between the AFGP/TLP chimeric genes located on either side of the AFGP gene containing region (Figure 14A). Loss of the AFGP gene containing region would have a detrimental affect as this would most likely reduce the total amount of AFGP serum protein produced by the fish. Reductions in AFGP serum concentrations may reduce the fitness of this organism in ice laden Antarctic waters.

AFGP Gene Genesis from an AFGP/TLP Chimeric Intermediate

Sequencing of the AFGP/TLP genomic locus allowed determination of the mechanism of AFGP/TLP chimeric to AFGP gene transition. AFGP has been shown to evolve from a functionally unrelated TLP gene, via an ancestral intermediate (AFGP/TLP chimeric gene) (Chen, DeVries, and Cheng 1997b; Cheng and Chen 1999). The AFGP lineage is hypothesized to have evolved by a deletion of E2 through I5 within an AFGP/TLP chimeric precursor. Comparisons of AFGP/TLP chimeric genes at their intron 5/exon 6 junction (3' deletion site) indicated that all genes contain an 18 nucleotide region that shared high sequence identity to the E2 coding region (Figure 15). Interaction of the template strand I5 region with E2 coding sequence on the nascent strand during genome replication would result in deletion of the trypsin containing region of the AFGP/TLP chimeric gene, and yield the AFGP gene structure observed within the locus. Highly repetitive regions (AFGP E2) are prone to DNA polymerase stalls during genome replication (Viguera, Canceill, and Ehrlich 2001). Polymerase stall and dissociation on the AFGP/TLP chimeric E2 region during replication could have allowed the template strand I5 region interaction with nascent strand E2 sequence within the open replication bubble. Such an interaction may have resulted in the deletion of the trypsinogen containing portion of the AFGP/TLP chimeric gene and AFGP gene genesis on the nascent strand upon resumption of DNA synthesis. This deletion during meiotic genome replication would result in the newly formed AFGP gene being present on the newly synthesized sister chromatid. Such a mutation could have been selected for and fixed within the Antarctic notothenioid population. The AFGP/TLP chimeric genes are thought to encode for both AFGP and TLP. Active AFGP/TLP chimeric transcripts detected in *D. mawsoni* could be translated to produce viable trypsin and AFGP peptides. Increases in chimeric gene transcription would most likely increase AFGP and trypsin protein products. AFGP gene genesis and propagation would allow the production of valuable AFGP serum protein product (beneficial within icy Antarctic waters) within the *D. mawsoni* without affecting TLP protein abundance.

Determination of the AFGP/TLP Locus Expansion Mechanisms

TLP, AFGP/TLP chimeric, and AFGP genes shared a tandem conserved gene orientation of a trypsinogen III gene (TLP and AFGP/TLP chimeric) or trypsinogen III pseudogene (AFGP) in a head to head orientation. This suggested segmental duplication by unequal crossing over to be the mode of TLP, AFGP/TLP chimeric, and AFGP gene duplication (Figure 14A). Segmental duplication events are associated with rapidly evolving gene families and also are prevalent in genes encoding for protein products that interact directly with the outside environment such as genes encoding for immune response and olfactory receptors (Eichler 1998; Eichler et al. 2001; Tellam et al. 2009).

To determine the molecular mechanism driving the AFGP/TLP genomic locus expansion by segmental duplication we located the AFGP, TLP, and AFGP/TLP chimeric gene genomic breakpoints. Genomic breakpoints were defined as the exact genomic location where a gene rearrangement occurred. AFGP and AFGP/TLP chimeric genes shared downstream sequence identity with TLP 2 (Figure 16A). The TLP or AFGP/TLP chimeric (T2/C1 and 2 from both haplotypes) and H2_C3 downstream breakpoint was observed at a region of low complexity approximately 850 bp downstream of the TLP and AFGP/TLP chimeric stop codon (Figure 16A). H2_C3 downstream sequence after this breakpoint shared sequence identity with all AFGP genes for an additional 1.5 kbp. This suggested the H2_C3AFGP/TLP chimeric gene shared a most recent common ancestor to the AFGP lineage. Subsequent AFGP/TLP chimeric gene duplication after this gene rearrangement and AFGP genesis by slipped mis-alignment (Figure 15) resulted in all AFGP genes sharing downstream sequence identity with this AFGP/TLP chimeric gene (H2_C3).

All AFGP, TLP and AFGP/TLP chimeric genes share this 850 bp downstream region in common as a result of multiple rounds of segmental duplication in which this sequence was duplicated with each gene. The transmission of this downstream sequence with each AFGP/TLP/chimeric gene allows determination of the AFGP gene family phylogeny using shared gene sequence (signal peptide, I1, and E6) and the 850 nucleotide downstream sequence shared between AFGP, TLP, and AFGP/TLP chimeric genes.

Maximum likelihood and Bayesian inferences are consistent with breakpoint analysis, showing the AFGP lineage is monophyletic, arising from a H2_C3 type AFGP/TLP chimeric gene (Figure 17A). A maximum likelihood tree of the 3’ flanking sequence (approximately 2 kbp) within all AFGP genes and the ancestral AFGP/TLP chimeric gene type (Figure 17B) showed the same tree topology as the 5’ AFGP tree (Figure 14B). This indicated crossing over within the AFGP gene did not occur and validated the use of 3’ flanking sequence to determine the AFGPs monophyletic origin from a single AFGP/TLP chimeric precursor.

The Molecular Mechanism of AFGP Expansion

Determination of the AFGP segmental duplicate breakpoints revealed the molecular mechanism responsible for extensive AFGP gene duplication. The H2_C3 and AFGP gene breakpoint was detected 100 bp down stream of a line remnant. LINE elements create sequence homology at non-homologous genomic regions, thus resulting in gene rearrangements and/or duplication (Zhang et al. 2005; Lemaitre and Sagot 2008). A LINE mediated gene rearrangement in the ancestral H2 C3-like AFGP/TLP chimeric gene or primordial AFGP gene likely resulted in the deposition of a $(CA)_{30\text{-}80}$ dinucleotide repeat stretch downstream of the primordial AFGP or AFGP/TLP chimeric gene in the *D. mawsoni* AFGP/TLP locus. AFGP segmental duplicates were created and propagated by repeated double stranded DNA (dsDNA) breakage at the $(CA)_{30\text{-}80}$ stretch (Figure 16B). Pyrimidine/purine stretches are prone to dsDNA breaks due to the non-β helix DNA conformation formed at these sites resulting in the formation of secondary structures that facilitate DNA breakage (Tautz, Trick, and Dover 1986; Wichman et al. 1992; Shaw and

Lupski 2004; Rowen et al. 2005; Szamalek et al. 2005; Zhang et al. 2005; O'Driscoll and Jeggo 2006; Lemaitre and Sagot 2008).

Double stranded DNA breakage (two of which can be observed in the CA_N (pyrimidine/purine) stretch) downstream of the AFGP gene appear to have resulted in duplication of the AFGP and trypsinogen 3 pseudogene segment by non-homologous unequal crossing over (Figure 14) (Figure 16B). Local pairing of AFGP genes/trypsinogen III pseudogenes during homologous recombination most likely constrained the gene rearrangements/duplications to the same locus. Enrichment of gene density has also been shown to be positively correlated with high regional duplication frequencies most likely attributing to unequal crossing over events among genes (Zhang et al. 2005). The extensive amount of AFGP gene duplication observed within this locus strongly suggests there was selection for gene duplication.

AFGP Segmental Duplicant Variability Encourages Rapid Gene Expansion

The AFGP locus consists of two distinct AFGP containing segmental duplicate types (modules). The AFGP gene and upstream trypsinogen 3 pseudogene containing regions are separated by two different spacer sequences (Type 1 and Type 2). The AFGP gene and trypsinogen 3 pseudogene containing regions are duplicated along with either the Type 1 or Type 2 spacer sequence we define as a Type 1 or Type 2 AFGP gene containing module. Type 1 and Type 2 modules are for the most part separated spatially within the locus (Figure 14) (Figure 17B). AFGP genes cluster within two separate clades in the AFGP phylogeny corresponding to Type 1 and Type 2 modules, as well as spatial separation with the locus (Figure 14) (Figure 17)**.** Figure 17B suggests the Type 1 breakpoint was the first to occur, breakage and rearrangement resulted in Type 2 genesis later. Interestingly the boundary genes in the tree are also next to each other in the locus (Figure 14A), providing a logical progression of locus growth by segmental duplication still visible today.

Type 1 and Type 2 AFGP gene containing modules vary in size from 17.8-22.8 kbp and 19.7-23.7 kbp, respectively (Figure 18) (Figure 19). Sequence identity within module duplicates ranged from 95-98%. The large module size range was attributable to large insertions/deletions within module types (Figure 18). Sequence identity between modules types was >95% within the AFGP and trypsinogen III pseudogene containing portion. However very little sequence identity (<50%) is observed after the $(CA)_{30\text{-}80}$ (Figure 19). Tandem repeats created by segmental duplication are highly unstable within genomes due to the high sequence similarity (Eichler 1998). The creation of two AFGP gene containing modules with low sequence identity for >17 kbp, introduced variability within the AFGP gene containing tandem duplicates. This may allow further amplification of the AFGP gene while reducing gene loss between long stretches of highly similar tandem repeat sequences. Thus low sequence identity between type 1 and type 2 modules can most likely be attributed to the rapid expansion and large AFGP gene family within the *D. mawsoni* AFGP/TLP locus.

AFGP Coding Region Expansion Contributes to High Serum Concentrations and Locus Stability

High serum concentrations observed in Antarctic notothenioids is not only achieved by increased gene dosage (see previous) but also by expansion of the AFGP polyprotein coding region (E2) (Figure 20). Although E2 is thought to have originally formed by replication slippage of an individual nine nucleotide coding element in an ancestral TLP gene (Chen, DeVries, and Cheng 1997b), there is very little slippage of this kind observed between AFGP genes (Figure 20). The region appeared to expand independently in paralogous AFGP genes by duplication of entire polypeptide encoding and flanking linker sequence region(s). Shared tree topology between 5' and 3' sequence (Figure 14B) (Figure 17B) ruled out E2 expansion by unequal crossing over. Most likely the AFGP E2 expanded by slipped misalignment on the nascent strand during genome replication. Long stretches of direct repeats have been shown to be unstable and prone to contraction in many taxa (Wierdl, Dominska, and Petes 1997; Harr and Schlotterer 2000; Xu, Peng, and Fang 2000; Lai and Sun 2003). Contraction of longer AFGP repeats in Antarctic fishes would result in the loss of valuable AFGP coding sequence necessary for organismal survival. The introduction of a three residue linker sequence early in AFGP/TLP chimeric evolution could stabilize the long repeat sequences, reducing contraction (loss of AFGP coding sequence) by slippage misalignment.

Although incorporation of linker sequences into AFGP E2 increases protein product per transcript by providing stabilizing effects, it appeared to inhibit the formation of larger polypeptide encoding molecules (which are more potent). This trend of inhibition of replication slippage within imperfect repeats has been noticed in many non-coding microsatellite sequences (Petes, Greenwell, and Dominska 1997; Bacon, Farrington, and Dunlop 2000; Eckert and Hile 2009) and trinucleotide repeat coding sequence diseases (Richards and Sutherland 1997). Slippage inhibition by linker sequence contamination may explain the high abundance of $(AAT)_{4-5}$ (less potent) molecules observed in the AFGP

E2 coding region and serum protein (Figure 13). Facing the need for increased serum freezing point depression (conferred by high AFGP serum presence) in cooling Antarctic waters and inhibition of coding sequence expansion by linker sequence contamination, AFGP E2 was expanded by the duplication of smaller polyprotein molecules that although less potent confer vital serum freezing point depression.

Comparisons of linker sequence between AFGP and AFGP/TLP chimeric genes show more variability within AFGP linker sequences (Table 4). Although linker sequence origin is unknown, evaluation of AFGP and AFGP/TLP chimeric linker sequence content indicated the ancestral linker sequence to be LFF or LIF. Added linker variability along with varying polyprotein content in the AFGP E2 region may have increased stability of the highly unstable tandem repeats by reducing similarity between segmental duplicates. Exon 2 variability would also explain the lack of crossing over observed between AFGP genes (Figure 14B) (Figure 16). Highly repetitive regions are prone to crossing over and gene rearrangements. A reduction in crossing over at this crucial coding region would reduce AFGP pseudogenization that could lower organsimal fitness. AFGP E2 linker presence may serve a dual function, increasing AFGP protein serum content and providing stability to highly similar AFGP tandem repeats. AFGP coding sequence composition can also be used to explain the AFGP serum phenotype, expanding the scope of our model from gene birth, through fixation of the observed phenotype.

AFGP/TLP Chimeric Genes May Function in D. mawsoni as an AFGP Gene

Comparisons of AFGP and AFGP/TLP chimeric polyprotein content, encoded in E2, indicated AFGP genes encode primarily for smaller AFGP molecules (AAT_{4-5}), consistent with observed serum presence (Figure 13). AFGP E2 encoded for mature peptides ranging from 2-39 AAT repeats, however serum AFGP polypeptides ranged from 4-79 repeats across *D. mawsoni* sampled (Figure 13) (Figure 21). Creation of larger polypeptides by incomplete cleavage at the AFGP polyprotein coding sequence, was determined to be unlikely as L and F residues (putative linker cleavage site for chymotrypsin-like protein) are conserved ruling out the inability to cleave at these sites (Chen, DeVries, and Cheng 1997a). Larger mature peptides are encoded in the AFGP/TLP chimeric E2, suggesting a possible function for this evolutionary intermediate. AFGP/TLP chimeric gene transcripts (Figure 22) are primarily detected in the stomach and pancreas, similar to AFGP expression (Cheng, Cziko, and Evans 2006). AFGP/TLP chimeric transcripts (which also encode an active TLP protein) lack non-digestive tissue expression observed in TLP transcripts, suggesting AFGP/TLP chimeric genes were maintained in the AFGP/TLP locus for their ability to act as an AFGP providing the larger transcripts observed in the serum of these animals (Figure 13).

CONCLUSIONS

We propose a mechanism by which the AFGP/TLP locus expanded by segmental duplication of AFGP, TLP, and AFGP/TLP chimeric genes (Figure 23). AFGP gene association near a recombination hotspot, and environmental selective pressures increased AFGP gene and serum concentration. The instability of highly similar tandem repeats was most likely circumvented by the creation of two dissimilar AFGP gene containing modules and variability introduced within the AFGP gene by expansion of the E2 coding region independently between genes by duplication of small polypeptide encoding units and variable flanking linker sequence.

It is widely accepted that the high Antarctic notothenioid AFGP serum concentrations, were driven by the need for notothenioid serum freezing point depression, causally linking this phenotype (high serum concentrations) to environmental selective pressures (cooling/ice-laden Antarctic waters). Study of the AFGP genomic locus provided an opportunity to visualize genomic adaptation to environmental selective pressures. Collectively the *D. mawsoni* AFGP/TLP genomic locus provided a uniquely complete genomic snapshot of *de novo* gene genesis and fixation within an organism in response to environmental selective pressures (cooling Antarctic waters). The large amount of literature documenting the AFGP phenotype also allowed explanation, on a genomic level, of AFGP phenotypic observances and functional classification of the AFGP/TLP chimeric gene type.

CHAPTER 3

TRYPSIN EVOLUTION: A COMPARATIVE STUDY OF COLD-RESPONSIVE TRYPSINS IN ANTARCTIC NOTOTHENIOID AND TEMPERATE WATER TELEOSTS.

ABSTRACT

Trypsins, digestive enzymes present in a wide range of organisms, are traditionally classified into three groups. Group I and group II are common to all vertebrates. These trypsins are predicted to function as the main digestive trypsins. Group III trypsins are a class of teleost specific psychrophilic (cold active) trypsins. Placement of newly discovered Antarctic notothenioid *Dissostichus mawsoni* trypsinogens within the current vertebrate trypsinogen phylogeny has allowed new insight into vertebrate trypsinogen evolution and teleost trypsinogen functionality. Bayesian and maximum likelihood phylogenetic inference of trypsinogen transcripts, obtained from NCBI EST libraries, across the vertebrate trypsinogen lineage indicated the presence of four main clades, clade I (teleost group I trypsins), clade II (teleost group II trypsins (except in the Antarctic notothenioid *D. mawsoni*) and tetrapod group I and II trypsins), clade IIIA (teleost group III trypsins), and clade IIIB (teleost group III trypsins). Phylogenetic inference suggested teleost and mammalian trypsinogens classified as group I evolved independently of one another. Mammalian trypsinogens arose from teleost group II trypsinogens after the loss of teleost group I ortholog at the base of the tetrapod lineage. Phylogenetic and genomic inference indicated clade IIIA and B trypsins, are more closely related to clade I (group I) trypsins, and represent a clade of teleost specific cold adapted trypsinogens. Clade III trypsinogens appear to have a

warm climate origin, not conducive to their known role as a psychrophilic trypsin. Characterization of *D. mawsoni* trypsinogen types by quantitative PCR and tissue expression profiling by PCR suggested clade III trypsinogens may be recruited for their predicted psychrophilic capabilities in Antarctic notothenioids. The function of clade III trypsinogens in temperate fishes is unclear, however our tissue distribution data indicated that clade III trypsins may perform an additional unknown function unique to clade III trypsinogens which is unrelated to their predicted psychrophilic activity. Biochemical data for one clade IIIB trypsin (*G. morhua* Y), and shared residue variations between all clade IIIB trypsinogens indicates that all clade IIIB trypsinogens may contain dual trypsinogen and chymotrypsinogen activity and reduced affinity for trypsinogen substrates observed in *G. morhua Y*. The evolution of a trypsinogen with variable enzyme and substrate specificity is a more plausible evolutionary selective pressure, as opposed to cold adaptation which may drive clade IIIB evolution in a time of warm climate.

INTRODUCTION

The ice-laden Southern Ocean is an isolated ecosystem, in which faunal taxa are dominated by a single suborder, Notothenioidei (Eastman 2005). Notothenioid survival in inhospitable Antarctic waters (present day average temperature -1.86°C), is largely due to the presence of antifreeze glycoproteins (AFGP), which confer protection against the freezing of hypoosmotic bodily fluids (compared to seawater) (DeVries 1971; Cheng and DeVries 1991; Cheng, Cziko, and Evans 2006). The AFGP gene evolved from a functionally unrelated trypsinogen like protease (TLP) gene (Chen, DeVries, and Cheng 1997b; Cheng and Chen 1999). While sequencing the AFGP/TLP genomic locus in the Antarctic notothenioid, *D. mawsoni*, in addition to TLP we found multiple genes in the locus that share high sequence similarity to the trypsin family of proteases. This led to the characterization of these putative trypsinogen genes and determination of their relationship to TLP.

Predominantly found in the vertebrate gut, digestive trypsins are synthesized in the pancreas and secreted into the intestine (Jeohn et al. 1995; Cheng, Cziko, and Evans 2006; Lilleeng et al. 2007). Together with pepsin and chymotrypsin, trypsin is one of the three principle digestive proteinases that function to cleave dietary proteins (Barrett and Rawlings 1995). Trypsins, as with all S1 serine peptidases are characterized by the presence of a catalytic triad (His-57, Asp-102, and Ser-195) within the enzyme's activation pocket. The catalytic triad when coupled with conserved substrate binding sites affords characteristic trypsin activity (Hedstrom 1996). Although trypsin mainly functions as a digestive enzyme, their expression in non-digestive tissues (stomach, liver, brain, skin, etc.) (Murray et al. 2004; Manchado et al. 2008) led to the discovery of additional processing functions such as activation of membrane associated receptors (Knecht et al. 2007) and propeptide processing (Koshikawa et al. 1998).

Vertebrate trypsinogens have been classified into three major groups which evolved sometime after agnathan divergence (group I, group II, and group III) (Roach et al. 1997). Group I and group II trypsins are classified as the main digestive trypsins in vertebrates (Roach et al. 1997; Spilliaert and Gudmundsdottir 1999; Murray et al. 2004; Knecht et al. 2007; Manchado et al. 2008). Teleost group I trypsins represent the traditional class of psychrophilic (cold adapted) digestive trypsin relative to teleost group II trypsins and mammalian group I and II trypsins (Smalas et al. 1994; Gudmundsdottir and Palsdottir 2005). Biochemical characterization of group I trypsins, as with many psychrophilic enzymes, have shown higher catalytic efficiencies and decreased thermal stability than their mesophilic counterparts (Asgeirsson, Fox, and Bjarnason 1989; Smalas et al. 1994; Leiros, Willassen, and Smalas 2000; Asgeirsson and Cekan 2006). Proteins adapt to cold in many ways however one common theme is increased flexibility due to less densely packed structures. Molecular studies of group I trypsins found decreased hydrophobic and increased polar residues aid in flexibility (Leiros, Willassen, and Smalas 1999). Asgeirsson and Cekan (2006) associated Atlantic cod cold adapted group I trypsinogen increased activity with increased substrate affinity. Comparative amino acid studies of group III trypsins from various teleosts indicated group III trypsins may contain residues that also confer increased cold activity, by the opposite trends (increased hydrophobicity, increased aromatic residues, and fewer polar residues) (Roach 2002). Recombinantly expressed trypsin Y, a *Gadus morhua* (Atlantic Cod) group III variant has been shown to be active at environmental temperatures lower than other trypsins (teleost group II and psychrophilic group I), thus supporting the previous classification of group III trypsins as an extremely psychrophilic trypsin (Palsdottir and Gudmundsdottir 2007b). Group III trypsins are presumed to function as a digestive trypsin when temperate fishes migrate through cold water in which the main digestive trypsin (group I) functionality would be reduced (Roach 2002). However, quantitation of *G. morhua* Y (group III) and group I trypsinogen transcript indicated group III trypsinogen levels to be very low with respect to group I (1:1340) **palsodotir**. While group III trypsins may posses the capability to act as a cold-responsive trypsin in teleosts, it is not clear by their transcript expression levels that they function in such a manner within these fishes.

Dissostichus mawsoni and *Pagothenia borchgrevinki* are Antarctic notothenioid fishes, geographically confined within the icy waters of the Southern Ocean (-1.86°C year round) (Eastman 2000; Eastman 2005). Antarctic waters are significantly colder than temperate waters, providing an opportunity to study trypsinogens of a cold stenothermal teleost. The discovery of three distinct trypsinogen gene types associated with the AFGP/TLP genomic locus in the Antarctic notothenioid *D. mawsoni* prompted further investigation into the vertebrate trypsinogen phylogeny and expression in Antarctic notothenioids. Bayesian and maximum likelihood phylogenetic analysis indicated the AFGP/TLP genomic locus contained one group I trypsinogen (we termed clade I) and two group III trypsinogens (we termed clade IIIA and IIIB). We present evidence which suggests a non-psychrophilic origin and may suggest a non-psychrophilic function for clade III trypsinogens in most teleosts and recruitment in Antarctic notothenioid fishes for their psychrophilic capabilities.

MATERIALS AND METHODS

Identification of D. mawsoni Trypsinogen Gene Sequences

D. mawsoni trypsinogen gene sequences were obtained from sequence alignment and analysis of the AFGP/TLP genomic locus isolated from a Bacterial Artificial Chromosomal (BAC) Library representing 10X coverage of the *D. mawsoni* genome (Cheng CH *et al.*, unpublished data). Gene presence was detected by BLAST searches of genomic sequence against the NCBI nucleotide database. Amino acid sequences of each gene were obtained by manual splicing of exon sequences from genomic DNA and translation in the Clustal function of MEGA 4 (Tamura et al. 2007). Appropriately located GT/AG splice sequences were identified to delineate intron/exon boundaries**.**

Sequence and Phylogenetic Analysis

Vertebrate trypsinogen sequences were obtained from GenBank, using text and homology based searches of EST and nucleotide databases. Table 5 contains the accession numbers for each of the trypsinogens used in our study. Taxa were restricted to select families that encompassed agnathans, teleosts, and tetrapods. Different types of serine-threonine proteases (trypsins, elastases, chymotrypsins) have been shown to evolve at different rates and occlude meaningful phylogenetic inference (Roach et al. 1997). For this reason we have limited our analysis to the three groups of characterized digestive trypsins and their orthologs identified in NCBI vertebrate EST/nucleotide databases and our *D. mawsoni* trypsinogens. Trypsinogen orthologs were identified by high sequence identity with previously characterized trypsinogens from each group (I, II, and III). *D. mawsoni* trypsinogens were also identified by high sequence identity to teleost digestive trypsinogens.

Multiple sequence alignments of *D. mawsoni* and other vertebrate trypsinogens were performed using the default settings in the MEGA4 clustal function, while constraining the nucleotide sequence alignment with the translated amino acid alignment (Tamura et al. 2007). Nucleotide sequence evolutionary models were determined to be GTR+I+G by Modeltest v3.7 (Posada 1998). Phylogenetic trees were produced from this alignment using Bayesian (Mr. Bayes 3.1.2)(Huelsenbeck and Ronquist 2001) and Maximum likelihood inference (PhyML) (Guindon and Gascuel 2003). Four independent Bayesian analyses were run with flat priors, four million generations, four chains, sampling every 100 generations. Three independent maximum likelihood trees were run in PhyML with 1000 bootstrap replicates. All trees were rooted with *P. marinus* trypsinogen sequence. Bayesian and maximum likelihood tree topology were identical, thus the trees were combined four our study. Bootstrap values and posterior probabilities are combined on one representative tree. Node support cutoff was 0.85 and 70 for Bayesian and maximum likelihood trees respectively. Nodes with bootstrap scores below the cut off are indicted by the abbreviation NS (not supported), and unlabeled nodes indicate no support by either method.

Phylogenetic clade groupings were verified by identification of shared intron/exon structure for *Petromyzon marinus, Danio rerio, Gasterosteus aculeatus, Oryzias latipes, Takifugu rubripes, Tetraodon nigroviridis, Xenopus laevis, Gallus gallus, Mus musculus* and *Homo sapiens* genes corresponding to their homologous trypsinogen transcripts within each phylogenetic clade. Gene sequence was identified by searching the ENSMBL genome database of each genome with the respective trypsinogen sequence. Intron/exon splice site junctions were determined as indicated above in *D. mawsoni* trypsinogen gene sequence identification.

The *D. mawsoni* AFGP/TLP locus contained two of the three trypsinogen groups (I and III) previously classified by Roach. We identified the *D. mawsoni* trypsinogen syntenic locus in two teleosts (*D. rerio* and *G. aculeatus*) and *Xenopus* and the group II locus in their respective ENSEMBL genome data bases. Loci were identified by significant BLAST identity to *D. mawsoni* trypsinogens and genomic microsynteny observed between the *D. mawsoni* genomic locus and orthologous loci. The group II locus was identified by significant sequence identity to teleost group II trypsinogens.

Fish Specimen and Tissue Collection

Antarctic notothenioids *D. mawsoni* and *P. borchgrevinki* were captured from McMurdo Sound, Antarctica, and kept in -1.6ºC flow through seawater aquarium facilities at the Crary science center of McMurdo Station. Native tissue samples for *D. mawsoni* and *P. borchgrevinki* were collected from anesthetized specimens and kept frozen at -80°C or in 90% ethanol at -20°C until use.

Warm-acclimated *P. borchgrevinki* were held at 4° C for 16 weeks according to Jin *et al.* (2006). Samples were collected and stored as above.

Temperate water adult *O. mykiss* specimens were obtained from a local hatchery (Harrietta Hills Trout Farm, Harrietta, MI). Trout were maintained in a 2000 liter fiberglass tank with constant non-chlorinated 14˚C well-water circulation under 12:12 h light: dark cycles (University of Indiana- South Bend Center, South Bend, IN). Tissues were collected from anesthetized specimens and stored in 95% ethanol at -20ºC until use.

RNA Extraction and cDNA Synthesis

RNA was obtained from three separate individuals for all tissues and treatments used. Purification of total RNA from brain, spleen, head kidney, caudal kidney, pancreatic tissue, gill, intestine, and liver from *O. mykiss, D. mawsoni,* and *P. borchgrevinki* were performed using Tri reagent (Molecular Research Center, Cincinnati, OH). First strand cDNA synthesis was achieved using Superscript II (Invitrogen, Carlsbad, CA) and one microgram of total RNA primed with an oligodT_{30} primer stock (10 micromolar) in a twenty microliter reaction volume according to the manufacturer's suggested protocol. The synthesized cDNA from each tissue was then divided into five microliter aliquots and stored at -20ºC until required.

Quantitative Real Time PCR Primer Design

Forward and reverse primers were designed for *D. mawsoni* and *O. mykiss* trypsinogen genes using the default parameters of AutoPrime (http://www.AutoPrime.de**)** (Table 3)**.** At least one primer (corresponding to cDNA transcript) in each pairing spans an intron/exon junction within genomic sequence, eliminating the possibility of priming to potential contaminating genomic DNA. AutoPrime requires ENSMBL genomic sequence to create primer pairs and thus *G. aculeatus* trypsinogen genomic sequences were first used to determine primer sites. Homologous primer sites in *D. mawsoni* trypsinogen sequences were then identified by sequence alignment to *G. aculeatus* in the Clustal function of MEGA(Tamura et al. 2007). Primers corresponding to *D. mawsoni* and *O. mykiss* trypsinogen transcript were synthesized (IDT Coralville, IA). *D. mawsoni* contain AFGP/TLP chimeric gene which shares identical coding sequence to clade IIIB trypsinogens. Primer pairs for *D. mawsoni* clade IIIB trypsinogen transcripts represent transcript from two distinct gene types (clade IIIB and AFGP/TLP chimeric genes). Both gene types are presumed to be capable of producing functional clade IIIB transcript and protein product.

Relative Real Time Quantitative PCR

Primers were validated for use with real-time iCycler (Bio Rad, Hercules, CA) PCR by determining optimal primer concentrations, and amplification efficiency following the method described by the system manufacturer. Primers were added to a twenty five microliter total reaction volume as per the protocol provided by the manufacturer ABgene Absolute QPCR SYBR Green Mix (Thermo, Epsom, UK). Final concentration of each primer pair were determined based on maximal product amplification and optimal amplification efficiency and are as follows: Dmaw clade I 600 nM, Dmaw clade IIIA 200 nM, Dmaw clade IIIB 300 nM, β-actin 200 nM.

Relative qPCR reactions were performed as follows: one microliter of *P. borchgrevinki* stomach first strand cDNA was used for all reactions (obtained as described above), diluted 1:3. Reactions were carried out on the iCycler iQ Real-Time thermocycler (BioRad, Hercules, CA). Conditions were set to the following parameters for all primer pairs: One cycle of 95ºC for 15 min, and forty cycles of 95ºC for 15 sec and 60 ºC for 20 sec. Each qPCR reaction was performed in triplicate. The Ct (defined as the cycle number at which the fluorescence surpasses a user defined threshold level) was determined for each reaction.

Quantification of transcript abundance was performed using the ΔΔ Ct method of quantitation for anterior stomach RT cDNA from native and warm-acclimated *P. borchgrevinki* (diluted 1:3) (Livak and Schmittgen 2001; Pfaffl 2001). The Ct values of each run were normalized against β-actin. Normalized mean Ct values for native and warm-acclimated *P. borchgrevinki* anterior stomach samples are represented as the mean Ct value of three individuals ± standard error of the mean. Each value is plotted as mean fold difference of warm-acclimated sample relative to the native sample.

Absolute Real Time qPCR

Absolute quantitation was employed to determine exact copy number differences between trypsinogen types. Purified PCR product for each *D. mawsoni* trypsinogen gene, amplified from pancreatic RT cDNA samples, was subcloned into the pGEM T-easy subcloning system (Promega, Madison, WI). Plasmids were *Sca*I linearized, purified using the Qiaquick PCR purification kit (Qiagen, Valencia, CA), and quantitated spectrophotometrically via the Quant-it ds-DNA assay kit (Invitrogen, Carlsbad, CA). Copy number was calculated using the following equation (Whelan, Russell, and Whelan 2003):

$$\text{DNA copy number} = \frac{6.02 \text{ X } 10^{23} \text{ (copy/mol) X DNA amount (g)}}{\text{DNA length (bp) X } 660\text{(g/mol/bp)}}$$

Plasmids were then diluted into ten fold dilutions series ranging from 1 X 10^9 to 1 X 10^3 copies/microliter and used to construct standard curves for each trypsinogen type. Ct values of each dilution were measured in duplicate using real-time qPCR with respective trypsinogen primer sets to generate standard curves. Ct values were plotted against the log of their initial copy numbers. Correlation coefficients (>0.95) and PCR efficiencies were determined using the icycler iQ optical system software v 3.1 (Bio-Rad, Hercules, CA). Standard curves were run along with experimental reactions. One microliter of *D. mawsoni* pancreatic RT cDNA, diluted 1:20 was quantitated, qPCR procedure is the same as described in the preceding relative qPCR section.

Ct values generated by a standard curve (representing plasmid copy number concentrations) were compared to experimental sample (*D. mawsoni*) Ct values using the iQ optical system software v3.1 (Bio-Rad, Hercules, CA) to determine trypsinogen copy number. The pancreas of teleosts is not a discrete organ. The harvested pancreatic tissue is heterogeneous, containing variable amounts of adipose tissue. As a result cDNA preparations from separate individuals are composed of a variable mixture of pancreatic and adipose tissue cDNA. Therefore transcript quantities cannot be averaged across individuals and thus directly compared to each other. Transcript copy numbers for each trypsinogen type ± SEM was plotted for each sample from each individual.

Detection of D. mawsoni *Genomic and RT First Strand Clade II Trypsinogen*

D. mawsoni clade II genomic sequence was identified using primers specific for *G. aculeatus* group II trypsinogen NCBI transcript (Table 5). Primer sequences are as follows:

Gacu 5' UTR- GTAAAATCAGTAGGAGCATCATGAAGCAGC- 3'; Gacu exon 3F: 5'-GCGTCTGGGTGAGCACAACATTG -3'; Gacu exon 3R: 5'- GTTTCCCCATCCAGAGAT CAGACAGCT–3'; Gacu 3': 5'-TTAGTTGGAGGACATGGTGCTGCGGAT-3'. *D. mawsoni* clade II RT cDNA was detected using *D. mawsoni* clade II specific primers, determined from *D. mawsoni* genomic exon 3 through 3' sequence. Primer sequences are as follows: Dmaw clade IIF: 5'-GATCGTCTGAGGTGCC- 3'; Dmaw clade IIR: 5'-AGAGTCTCCCTGGCAGGAGTCCTT -3'. PCR products were purified using Qiaquick PCR purification spin columns according to the manufacturer's' instructions (Qiagen, Valencia, CA). Product identity and primer specificity were verified by sequencing (Big Dye v3, Invitrogen, Carlsbad, CA). Sequences were read on the ABI3730*xl* sequence analyzer (Applied Biosystems, Foster City, CA) by the Sequencing unit at the Keck Center for Comparative and Functional Genomics (UIUC, Urbana, IL). *D. mawsoni* group II primers for tissue distribution were determined according to genomic sequence for encompassing exon 3 and exon 4.

Tissue Distribution Pattern Determination in Antarctic and Temperate Teleost

In order to detect trypsinogen transcript presence/absence across our tissue series, primers and PCR conditions were optimized to produce fragments ranging from 124-187 bp specific for each trypsinogen group (Table 3). *D. mawsoni* qPCR primers were used for *D. mawsoni* and *P. borchgrevinki* tissue series. *O. mykiss* primers were made according to the method described in the preceding qPCR primer design section. Accession numbers for *O. mykiss* trypsinogen sequences used to generate PCR primers can be found in Table 5. All PCR products were amplified using 0.5 μL of RT template cDNA purified from *D.*

mawsoni, P. borchgrevinki and *O. mykiss* brain, spleen, head kidney, caudal kidney, pancreas, gill, intestine, and liver. First strand cDNA was added to a mixture containing 200 nanomolar final concentration primer and deoxynucleotide (dNTP) and 0.5 microliter of Taq polymerase, in a 50 microliter PCR reaction (TaKaRa, Otsu, Shiga, Japan). The following thermal cycling conditions were used: one cycle of 95ºC for 5 min, and 35 cycles of 95ºC for 15 sec, and 60 ºC for 20 sec. Ten microliters of each reaction were electrophoresed on a 2% agarose III gel (Amresco, Solon, OH) for photo documentation**.** PCR products were purified and sequenced as described in the preceding section.

RESULTS

D. mawsoni Trypsinogen Genes

Figure 24 shows the predicted amino acid translation of the 11 *D. mawsoni* trypsinogen genes identified from our sequence reconstruction of the AFGP/TLP containing portion of the *D. mawsoni* genome. All *D. mawsoni* trypsinogens contained intact intron/exon structures. These trypsinogen genes were determined to be free of frameshift mutations and stop codons that would most likley render the protein product inactive. All trypsinogens contained the conserved catalytic triad (His-57, Asp-102, and Ser-195) (numbering according to chymotrypsin), necessary for trypsin functionality (Hedstrom 1996). All twelve Cysteine residues responsible for the six disulfide bridges giving the trypsin its proper conformation were also conserved (de Haen, Neurath, and Teller 1975). Tyrosine 172, Glycines 216 and 226 as well as Aspartic acid 189 (chymotrypsin number system (Zwilling et al. 1975), known to be involved in trypsin specificity were also conserved among all *D. mawsoni* trypsins (Hedstrom 1996).

Glycine 226, present at the entrance of the specificity pocket, was replaced by a Valine residue in *D. mawsoni* TLP2. This may affect substrate binding/specificity, as the biochemical properties of these residues are quite different.

Classification of D. mawsoni Trypsinogens within the Vertebrate Trypsinogen Phylogeny Reveals Four Distinct Trypsinogen Clades.

In order to determine the relatedness of *D. mawsoni* trypsinogen sequences to each other and to vertebrate trypsinogens, a search for homologous sequences across vertebrate taxa (primarily centered in teleosts) was carried out. A total of 85 sequences (20 teleost species, as well as four tetrapod species) were identified across the three previously proposed trypsinogen groups (I, II, III) (Roach 2002) (Table 5). The translated amino acid trypsinogen sequences for all ESTs appeared to be functional with the exception of Ipun 4 which contained a mutation H57Q and S195A (chymotrypsin number system) in the catalytic triad(Figure 25). These mutations likely result in loss of trypsin activity of the protein product from this individual transcript. The Gacu 4 ENSMBL predicted amino acid translation of the genomic sequence (Figure 25) contained an S195L mutation (chymotrypsin number system) in the catalytic triad. This mutation was not observed in the NCBI *G. aculeatus* transcript (DW651310.1). This may suggest Gacu4 (clade IIIB) mutations are variable among *G. aculeatus* individuals.

The Bayesian and maximum likelihood tree topologies were congruent, thus our tree is from Bayesian Inference with maximum likelihood bootstrap replicate values added. Bayesian and maximum likelihood phylogenetic trees of the mature trypsinogen trypsinogen nucleotide coding sequence showed four distinct clades we termed I, II, IIIA and IIIB (Figure 26). Most nodes leading to the four major trypsinogen groupings within the Bayesian and maximum likelihood trees had good support indicated by Bayesian Posterior Probability (>0.86) and maximum likelihood bootstrap replicate (>70) values. Neighbor-joining analysis of predicted amino acid sequence produced the same tree topology (data not shown).

Gene intron/exon structure was available for all trypsinogens represented within Figure 26 for *P. marinus, D. rerio, D. mawsoni, G. aculeatus, T. rubripes, X. laevis, G. gallus, M. musculus, H. sapiens*. Analyses of the intron/exon structures showed the four major clade groupings (identified as clade I, II, IIIA, and IIIB in Figure 26) are supported by intron/exon structures and represent a parsimonious order of intron gain between clade groupings. The ancestral intron/exon structure is the 5-exon type, present in all *P. marinus* and clade II trypsinogens. Teleost fishes underwent two separate gain of intron events in clade I and IIIB trypsinogens. Clade IIIB and clade I differ in intron/exon structure. Clade I trypsinogens gained an intron in the fifth exon of the 5-exon ancestor (most closely related to clade II trypsinogens) to form the 6-exon trypsinogen observed in all teleosts. Clade IIIB trypsinogens gained an intron in the third exon of their 5-exon ancestor (most closely related to clade IIIA) to form the 6-exon trypsinogen observed in all teleosts.

P. marinus contains one trypsinogen type (5-exon), which was proposed to predate the trypsinogen duplication events of teleosts and tetrapods (Roach et al. 1997). The *P. marinus* sequence served as an outgroup, sharing a most recent common ancestor with 5-exon clade II trypsinogens. Teleost group II trypsinogens, as well as tetrapod group I and II trypsinogens, again previously classified by Roach (1997) were all contained within our clade II. Roach's teleost group I trypsinogens fall within our teleost specific clade I (Figure 26). All teleost fishes screened contained clade I and II trypsinogens. Teleost trypsinogens previously classified as a single group III, formed two subclades we termed clades IIIA (5-exon) and IIIB (6-exon). Fishes, within the orders Siluriformes to Tetraodontiformes contained at least one trypsinogen from each clade (four trypsinogen types), with the exception of medaka. Clade III trypsinogens could not be found via BLAST searches of medaka NCBI EST libraries or within its preENSEMBL genome assembly.

The *D. mawsoni* trypsinogens from the AFGP/TLP locus fell into two of the three previously characterized trypsinogen groups (I and III) put forth by Roach (1997). Seven trypsinogens (Dmaw 1a-g) clustered with teleost specific clade I trypsinogens (Roach's group I). Two of these trypsinogens 1d and 1 g were located within a long branch clade I subclade along with Gacu 1b and 1d (Figure 26). Transcripts corresponding to these *G. aculeatus* trypsinogens were not found in any of the *G. aculeatus* NCBI EST libraries. This may suggest that these genes in *G. aculeatus* and their orthologs in *D. mawsoni* are pseudogenes. Thus *D. mawsoni* may only contain five functional clade I trypsinogens. The remaining four *D. mawsoni* trypsinogens genes (not within clade I) clustered with other trypsinogens previously identified as Roach's group III trypsinogen type. This larger group III clade was composed of two smaller subclades we termed clade IIIA ((clade IIIA (contains Dmaw 3a-c), and clade IIIB (contains Dmaw TLP 1 and 2)). Clade IIIA has the ancestral 5-exon gene structure. The most recent common ancestor to clade IIIB gained an intron in the third exon of the ancestral 5-exon structure, leading to a 6-exon structure in the lineage. Roach's group II trypsinogens, found in all vertebrates examined (except *P. marinus* which only has one trypsinogen type) were not detected within our *D. mawsoni* trypsinogen genomic locus.

Syntenic Locus Identification Supports the Vertebrate Trypsinogen Phylogeny.

We identified *D. rerio, G. aculeatus,* and *X. laevis* clade I, II and III trypsinogen loci within their respective ENSMBL genome databases that further support our vertebrate trypsinogen phylogeny, by gene presence or absence in associated loci. *D. rerio* and *G. aculeatus* digestive trypsinogen genes were distributed between two loci on chromosome 16 and group X, respectively (Figure 27). *G. aculeatus* clade I (1a-d), IIIA (3a), and IIIB (3b) trypsinogen genes were found in the same locus while teleost clade II trypsinogens (2a, 2b) were within a separate locus eight Mbps away. Gacu 2b was found to be a 5' truncated pseudogene (Figure 27). *D. rerio* clade I (two genes) and clade II (3a and 3b) trypsinogen gene loci were separated by 0.5 Mbp. *D. rerio* 3a and 3b trypsinogens share 61 % nucleotide identity. The *D. rerio* 3b trypsinogen EST was excluded from our phylogenetic inference because of low shared sequence identity to with other *D. rerio* digestive trypsinogens in our analysis. Incorporation of *D. rerio* 3b resulted in long branch attraction with group III trypsinogens (54-58 % nucleotide identity w/group III trypsinogens). No clade III genes were found close to *D. rerio* clade I trypsinogens nor were they detected elsewhere within the genome of *D. rerio*. Further BLAST searches of the *D. rerio* and *G. aculeatus* ENSEMBL genome databases with their respective digestive trypsinogen gene sequence did not reveal other loci containing the genes of interest. *D. mawsoni* clade I and III trypsinogen genes were present in 1 locus, consistent with the *G. aculeatus* syntenic locus. In all three species clade I (*D.*

rerio only clade I, clade I and III *G. aculeatus* and *D. mawsoni*) trypsinogen loci were associated with the translocase of the outer mitochondrial membrane 40 (TOMM40) gene (Figure 27).

The *Xenopus* ENSEMBL genome is not assembled into chromosomes, however we identified two trypsinogen loci, corresponding to teleost clade I and II trypsinogens, within the *Xenopus* genome on separate scaffolds. Scaffold 1481 contained four tandem trypsinogens sharing high nucleotide sequence identity to each other and previously identified *Xenopus* trypsinogen 10 (NP001011209) (Figure 27) (Table 6). Two complete genes, trypsinogen 10b and c, were included in our analysis and were associated with clade II (Figure 26). Scaffold 701 contained three additional tandem trypsinogens (NP_001011251) that share 93% nucleotide identity to previously identified *Xenopus* PRSS1, a digestive trypsinogen (NP_001011204). These trypsinogens were associated with the TOMM40 gene (Figure 27). PRSS1 TOMM40 microsynteny indicated these genes are the teleost clade I trypsinogen orthologs. These trypsinogens shared 54% nucleotide sequence identity with *Xenopus* trypsinogen 10b and c sequences, and were excluded from phylogenetic analyses due to low sequence identity and long branch attraction (52-54 % nucleotide sequence identity) with group III trypsinogens. *Xenopus* PRSS1 shares 61% nucleotide identity with *D. rerio* outlier 3b.

D. mawsoni Genome Contains Clade II Trypsinogen Pseudogenes

Teleost clade II trypsinogens and clade I (*D. rerio*) clade I/III (*G. aculeatus*) were found in separate genomic loci. As the clade II locus is not associated with the clade I/clade III locus in other teleosts, this region was most likely not represented in our *D.*

mawsoni sequence reconstruction of the clade I/clade III locus. Thus *D. mawsoni* genomic DNA and pancreatic RT cDNA were screened by PCR for the presence of clade II trypsinogen genes and transcripts. Genomic sequence containing *D. mawsoni* clade II 3' sequence (intron two through exon five) were amplified from genomic DNA using primers designed from *O. mykiss* clade II trypsinogens (Table 3). Efforts to amplify the 5' portion of clade II trypsinogens, which was highly conserved across teleost species, were unsuccessful (data not shown). This indicated the 5' truncation of this gene or extensive sequence modification. Primers specific for the exon three and exon five portion of clade II trypsinogens, according to *D. mawsoni* clade II genomic DNA sequence, were unable to detect expression in *D. mawsoni* pancreatic cDNA, further supporting pseudogenization of clade II trypsinogens within *D. mawsoni*.

Teleost Clade III Trypsinogens Contain Conserved Variations From Teleost Clade I Trypsinogens In Highly Conserved Loop Regions.

The discovery of two distinct clade III trypsins (A and B) prompted the reanalysis of clade III trypsin amino acid residue properties to reflect the presence of two clades. Molecular properties were assessed based on residue conservation of sequence alignments and the analysis of amino acid composition of catalytically active clade IIIA, IIIB, and I trypsinogens (Figure 25). The sequence alignment of the amino acid translation of teleost trypsinogens showed that clade IIIB trypsinogen S190A and K188S substitutions with respect to clade I (chymotrypsin numbering system) is absolutely conserved in clade IIIB. These variations are located in the trypsinogen loop 1 region, corresponding to residues 129-135 in our alignment (Figure 25). Loop 1 and 2 of serine proteases are extensions of the specificity pocket and indirectly control substrate specificity by altering substrate binding specificity or positioning of substrate within the catalytic triad (Hedstrom 1996). The S190A mutation was not conserved in clade IIIA trypsinogens (Figure 25) and does not appear to be associated with organismal environmental temperature. *D. mawsoni* and all teleosts contain an S/K222P variation conserved among clade III trypsinogens. Most clade III trypsinogens also contain a K/H224Y variation. Both these variations are shared with the basal *P. marinus* trypsinogen and located within the trypsinogen loop 2 region corresponding to residues 203-208 (Figure 25).

D. mawsoni Trypsinogen Expression Characterization

In an effort to characterize the expression patterns of the clade IIIA trypsinogen type and infer clade III trypsinogen functionality as a cold-responsive trypsinogen (by comparison to temperate fish data already present within the literature) we characterized digestive trypsinogen expression levels in native and warm-acclimated Antarctic notothenioid digestive tissue (pancreas) (Figure 28). Native *D. mawsoni* clade IIIA and IIIB pancreatic trypsinogen transcript copy number are significantly higher than clade I pancreatic transcript abundance within the three individuals tested. Trypsinogen transcript copy number is not significantly different between clade IIIA and IIIB trypsinogen transcripts (p>0.01 between clades I and III within individuals). Clade I: clade IIIA ratios are as follows: sample 1-1:25.9, sample 2-1:35.7, sample 3-1:23.2. Clade I: clade IIIB ratios are as follows: sample 1-1:24.8, sample 2-1:63.4, sample 3-1:57.2. While copy number between samples are variable due to teleost pancreatic anatomy (variable adipose tissue contamination) the trend of higher clade III transcript compared to clade I appears to be constant across all individuals. Clade IIIA and IIIB were not significantly different from each other within any of the three individuals tested.

We compared clade I, II, and III trypsinogen transcript tissue distribution patterns between temperate water trout (*O. mykiss*) and Antarctic notothenioids (*D. mawsoni* and *P. borchgrevinki*) by PCR analysis of cDNA from select tissues (Figure 29). O. *mykiss* clade I and II trypsinogens were detected in all tissues sampled; expression appears to be highest in pancreas. Clade IIIA and IIIB trypsinogen transcript in trout was primarily detected in stomach and pancreas, however low levels of expression can be detected in most other tissues sampled. *D. mawsoni* notothenioid clade I trypsinogen transcripts are detectable in the spleen, head kidney and pancreas only. Clade IIIA trypsinogen transcript is detected in all tissues tested, except the brain and liver. TLP (clade IIIB) transcript is expressed in all tissues but gill and caudal kidney. Clade II trypsinogen transcripts could not be detected in any of the *D. mawsoni* tissues examined. Antarctic *P. borchgrevinki* trypsinogen transcript patterns display the same general trends as *D. mawsoni* trypsinogens, i.e. clade III transcripts have a wider distribution pattern than clade I (Figure 29).

To further analyze trypsinogen type functionality with respect to cold, we assessed the effect of temperature on trypsinogen transcript expression in the Antarctic notothenioid *P. borchgrevinki* by relative qPCR (Figure 30). Pancreatic samples were not available for warm-acclimated (4°C) individuals, thus anterior stomach samples were analyzed. As stomach is a discrete organ, relative qPCR was carried out for stomach mucosa cDNA. Anterior stomach cDNA from warm-acclimated individuals were found to have clade I trypsinogen expression, while clade I transcript could not be detected in the majority (2 of 3) of unacclimated controls (-1.6°C) (Figure 30A). Anterior stomach samples of warm-acclimated *P. borchgrevinki* had significantly lower clade IIIA ($p<0.05$) and IIIB ($p<0.001$) trypsinogen transcript levels as compared to unacclimated samples (Figure 30B and C). Equivalent to a decrease of expression in warm-acclimated individuals by 75% (clade IIIA) and 94% (clade IIIB) compared to native samples (Figure 30).

DISCUSSION

The Vertebrate Trypsinogen Phylogeny

The association of three distinct trypsinogens with the Antarctic notothenioid *D. mawsoni* AFGP/TLP genomic locus prompted study of the Antarctic trypsinogen cold adaption with respect to their warmer water teleost counterparts. Our phylogenetic analysis indicated there to be four distinct types of teleost trypsinogens, clustering into 4 clades (I, II, IIIA, IIIB) (Figure 26). Previously trypsinogens were classified into three distinct groups by their inferred physiochemical properties and sequence homology (group I (anionic), group II (cationic), group III (cold active)) (Scheele, Bartelt, and Bieger 1981; Roach et al. 1997; Spilliaert and Gudmundsdottir 1999; Roach 2002). Teleost trypsinogens classified under Roach's group II and tetrapod groups I and II trypsinogens are present with in our clade II. Teleost trypsinogens corresponding to Roach's group I trypsinogens are found in our clade I. *D. mawsoni* and most teleosts appear to contain two types of group III trypsinogens we termed clade III A and B

(Figure 26), concurring with recent findings in the Senegalese sole (Manchado et al. 2008).

Due to variations between paralogous trypsinogen gene evolution rates and trypsinogen gene conversion in mammals, it was previously suggested that the vertebrate trypsinogen phylogeny could not be resolved (Roach et al. 1997). We resolved the vertebrate trypsinogen phylogeny through the exclusion of highly divergent trypsinogens (Xtro PRSS1 (a-c) and Drer 3b) from our phylogenetic analysis, and the use of genomic microsynteny and intron/exon structure to locate and classify divergent trypsinogens as well as verify trypsinogen phylogenetic inference (Figure 26). Gene conversion was not observed within the teleost trypsinogen phylogeny, most likely due to the separation of clade I and II genomic loci early in teleost evolution. Our phylogenetic analysis indicated the vertebrate trypsinogen lineage evolved from a single ancestral gene type (most similar to present day clade II trypsinogens), duplication of this gene some time before the teleost radiation, resulted in two trypsinogen genes (clade I and II trypsinogens) (Figure 26). Clade I and II trypsinogens persisted in *Xenopus* (Figure 27), however they appear to have been lost in reptiles (*G. gallus*) and mammals. Roach's mammalian group I and II trypsinogens (clustering within clade I of our tree) most likely arose from a clade II ancestral trypsinogen at the base of the tetrapod lineage independently of the teleost clade I (roach's group I) trypsins. Tetrapod and teleost development of Roach's group I trypsinogens (clade I- teleosts, clade II-tetrapods) occurred independent of each other. Further analyses should take in to account that teleost clade I trypsins, and tetrapod clade II trypsinogens (composed of Roach's groups I and II) do not share a most recent common ancestor, as all tetrapod trypsinogens most likely arose from a common clade II-like ancestral precursor.

Cationic and Anionic Trypsinogen Presence Is Conserved in Amphibians through Mammals

Roach's group I trypsinogens (teleost clade I trypsinogens) are usually but not always anionic, while Roach's group II trypsinogens (teleost clade II trypsinogens) are usually but not always cationic (Roach et al. 1997). Teleosts maintain a separation of predicted trypsinogen charge (clade I predicted pI <7, clade III predicted pI >7) between phylogenetic clades and loci, not observed in mammals (Figure 26) (Figure 27) (Table 6). Both clades I and II trypsinogens can be identified in *Xenopus* by sequence identity and/or genomic microsynteny at their respective trypsinogen loci. Examination of the amino acid translation of *Xenopus* clade II trypsinogens (Scaffold 1481;10b and c) (Figure 27) predicted charge at physiological pH shows the development of cationic and anionic trypsinogens (characteristic of Roach's group I and group II trypsinogen types) within the clade II locus (Table 6). Presumably high sequence variation (54% identity to *Xenopus* clade II digestive trypsinogen) in the amino acid translation of *Xenopus* clade I orthologous (anionic) trypsinogens (Scaffold 701; Xtro a,b,c) (Figure 27) resulted in the loss of *Xenopus* anionic digestive trypsinogens. The amino acid translation of *Xenopus* clade II trypsinogen charges appear to have drifted to maintain anionic digestive trypsinogen presence, both trypsinogens identified previously as group I and II (Roach et al. 1997) are within one locus. This locus is presumed by sequence identity to be orthologous to clade II teleost trypsinogens. Thus, cationic and anionic trypsinogens within tetrapods most likely evolved from a common ancestral clade II (cationic) trypsinogen precursor. Clade II (cationic) charge appeared to drift (Xtro 10b) (Table 6) possibly to maintain the presence of both cationic and anionic charged trypsins in *Xenopus* after the loss of the orthologous clade I trypsinogen. Two trypsinogens similar in sequence, but varying in charge are found in many tetrapods (Roach et al. 1997). The conservation of two distinct trypsinogen genes encoding different charged trypsin products may suggest an evolutionary drive for the existence of both anionic and cationic trypsins in teleosts and tetrapods.

Clades I and III Form a Large Teleost Specific Clade Composed of Three Distinct Trypsinogen Gene Types.

Within the teleost lineage clade III trypsinogens arose from a common ancestor to clade I trypsinogens (Figure 26). Clades I and III form a teleost specific clade of psychrophilic trypsinogens, most likely sharing some cold adapted properties related to trypsin functionality within an ectotherm living in cool waters (-2 to 24°C) compared to their mesophilic trypsin analogs (of endotherms) (37°C). Clade I trypsinogens appear to be more ancient evolving before Cypriniformes (*D. rerio* and *P. promephales*) divergence. Clade III trypsinogens, not found in genomic (*D. rerio*) (Figure 27) or transcriptomic (*D. rerio* and *P. promephales*) (Figure 26) screens of Cypriniformes, were detectable in the transcriptomes of Siluriformes *I. punctuatus* and *I. furicatus* (Figure 26). Clade III evolution can be place around the time of siluriform divergence roughly 150 mya. Siluriformes and other Otophysi originated around South America, ancient waters of this time were estimated to be comparable to tropical climates, where the driving force for the evolution of a cold active trypsinogen did not exist (Loyd 1984; Briggs 2005). The association of clade III trypsinogen evolution with a time of warm climate is not conducive with a putative cold adapted function attributed by Roach (2002). Clade IIIB cold adaptation observed in recombinant *G. morhua* Y, and predicted in all clade IIIB trypsinogens by shared sequence identity may then not likely to be the driving force that brought about clade III evolution. The large difference in tyrosine residues between clade I and III trypsinogens in loop regions and observed variations in hydrophilicity (Spilliaert and Gudmundsdottir 1999; Roach 2002) could alter protein loop conformations affecting enzyme substrate specificity or processing functions. It is not known why teleosts evolved three distinct trypsinogen types, however formation in warm climate may suggest evolution for variations in trypsin processing required for new food sources or some other function unrelated to cold.

Antarctic Teleost Trypsinogen Cold Adaptation

The Antarctic marine environment (-1.86 ° C) is colder than the observed temperature range of biochemically determined trypsin activity of temperate *G. morhua* and *S. salar* (Outzen et al. 1996, (Asgeirsson, Fox, and Bjarnason 1989; Asgeirsson and Cekan 2006) clade I trypsin (activity range: 4-65°C, optimal activity: 55°C), and recombinantly expressed cod Y trypsin, a clade III *G. morhua* variant (activity range: 2-30°C, optimal activity 21°C) (Gudmundsdottir and Palsdottir 2005). It is expected that *D. mawsoni* trypsinogens adapted to maintain activity at lower temperatures. This is supported by pyloric cecae purified clade I trypsin of the sub-Antarctic notothenioid

Paranotothenia magellanica which was shown to have higher catalytic efficiencies and reduced thermal stability compared to pyloric ceca purified *S. gairdneri* (temperate teleost) and *B. taurus* (mammal) trypsin orthologs (Genicot, Feller, and Gerday 1988). We compared our *D. mawsoni* genomic sequence and NCBI clade I *P. magellanica*, clade IIIB *N. coriiceps* nucleotide sequences to those of temperate and basal teleost clade III trypsinogens to identify properties that may be associated with Antarctic trypsinogen adaptation to extreme cold (Figure 25). The *D. mawsoni* trypsinogen types group well with temperate fishes within clades, indicative of high shared sequence identity (Figure 26). The examination of amino acid sequence identity within clades does not reveal any large changes unique to Antarctic *D. mawsoni* trypsinogens (Figure 25). Specifically within the catalytic triad and loop regions, Glycine 216 and 226 which confer trypsin activity and binding specificity (Hedstrom 1996). *D mawsoni* clade IIIB trypsinogens contain the S190A mutation and the K188S mutation that are speculated to confer dual trypsin and chymotrypsin enzyme activity to *G. morhua* Y. This would suggest all of the *D. mawsoni* trypsinogens clustering within clades I, IIIA, and IIIB are most likely capable of performing functions similar to those observed in temperate teleost clades I and IIIB trypsinogens (Asgeirsson, Fox, and Bjarnason 1989; Outzen et al. 1996; Gudmundsdottir and Palsdottir 2005).

Our genomic and transcriptomic screens suggest *D. mawsoni* may not contain functional clade II trypsinogens within their genome or transcriptome. It is unclear how Antarctic notothenioids adapted to the loss of this ancient highly conserved trypsinogen gene type. However, clade II trypsinogen transcript is detectable in the temperate water basal non-Antarctic notothenioid *Eleginops maclovinus* (data not shown) and all other teleosts screened, suggesting Antarctic clade II trypsinogen loss may be a result of the cold Antarctic environment.

Antarctic Notothenioid Trypsinogen Expression Profiles: Trypsinogen Response to Extreme Cold.

Antarctic notothenioids may have adapted to life in the cold Southern Ocean by increasing the amount of clade III trypsinogen transcript compared to clade I trypsinogen transcript abundance (temperate water digestive trypsin) (Figure 28). We speculate Antarctic clade III trypsinogens were recruited based on their observed increase in activity at cold temperatures in recombinantly expressed *G. morhua* Y (clade IIIB variant), compared to their temperate counterparts (clade I). Recombinantly expressed cod Y has been shown to be more cold-adapted than their clade I counterparts in *G. morhua* having an activity range of 2-30°C with optimal activity at 21°C, compared to clade I activity from 4-66°C with optimal activity at 56°C (Spilliaert and Gudmundsdottir 1999; Gudmundsdottir and Palsdottir 2005; Palsdottir and Gudmundsdottir 2007b). Molecular analyses have indicated *G. morhua* and all clade III trypsinogens may share the cold adaptive properties that confer increased cold activity to *G. morhua* Y (Spilliaert and Gudmundsdottir 1999; Roach 2002).

Although many teleost fishes have been found to contain cold adapted trypsinogens, Antarctic expression profiles represent the first time clade III trypsinogen transcript abundance was elevated above that of the main digestive trypsinogen (clade I) in any teleost. Previous analyses in warmer water teleosts reported the clade IIIB trypsinogen transcript abundance in the warmer water teleost *G. morhua* (Cod Y) (Palsdottir and Gudmundsdottir 2007b) and clade IIIA (Ssen3) and IIIB (SsenY) trypsinogen transcript abundance in *S. senegalensis* (Manchado et al. 2008) were much lower than that of clade I. Under cold stress, one would expect clade III levels trypsinogen transcript levels (and presumable protein product) to be equal to or higher than that of clade I trypsinogen (less cold adapted digestive trypsin), as seen in *D. mawsoni*. Based on our trypsinogen transcript abundance, it appears Antarctic clade III trypsinogens (both A and B) may act as the main digestive trypsinogen in Antarctic notothenioids.

Antarctic notothenioid non-digestive trypsinogen tissue distributions followed the same temperature dependant trend observed between temperate and Antarctic pancreatic trypsinogen transcript abundance (Figure 29). Our comparative study in warm-acclimated and native *P. borchgrevinki* also showed trypsinogen transcript between the two temperatures varied in a manner correlating higher clade III (A and B) trypsinogen transcript (and presumably higher protein product) to cold temperature (Figure 30). Aside from digestion within the small intestine, trypsins process protease activating receptors (PARs) in multiple non-digestive tissues which control such processes as inflammation and pain sensation (Dale and Vergnolle 2008; Larsen et al. 2008). Trypsins were also associated with processing functions within immune system cells and host defense (Koshikawa et al. 1998; Bajaj-Elliott 2003). Increased clade III transcript (above clade I) in non-digestive tissues has not to our knowledge been observed in any teleosts to date, and represents the first physiological example suggesting clade III trypsinogens (both A and B) utilization for their putative psychrophilic functions in both digestive and non-digestive tissues.

Antarctic clade I trypsinogens have also been shown to be cold adapted (Genicot, Feller, and Gerday 1988; Genicot et al. 1996) it is unclear why clade III trypsinogens were recruited as the main Antarctic notothenioid digestive trypsinogen. As we have not quantitated temperate teleost clade I transcript we do not claim that Antarctic clade I transcript is decreasing in Antarctic notothenioids compared to temperate clade I transcript abundance. Our comparisons are limited to Antarctic clade I trypsinogens with respect to Antarctic clade III transcript abundance and tissue distribution. We speculate from biochemical analyses in temperate teleost trypsinogen orthologs (*G. morhua*-clade I and IIIB and *S. salar*-clade I), that perhaps the observed higher innate activity of clade III trypsinogens at lower temperatures predisposed them to Antarctic cold selective pressures. Our Antarctic notothenioid trypsinogen expression profiles present the first example of trypsinogen expression profiles in response to extreme cold and may suggest that clade III trypsins were recruited in Antarctic notothenioids for their putative psychrophilic properties.

Clade III Trypsinogens May Have an Additional Function Unrelated to Their Psychrophilic Nature.

High expression of Antarctic clade III trypsinogens indicated these trypsinogens may serve as the main digestive and processing trypsin in teleosts (Figure 28) (Figure 29) (Figure 30) however, low expression levels in temperate fishes does not support clade III usage in this capacity (Palsdottir and Gudmundsdottir 2007a; Manchado et al. 2008).

The observed maintenance of functional trypsinogen presence within a given tissue in temperate and Antarctic teleosts highlighted the necessity of trypsin activity within digestive and non-digestive tissues (Figure 29) (Figure 30). However, the retention of clade III trypsinogen transcript in many tissues may indicate an additional processing function (outside of typical trypsin processing-capable of being performed by clade I trypsins) unable to be performed by clade I trypsins (Figure 29). The recombinant Cod Y protein, clade IIIB variant, has been shown to contain trypsin activity (redundant to clade I) (Palsdottir and Gudmundsdottir 2007b). This protein was also shown to have chymotrypsin capability (unique to clade IIIB) (Palsdottir and Gudmundsdottir 2007b). Dual enzyme functionality was attributed to S190A and K188S mutations that may broaden the enzyme specificity by altering positioning of the substrate within the specificity pocket and varying substrate binding preferences, both of which were observed in recombinant *G .morhua* Y (Palsdottir and Gudmundsdottir 2007b). As the structure of chymotrypsin and trypsin are known to be similar residue variability within the loop regions has been shown to indirectly affect trypsin or chymotrypsin activity by altering substrate positioning within the catalytic pocket (Hedstrom et al. 1994; Hedstrom 1996). Clade IIIB trypsinogens evolution around the time of Siluriformes diversification may be associated with its dual enzyme functionality observed in recombinant *G. morhua* Y (Palsdottir and Gudmundsdottir 2007b) and may be shared across all clade IIIB trypsinogens due to high sequence conservation within all clade IIIB trypsinogens and specific residue conservation mentioned above (Figure 25) (Figure 26). Dual enzyme functionality may be one selective pressure driving clade IIIB evolution and maintaining clade IIIB genomic/transcriptomic presence in teleosts. It is unknown if clade IIIA trypsins also have dual enzyme activity, however it is unlikely as the key residues attributed to this activity in the *G. morhua* Y (clade IIIB) are not conserved in clade IIIA trypsinogens. Further study into clade IIIB and newly characterized clade IIIA function in temperate fishes is warranted to examine the possibility of alternate functions for these trypsinogens unrelated to their psychrophilic tendencies. Antarctic (*D. mawsoni* and *P. borchgrevinki*) and temperate (*O. mykiss*) teleost trypsinogen expression profiles may indicate that low constitutive levels of clade III (both A and B) trypsinogen transcript observed in our study as well as in other temperate (Palsdottir and Gudmundsdottir 2007a) and tropical (Manchado et al. 2008) fishes may reflect a vital non-redundant processing capability unique to clade III trypsins.

CONCLUSIONS

Antarctic *D. mawsoni* trypsinogens provide a valuable set of stenothermal trypsinogens, currently missing from the trypsinogen repertoire. Classification of Antarctic trypsinogens within the current vertebrate trypsinogen phylogeny allowed increased resolution at the base of the vertebrate lineage by the usage of multiple teleost trypsinogens which do not appear to experience gene conversion (observed in mammals). Mammalian gene conversion has been predicted to obscure meaningful trypsinogen phylogenetic inference (Roach et al. 1997). Our phylogenetic analysis also indicated that tetrapod trypsinogens previously classified as group I and II trypsinogens by Roach (1997) evolved from a common teleost group II trypsinogen (termed clade II in our analysis) upon loss of the teleost group I ortholog (termed clade I in our analysis) in the tetrapod lineage. Phylogenetic inference and intron/exon structure supported the presence of two clade III trypsinogen gene types in teleosts (we termed clade IIIA and B). Characterization of Antarctic clade I, clade III trypsinogen types provided the first example of clade I and III trypsinogen expression in response to extreme cold. Antarctic trypsinogen transcript abundance and tissue distribution patterns suggested that clade III trypsins may be recruited for usage as the main digestive trypsin in Antarctic notothenioids. Clade IIIA trypsinogens not previously characterized may represent a new form of cold adapted trypsin in Antarctic notothenioids. Disparity between trypsinogen clade type transcript abundance in warm and polar fishes, coupled with the discovery of clade III trypsinogens origin in a time of warm climate may suggest an additional clade III trypsin function.

LITERATURE CITED

Amemiya, C. T., T. Ota, and G. W. Litman. 1996. Construction of P1 artificial chromosome (PAC) libraries from lower vertebrates. Pp. 223-256 *in* E. Lai, and B. Birren, eds. Analysis of Nonmammalian Genomes. Academic Press, San Diego, CA.

Amemiya, C. T., T. P. Zhong, G. A. Silverman, M. C. Fishman, and L. I. Zon. 1999. Zebrafish YAC, BAC, and PAC genomic libraries. Meth. in Cell Biol. **60**:235-258.

Anderson, J. B. 1999. Antarctic Marine geology. Cambridge University Press, Cambridge, MA.

Asgeirsson, B., and P. Cekan. 2006. Microscopic rate-constants for substrate binding and acylation in cold-adaptation of trypsin I from Atlantic cod. FEBS Lett. **580**:4639-4644.

Asgeirsson, B., J. W. Fox, and J. B. Bjarnason. 1989. Purification and characterization of trypsin from the poikilotherm *Gadus morhua.* Eur. J. Biochem. **180**:85-94.

Bacon, A. L., S. M. Farrington, and M. G. Dunlop. 2000. Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Hum Mol Genet **9**:2707-2713.

Bajaj-Elliott, M. 2003. Trypsin and host defense: a new role for an old enzyme. Gut **52**:166-167.

Barrett, A. J., and N. D. Rawlings. 1995. Families and clans of serine peptidases. Arch. Biochem. Biophys. **318**:247-250.

Bilyk, K. T., and A. L. DeVries. 2009. Freezing avoidance of the Antarctic icefishes (Channichthyidae) across thermal gradients in the Southern Ocean. Polar Biol. **33**:203-213.

Briggs, J. C. 2005. The biogeography of otophysan fishes (Ostariophysi: Otophysi): a new appraisal. J. Biogeogr. **32**:287-294.

Chen, L., A. L. DeVries, and C. H. Cheng. 1997a. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc. Natl. Acad. Sci. U.S.A. **94**:3817-3822.

Chen, L., A. L. DeVries, and C. H. Cheng. 1997b. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. U.S.A. **94**:3811-3816.

Chen, Z., C.-H. Cheng, J. Zhang, L. Cao, L. Chen, L. Zhou, J. Yudong, Y. Hua, C. Deng, Z. Dai, Q. Xu, S. Sun, Y. Shen, and L. Chen. 2008. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. U.S.A. **105**:12944-12949.

Cheng, C.-H., and A. L. DeVries. 1991. The role of antifreeze glycopeptides in the freezing avoidance of cold water fishes. *in* G. di Prisco, ed. Life under extreme conditions. Springer-Verlag, Heidelberg, Berlin.

Cheng, C. H., and L. Chen. 1999. Evolution of an antifreeze glycoprotein. Nature **401**:443-444.

Cheng, C. H., L. Chen, T. J. Near, and Y. Jin. 2003. Functional antifreeze glycoprotein genes in temperate-water New Zealand nototheniid fish infer an Antarctic evolutionary origin. Mol. Biol. Evol. **20**:1897-1908.

Cheng, C. H., P. A. Cziko, and C. W. Evans. 2006. Nonhepatic origin of notothenioid antifreeze reveals pancreatic synthesis as common mechanism in polar fish freezing avoidance. Proc. Natl. Acad. Sci. U.S.A. **103**:10491-10496.

Cheng, C. H., and H. W. Detrich, 3rd. 2007. Molecular ecophysiology of Antarctic notothenioid fishes. Philos. Trans. R. Soc. Lond., B, Biol. Sci. **362**:2215-2232.

Dale, C., and N. Vergnolle. 2008. Protease signaling to G protein-coupled receptors: implications for inflammation and pain. Journal of Receptor and Signal Transduction Research **28**:29-37.

de Haen, C., H. Neurath, and D. C. Teller. 1975. The phylogeny of trypsin related serine proteases and their zymogen: new methods for the investigation of distant evolutionary relationship. J. Mol. Biol. **92**:225-259.

DeVries, A. L. 1971. Glycoproteins as biological antifreeze agents in Antarctic fishes. Science **172**:1152-1155.

DeVries, A. L., S. K. Komatsu, and R. E. Feeney. 1970. Chemical and physical properties of freezing point-depressing glycoproteins from Antarctic fishes. J. Biol. Chem. **245**:2901-2908.

DeVries, A. L., J. Vandenheede, and R. E. Feeney. 1971. Primary structure of freezing point-depressing glycoproteins. J. Biol. Chem. **246**:305-308.

Dolezel, J., J. Bartos, H. Voglmayr, and J. Greilhuber. 2003. Nuclear DNA content and genome size of trout and human.

Cytometry. Part A : The Journal of the International Society for Analytical Cytology **51**:127-128; author reply 129.

Eastman, J. M., AR. 2000. Fishes of the Antarctic continental shelf: evolution of a marine species flock? J. Fish Biol. **57**:84-102.

Eastman, J. T. 2005. The nature of diversity of Antarctic fishes. Polar Biol. **28**:93-107.

Eastman, J. T. 1993. Antarctic Fish Biology; Evolution in a Unique Environment. Academic, San Diego.

Eckert, K. A., and S. E. Hile. 2009. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol Carcinog **48**:379-388.

Eichler, E. E. 1998. Masquerading repeats: paralogous pitfalls of the human genome. Genome Res. **8**:758-762.

Eichler, E. E., M. E. Johnson, C. Alkan, E. Tuzun, C. Sahinalp, D. Misceo, N. Archidiacono, and M. Rocchi. 2001. Divergent origins and concerted expansion of two segmental duplications on chromosome 16. The Journal of Heredity **92**:462-468.

Foster, B. 1984. The marine environment. Pp. 345-371 *in* R. M. Laws, ed. Antarctic Ecology. Academic Press, London.

Genicot, S., G. Feller, and C. H. Gerday. 1988. Trypsin from Antarctic fish (*Paranotothenia magellanica Forster*) as compared

with trout (*Salmo gairdneri*) trypsin. Comp. Biochem. Physiol. B: Biochem. Mol. Biol. **90**:601-609.

Genicot, S., F. Rentier-Delrue, D. Edwards, J. VanBeeumen, and C. Gerday. 1996. Trypsin and trypsinogen from an Antarctic fish: molecular basis of cold adaptation. Biochim Biophys Acta **1298**:45-57.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc. Natl. Acad. Sci. U.S.A. **103**:11240-11245.

Gudmundsdottir, A., and H. M. Palsdottir. 2005. Atlantic cod trypsins: from basic research to practical applications. Mar. Biotechnol. **7**:77-88.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696-704.

Harr, B., and C. Schlotterer. 2000. Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155**:1213-1220.

Hedstrom, L. 1996. Trypsin: a case study in the structural determinant of enzyme specificity. Biol. Chem. **377**:465-470.

Hedstrom, L., S. Farr-Jones, C. A. Kettner, and W. J. Rutter. 1994. Converting trypsin to chymotrypsin: ground-state binding does not determine substrate specificity. Biochemistry **33**:8764-8769.

Hsiao, K. C., C. H. Cheng, I. E. Fernandes, H. W. Detrich, and A. L. DeVries. 1990. An antifreeze glycopeptide gene from the Antarctic cod *Notothenia coriiceps neglecta* encodes a polyprotein of high peptide copy number. Proc. Natl. Acad. Sci. U.S.A. **87**:9265-9269.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Jacobs, G., D. Dechyeva, T. Wenke, B. Weber, and T. Schmidt. 2009. A BAC library of *Beta vulgaris L.* for the targeted isolation of centromeric DNA and molecular cytogenetics of Beta species. Genetica **135**:157-167.

Jeohn, G. H., S. Serizawa, A. Iwamatsu, and K. Takahashi. 1995. Isolation and characterization of gastric trypsin from the microsomal fraction of porcine gastric antral mucosa. J. Biol. Chem. **270**:14748-14755.

Jin, Y., and A. L. DeVries. 2006. Antifreeze glycoprotein levels in Antarctic notothenioid fishes inhabiting different thermal environments and the effect of warm acclimation. Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **144**:290-300.

Kalbe, M., C. Eizaguirre, I. Dankert, T. B. Reusch, R. D. Sommerfeld, K. M. Wegner, and M. Milinski. 2009. Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. Proc Biol Sci **276**:925-934.

Kennett, J. 1982. Marine Geology. Prentice-Hall, Englewood, NJ.

Kiss, A. J., A. Y. Mirarefi, S. Ramakrishnan, C. F. Zukoski, A. L. Devries, and C. H. Cheng. 2004. Cold-stable eye lens crystallins of the Antarctic nototheniid toothfish *Dissostichus mawsoni Norman.* J. Exp. Biol. **207**:4633-4649.

Knecht, W., G. S. Cottrell, S. Amadesi, J. Mohlin, A. Skaregarde, K. Gedda, A. Peterson, K. Chapman, M. D. Hollenberg, N. Vergnolle, and N. W. Bunnett. 2007. Trypsin

IV or mesotrypsin and p23 cleave protease-activated receptors 1 and 2 to induce inflammation and hyperalgesia. J. Biol. Chem. **282**:26089-26100.

Koshikawa, N., S. Hasegawa, Y. Nagashima, K. Mitsuhashi, Y. Tsubota, S. Miyata, Y. Miyagi, H. Yasumitsu, and K. Miyazaki. 1998. Expression of trypsin by epithelial cells of various tissues, leukocytes, and neurons in human and mouse. American Journal of Pathology **153**:937-944.

Lai, Y., and F. Sun. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol **20**:2123-2131.

Larsen, A. K., O. M. Seternes, M. Larsen, L. Aasmoe, and B. Bang. 2008. Salmon trypsin stimulates the expression of interleukin-8 via protease-activated receptor-2. Toxicol. Appl. Pharmacol.

Leiros, H. K., N. P. Willassen, and A. O. Smalas. 2000. Structural comparison of psychrophilic and mesophilic trypsins. Elucidating the molecular basis of cold-adaptation. Eur. J. Biochem. **267**:1039-1049.

Leiros, H. K., N. P. Willassen, and A. O. Smalas. 1999. Residue determinants and sequence analysis of cold-adapted trypsins. Extremophiles **3**:205-219.

Lemaitre, C., and M. Sagot. 2008. A small trip in the untraquil world of genomes A survey on the detection and analysis of genome rearrangement breakpoints. Theoretical Computer Science **395**:171-192.

Lilleeng, E., M. K. Froystad, G. C. Ostby, E. C. Valen, and A. Krogdahl. 2007. Effects of diets containing soybean meal on trypsin mRNA expression and activity in Atlantic salmon (*Salmo salar L*). Comp. Biochem. Physiol., Part A Mol. Integr. Physiol. **147**:25-36.

Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods **25**:402-408.

Loyd, C. R. 1984. Pre-Pleistocene paleoclimates: The geological and paleontological evidence; modeling strategies,boundary conditoins and some preliminary results. Academic Press, Inc, New York.

Manchado, M., C. Infante, E. Asensio, A. Crespo, E. Zuasti, and J. P. Canavate. 2008. Molecular characterization and gene expression of six trypsinogens in the flatfish Senegalese sole (*Solea senegalensis Kaup*) during larval development and in tissues. Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **149**:334-344.

Mazzei, F., L. Ghigliotti, J. Coutanceau, H. Detrich, V. Prirodina, C. Ozouf-Costaz, and E. Pisano. 2007. Chromosomal characteristics of the temperate notothenioid fish *Eleginops maclovinus* (Cuvier). Polar Biol. **31**:629-634.

Metzker, M. L. 2005. Emerging technologies in DNA sequencing. Genome Res. **15**:1767-1776.

Murray, H. M., J. C. Perez-Casanova, J. W. Gallant, S. C. Johnson, and S. E. Douglas. 2004. Trypsinogen expression during the development of the exocrine pancreas in winter flounder (*Pleuronectes americanus*). Comp. Biochem. Physiol., Part A Mol. Integr. Physiol. **138**:53-59.

O'Driscoll, M., and P. A. Jeggo. 2006. The role of double-strand break repair- insights from human genetics. Nat. Rev. Genet. **7**:45-54.

O'Grady, S. M., J. C. Ellory, and A. L. DeVries. 1982. Protein and glycoprotein antifreezes in the intestinal fluid of polar fishes. J. Exp. Biol. **98**:429-438.

Osoegawa, K., P. J. de Jong, E. Frengen, and P. A. loannou. 2001. Construction of bacterial artificial chromosome (BAC/PAC) libraries. Wiley-Intersciences.

Outzen, H., G. I. Berglund, A. O. Smalas, and N. P. Willassen. 1996. Temperature and pH sensitivity of trypsins from Atlantic salmon (*Salmo salar*) in comparison with bovine and porcine trypsin. Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **115**:33-45.

Palsdottir, H. M., and A. Gudmundsdottir. 2007a. Development of a qRT-PCR assay to determine the relative mRNA expression of two different trypsins in Atlantic cod (*Gadus morhua*). Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **146**:26-34.

Palsdottir, H. M., and A. Gudmundsdottir. 2007b. Expression and purification of a cold-adapted group III trypsin in *Escherichia coli*. Protein Expression Purif. **51**:243-252.

Petes, T. D., P. W. Greenwell, and M. Dominska. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. Genetics **146**:491-498.

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

Posada, D. a. C., KA. 1998. Modeltest: testing the model of DNA substitiution. Bioinformatics **14**:817-818.

Powers, J. C., S. Odake, J. Oleksyszyn, H. Hori, T. Ueda, B. Boduszek, and C. Kam. 1993. Proteases--structures, mechanism and inhibitors. Agents Actions Suppl. **42**:3-18.

Reusch, T. B., A. Ehlers, A. Hammerli, and B. Worm. 2005. Ecosystem recovery after climatic extremes enhanced by genotypic diversity. Proc Natl Acad Sci U S A **102**:2826-2831.

Richards, R. I., and G. R. Sutherland. 1997. Dynamic mutation: possible mechanisms and significance in human disease. Trends Biochem. Sci. **22**:432-436.

Roach, J. C. 2002. A clade of trypsins found in cold-adapted fish. Proteins **47**:31-44.

Roach, J. C., K. Wang, L. Gan, and L. Hood. 1997. The molecular evolution of the vertebrate trypsinogens. J. Mol. Evol. **45**:640-652.

Ronaghi, M. 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. **11**:3-11.

Rowen, L., E. Williams, G. Glusman, E. Linardopoulou, C. Friedman, M. E. Ahearn, J. Seto, C. Boysen, S. Qin, K. Wang, A. Kaur, S. Bloom, L. Hood, and B. J. Trask. 2005. Interchromosomal segmental duplications explain the unusual structure of PRSS3, the gene for an inhibitor-resistant trypsinogen. Mol. Biol. Evol. **22**:1712-1720.

Scheele, G., D. Bartelt, and W. Bieger. 1981. Characterization of human exocrine pancreatic proteins by two-dimensional isoelectric focusing/sodium dodecyl sulfate gel electrophoresis. Gastroenterology **80**:461-473.

Shaw, C. J., and J. R. Lupski. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum. Mol. Genet. **13 Spec No 1**:R57-64.

Smalas, A. O., E. S. Heimstad, A. Hordvik, N. P. Willassen, and R. Male. 1994. Cold adaption of enzymes: structural comparison between salmon and bovine trypsins. Proteins **20**:149-166.

Smith, M. J. 1998. Evolutionary genetics, Oxford.

Soderlund, C., S. Humphray, A. Dunham, and L. French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. **10**:1772-1787.

Soderlund, C., I. Longden, and R. Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. **13**:523-535.

Somero, G. N. 2003. Protein adaptations to temperature and pressure: complementary roles of adaptive changes in amino acid sequence and internal milieu. Comp. Biochem. Physiol. B, Biochem. Mol. Biol. **136**:577-591.

Spilliaert, R., and A. Gudmundsdottir. 1999. Atlantic Cod Trypsin Y-Member of a Novel Trypsin Group. Mar Biotechnol (NY) **1**:598-607.

Swofford, D. L. 2001. PAUP*: Phylogenetic Analysis Using Parsimony(**and Other Methods)*. Sinauer Associates, Suderland, MA.

Szamalek, J. M., V. Goidts, N. Chuzhanova, H. Hameister, D. N. Cooper, and H. Kehrer-Sawatzki. 2005. Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome. Hum. Genet. **117**:168-176.

Szinay, D., S. B. Chang, L. Khrustaleva, S. Peters, E. Schijlen, Y. Bai, W. J. Stiekema, R. C. van Ham, H. de Jong, and R. M. Klein Lankhorst. 2008. High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. Plant J. **56**:627-637.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**:1596-1599.

Tautz, D., M. Trick, and G. A. Dover. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**:652-656.

Tellam, R. L., D. G. Lemay, C. P. Van Tassell, H. A. Lewin, K. C. Worley, and C. G. Elsik. 2009. Unlocking the bovine genome. BMC Genomics **10**:193.

Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. **20**:2587-2595.

Whelan, J. A., N. B. Russell, and M. A. Whelan. 2003. A method for the absolute quantification of cDNA using real-time PCR. J Immunol Methods **278**:261-269.

Wichman, H. A., R. A. Van Den Bussche, M. J. Hamilton, and R. J. Baker. 1992. Transposable elements and the evolution of genome organization in mammals. Genetica **86**:287-293.

Wicker, T., E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics **7**:275.

Wierdl, M., M. Dominska, and T. D. Petes. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146**:769-779.

Xu, X., M. Peng, and Z. Fang. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet **24**:396-399.

Yasui, Y., M. Mori, D. Matsumoto, O. Ohnishi, C. G. Campbell, and T. Ota. 2008. Construction of a BAC library for buckwheat genome research - an application to positional cloning of agriculturally valuable traits. Genes Genet. Syst. **83**:393-401.

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

Zwilling, R., H. Neurath, L. H. Ericsoon, and D. L. Enfield. 1975. The amino-terminal sequence of an invertebrate trypsin (crayfish *Astacus leptodactylus*): homology with other serine proteases. FEBS Lett. **60**:247-249.

FIGURES

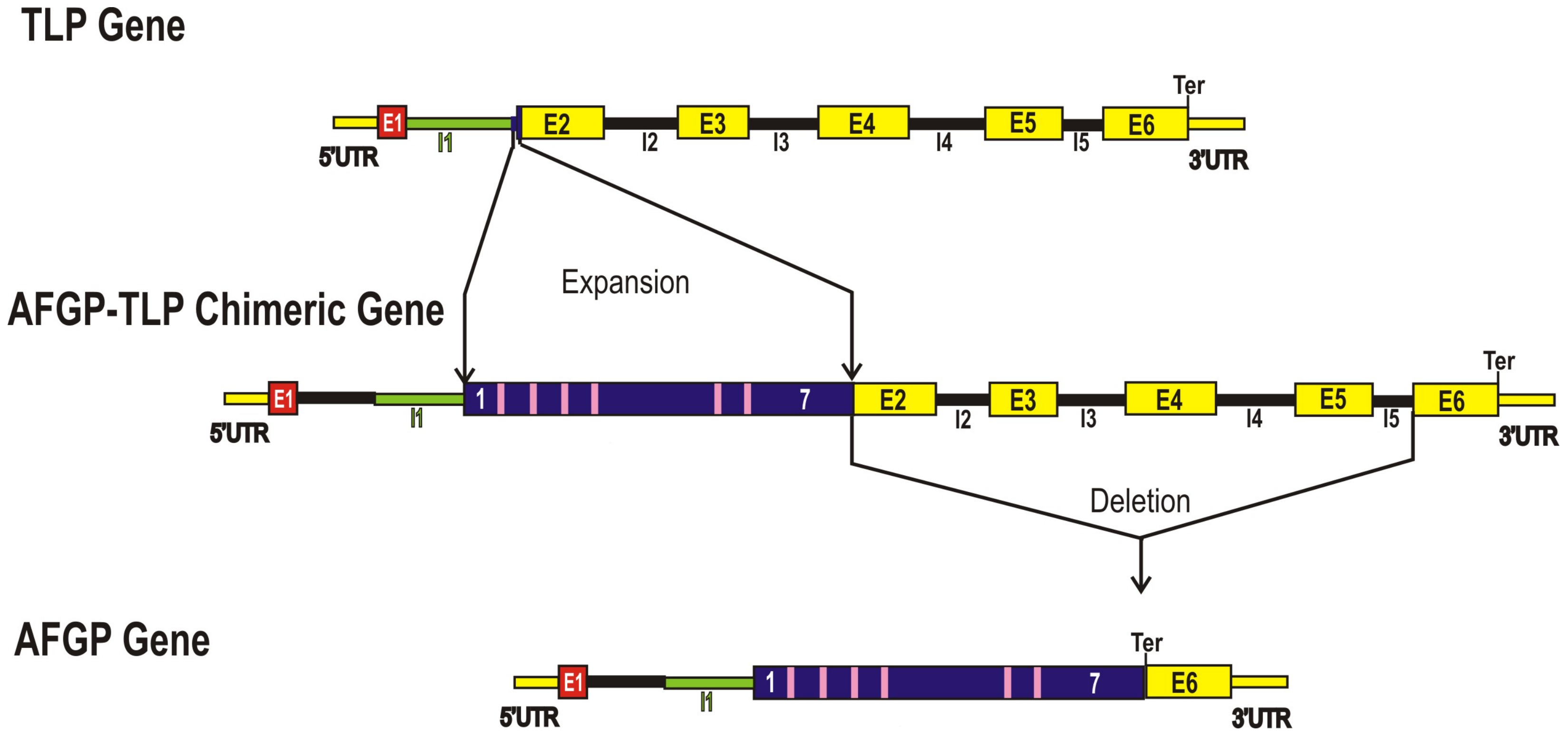


Figure 1. Antifreeze glycoprotein gene evolution. The Antifreeze Glycoprotein gene (bottom) evolved from a trypsinogen- like protein (top) precursor through an AFGP/TLP chimeric (middle) ancestral intermediate through expansion of a single TLP intron 1/exon 2 acagcggca (Thr-Ala-Ala) motif (purple) forming the AFGP-TLP chimeric gene. Recruitment of the 5' and 3' ends of the chimeric gene and deletion of the majority of the chimeric TLP containing region formed the AFGP gene. Exons (E) are represented as boxes introns (I) as lines. All gene introns and exons are color coded to indicate shared sequence identity within introns and exons. AFGP Ala-Ala-Thr coding sequence is represented as a purple box in TLP, AFGP/TLP chimeric and AFGP genes.

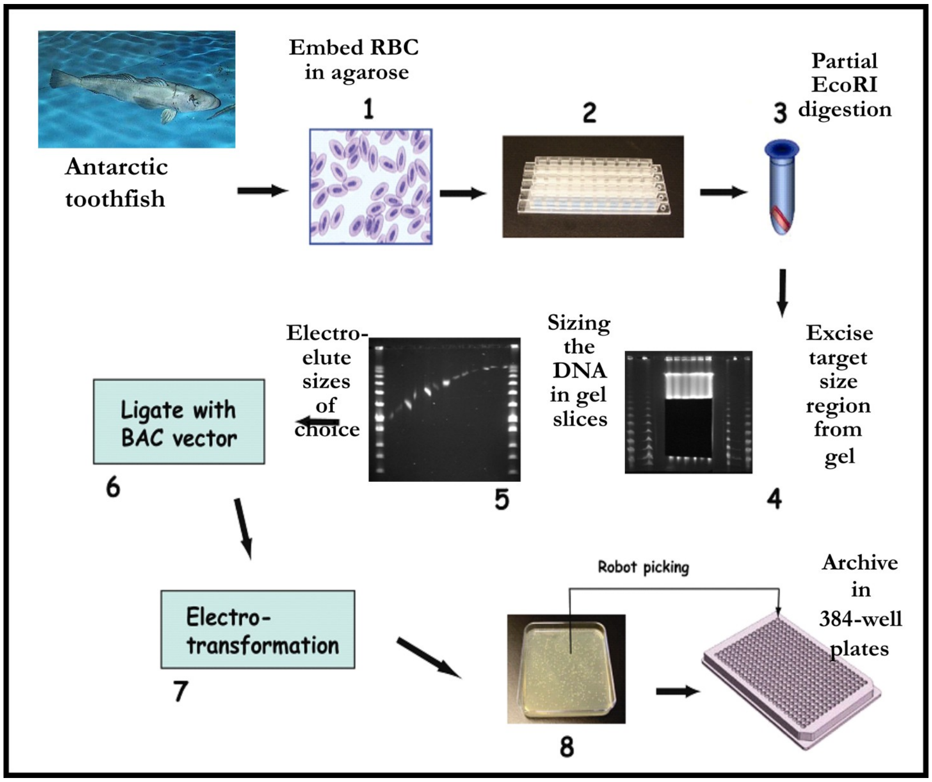


Figure 2. Bacterial Artificial Chromosome (BAC) Library Construction Overview. Schematic depicting the major steps necessary for constructing and archiving a BAC library. Modified after Miyake and Amemiya (2004)

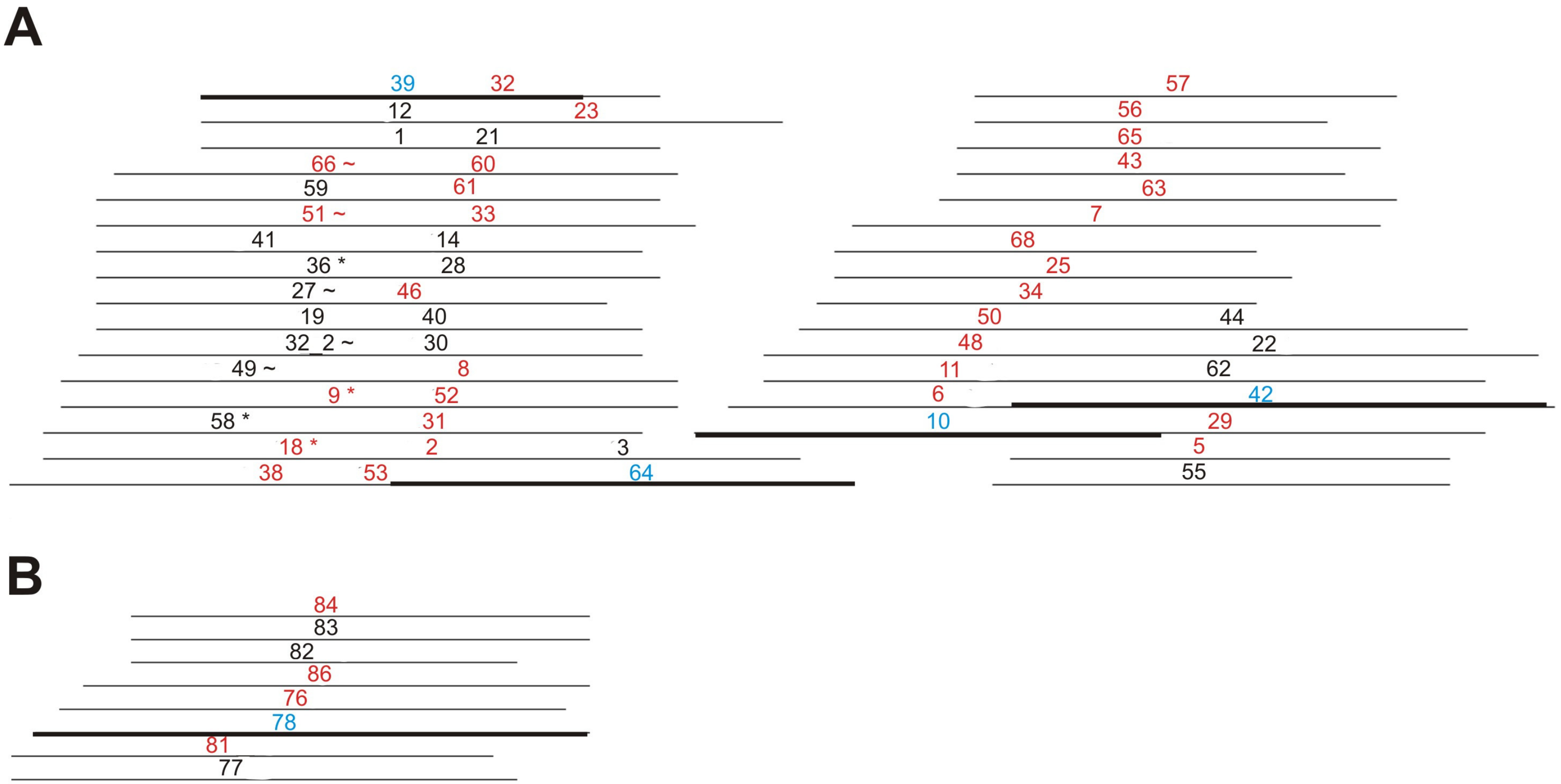


Figure 3. ***D. mawsoni*** **Bacterial Artificial Chromosome (BAC) library predicted alignment of AFGP/TLP positive clones using the algorithm of finger printing contig.** A. Representation of AFGP/TLP locus contig 1 containing only BAC clones positive for both AFGP and TLP. B. AFGP/TLP locus contig 2 containing AFGP positive BAC clones. The minimal tiling path (MTP), or the smallest number of overlapping clones that span the entire AFGP/TLP locus, are shown as bolded black lines, clone numbers are blue. Sequence data (at 8-11 X coverage) was obtained for clones within the MTP. Non-MTP BAC clones whose positions are verified by paired end matching are indicated in red.

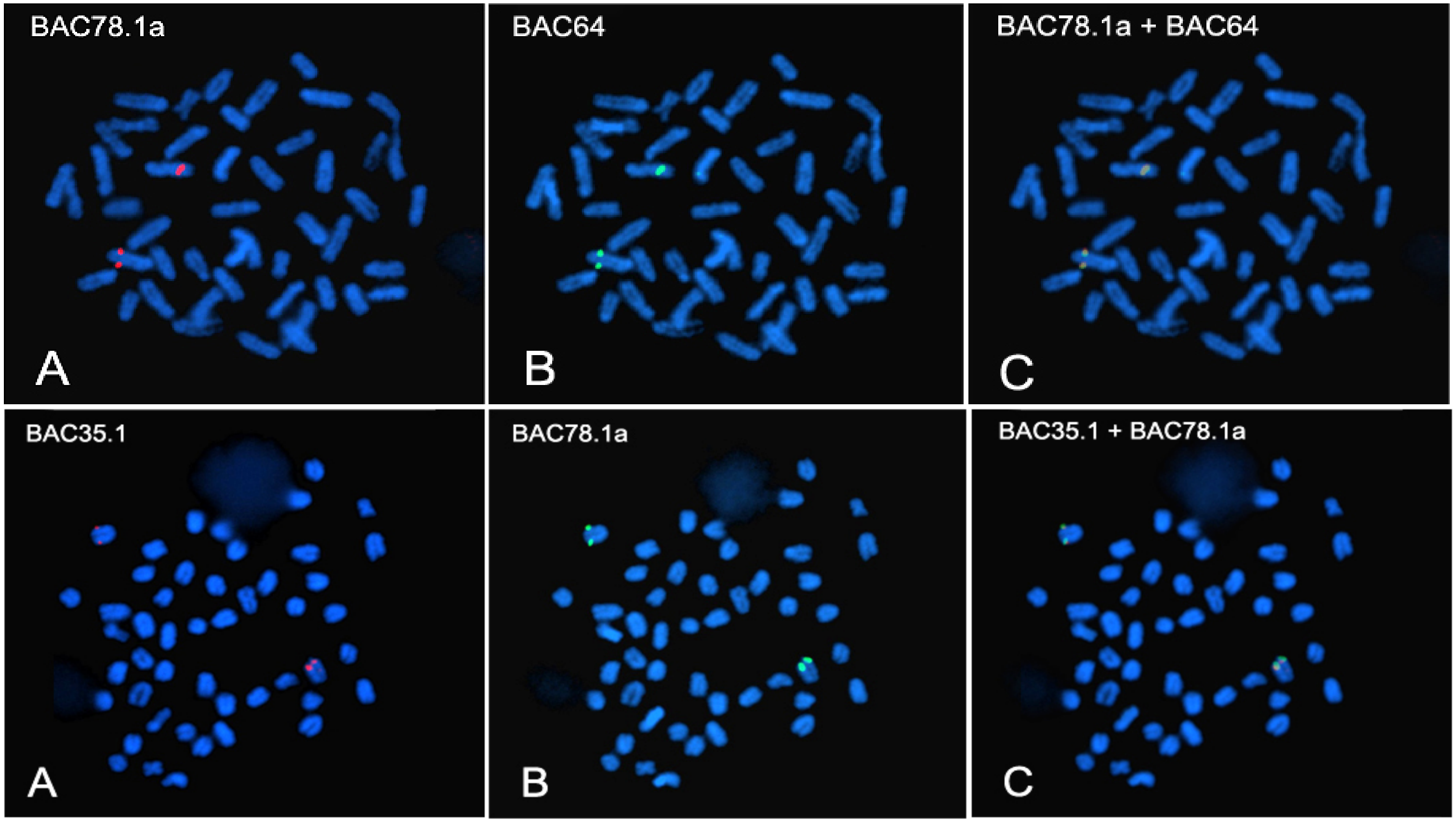


Figure 4. FISH (fluorescent in situ hybridization) of *Dissostichus mawsoni* minimal tiling path BAC clones from the three separate AFGP/TLP loci created using finger printing contig. The top three panels show FISH with BAC 78.1 alone as probe (A), stripped and re-hybridized with DmBAC 64 (B), and (C) superposition of the two images A and B. The bottom three panels likewise show FISH with DmBAC 35.1 alone (A), stripped and re-hybridized with DmBAC 78.1 (B), and (C) superposition of the two images A and B. Within the distance resolution of chromosomal FISH, all TLP and AFGP genes are contained with a single chromosomal region in the toothfish genome. Thus the non-overlapping BAC clones 35.1 and 78.1 are close neighbors of the major TLP/AFGP gene cluster. Chromosomal fish carried out by C.-H. Cheng and collaborators.

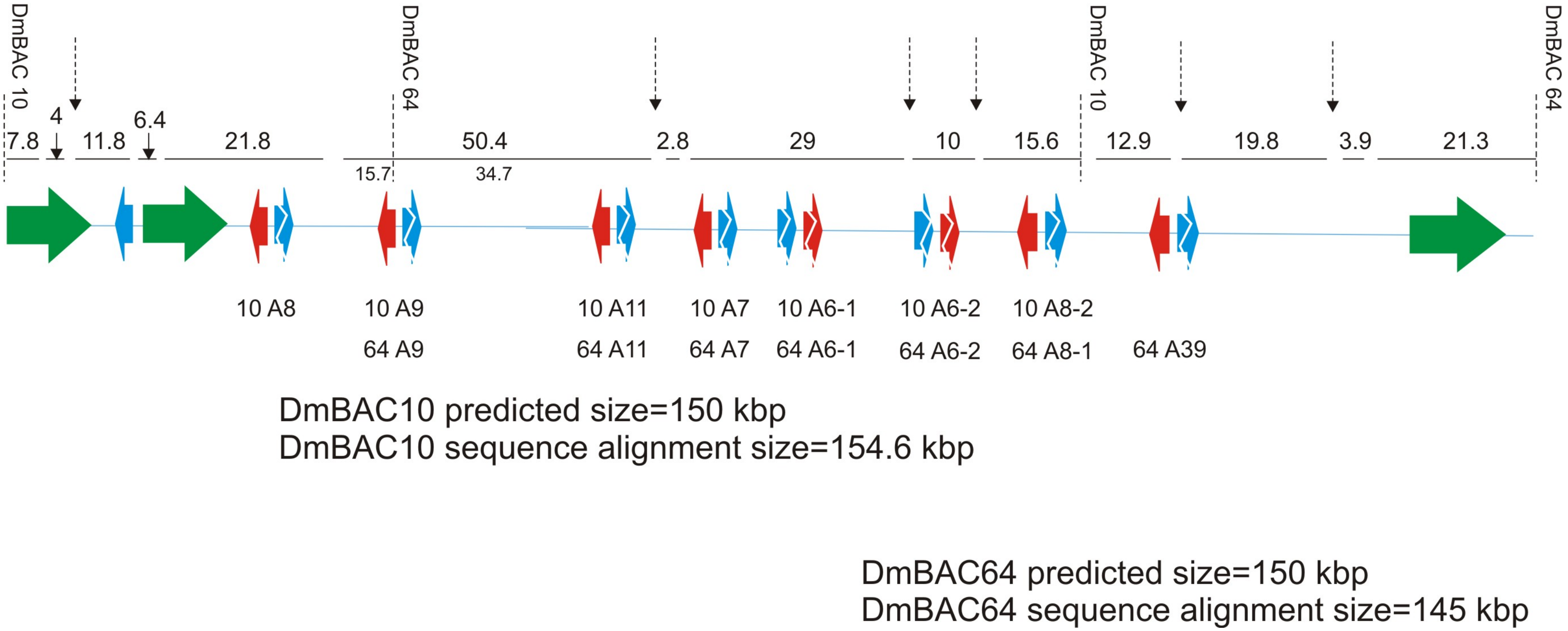
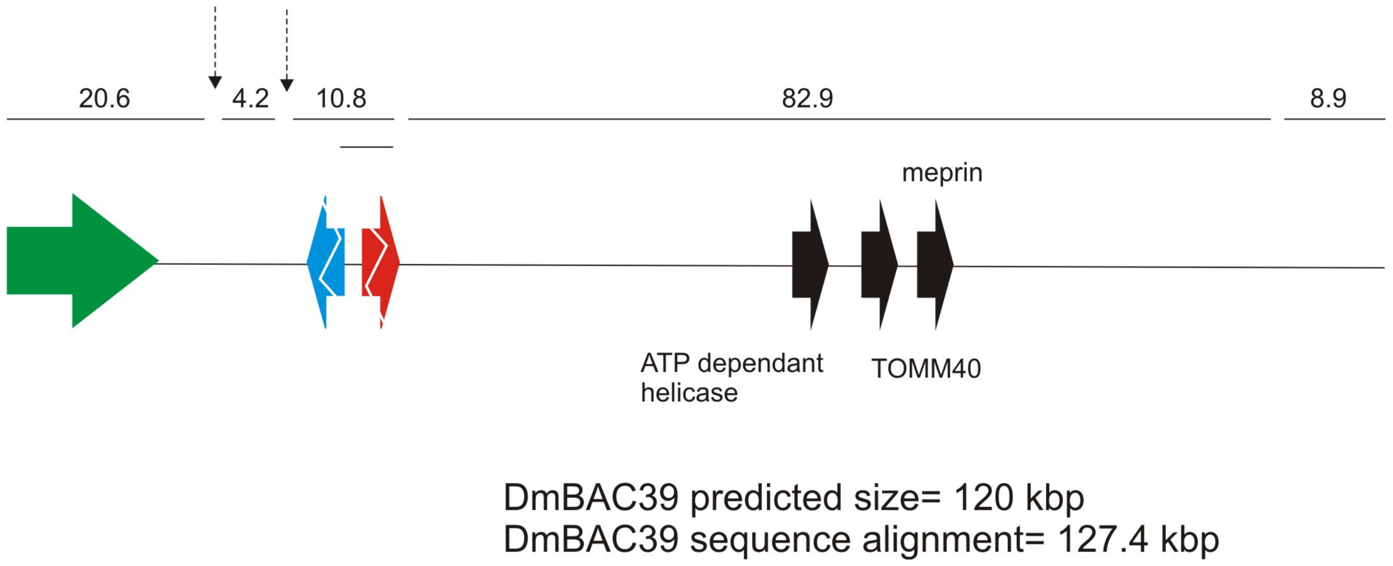
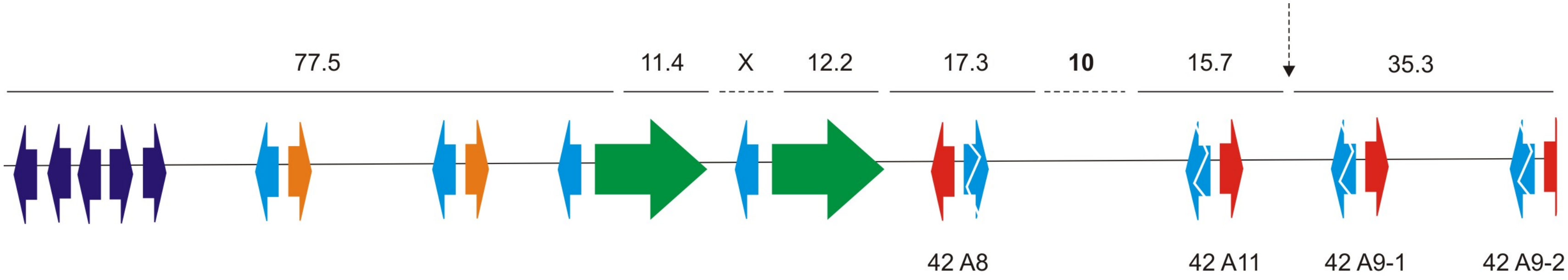


Figure 5. *D. mawsoni* BAC clone 64 and 10 (DmBAC10/64) shotgun subclone library sequence reconstruction. Arrows show gene orientation from 5' to 3' in the direction of the arrow. Broken arrows indicate a pseudogene. Color variations in arrows represent different genes, and correspond as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange- TLP, green-chimeric, red AFGP. Contig locations and sizes (kbp) are indicated above the assembly. AFGP (red- A) genes are labeled according to the last exon 2 polyprotein encoded in each gene and the BAC clone in which they are found (ex: 10 A8, AFGP gene A8 is found in DmBAC 10). AFGP gene names indicated below each AFGP gene. Spaces between contigs are gaps (<2 kbp) in the alignment. Vertical arrows indicate gaps of unknown size. Vertical dashed lines show the end of respective BAC clones. DmBAC clone predicted size, according to *Not*I enzymatic digestion and PFG electrophoretic analysis, is indicated below the alignment as is the size of the shotgun subclone sequence assembly.



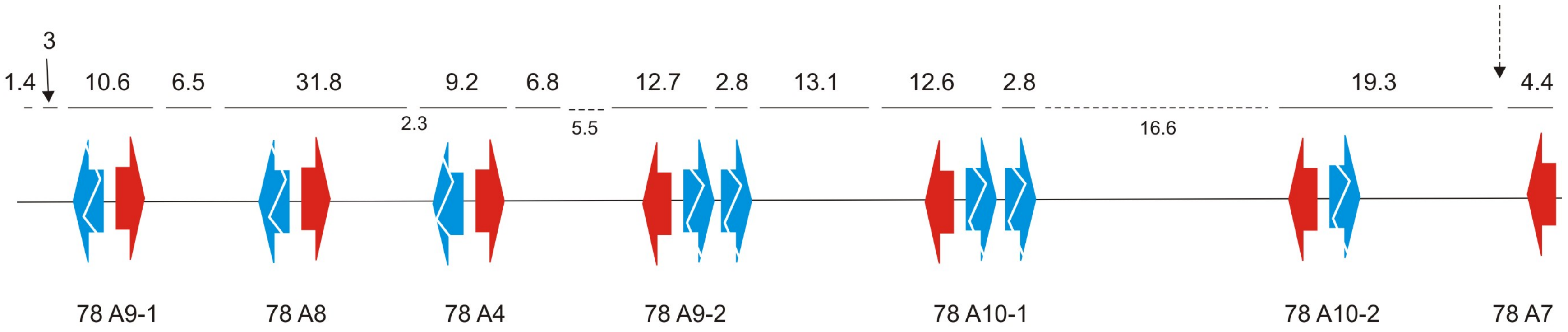
DmBAC39 predicted size= 120 kbp
DmBAC39 sequence alignment= 127.4 kbp

Figure 6. ***D. mawsoni*** **BAC clone 39 (DmBAC39) shotgun subclone library sequence reconstruction.** Genes are represented by arrows in a 5' to 3' directionality. Pseudogenes are shown as broken arrows. Color variations in arrows represent different genes, and correspond as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange- TLP, green-chimeric, red AFGP. Contigs locations and sizes (kbp) are indicated above the assembly. Spaces between contigs show the position of gaps in the alignment. Black arrow/gene identity is indicated above or below the gene. Vertical arrows indicate gaps of unknown size. DmBAC 39 predicted size according to *Not*I enzymatic digestion and PFG electrophoretic analysis is indicated below the alignment as is the size of the sequence assembly alignment. TOMM40 translocase of outer mitochondrial membrane 40.



DmBAC 42 predicted size=155 kbp
DmBAC 42 sequence assembly size=179 kbp

Figure 7. ***D. mawsoni*** **BAC clone 42 (DmBAC 42) shotgun subclone library assembly.** Genes are represented as arrows pointing toward the 3' end of the gene. Gene type corresponds as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange- TLP, green-chimeric, red AFGP. The presence of a pseudogene is indicated by a fragmented gene. Contig locations and sizes (kbp) are indicated above the assembly. Spaces between contigs or dotted lines represent gaps in the alignment. Gap sizes are smaller than 2 kbp unless otherwise indicated. The gap of size X represents a region within DmBAC 42 of unknown size. Vertical arrows indicate gaps of unknown size. BAC clone predicted size, determined by *Not*I enzymatic digestion and PFG electrophoretic analysis, is indicated below the alignment as is the size of the shotgun subclone sequence assembly alignment. AFGP (red- A) genes are labeled according to the last exon 2 polyprotein encoded in each gene (ex A8) and indicated below each AFGP gene.



DmBAC78 predicted size= 150 kbp
DmBAC78 sequence assembly size=161.4 kbp

Figure 8. *D. mawsoni* BAC library clone DmBAC78 shotgun subclone library assembly. Genes are represented by arrows in a 5' to 3' directionality. Pseudogenes are represented as fragmented arrows. Gene type corresponds as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange-TLP, green-chimeric, red AFGP. AFGP (red- A) genes are labeled according to the last exon 2 polyprotein encoded in each gene (ex A8) and indicated below each AFGP gene. Shotgun subclone assembly contig locations and sizes (kbp) are indicated above the assembly. Spaces between contigs and dotted lines represent gaps in the alignment. Gaps are smaller than 2 kbp unless indicated. Vertical arrows indicate gaps of unknown size. The predicted size of DmBAC 64 according to *Not*I enzymatic digestion and PFG electrophoretic analysis is indicated below the alignment as is the size of the sequence assembly alignment.

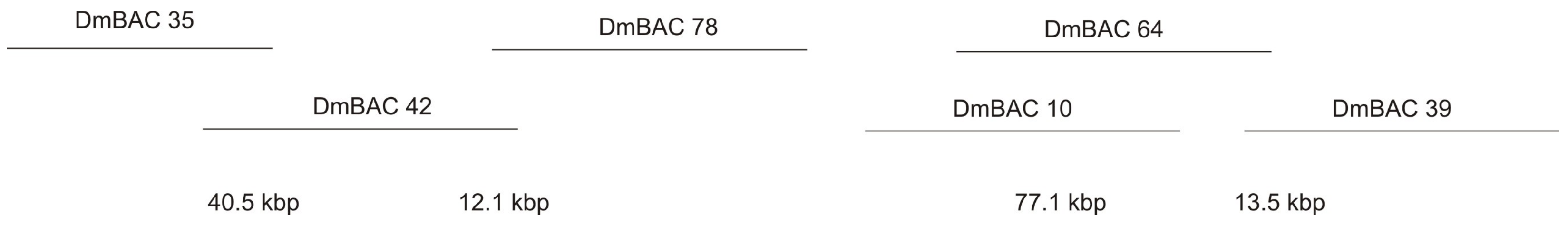


Figure 9. AFGP/TLP locus minimal tiling path. DmBAC shotgun sequence clone alignments, aligned together according to shared sequence identity indicative of overlapping regions. Amount of overlap is indicated below the alignment. DmBAC78 and 10 do not have any detectable overlapping regions producing a gap of unknown size within the AFGP/TLP locus.

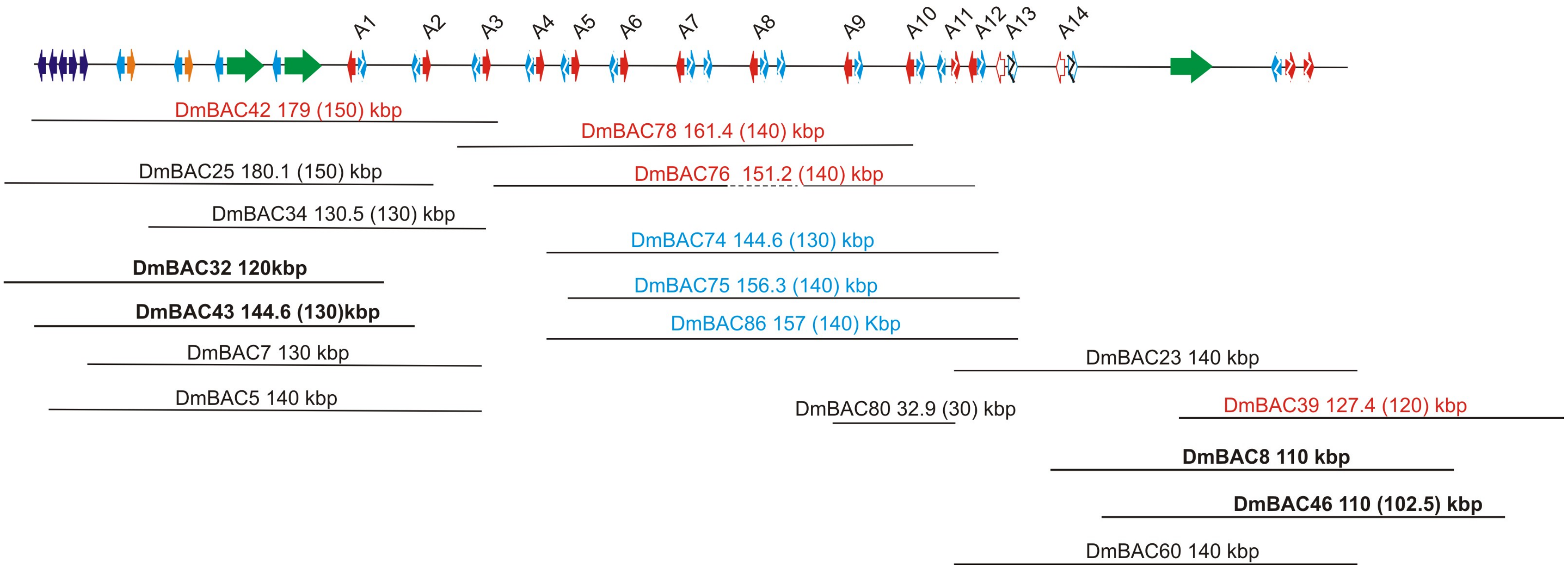


Figure 10. AFGP/TLP locus haplotype 1 pictoral representation. Genes are shown as arrows pointing toward the 3' end of the gene. Pseudogenes are shown as fragmented arrows. Color variations in arrows represent different genes, and correspond as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange- TLP, green-chimeric, red AFGP. AFGP gene names are indicated above each gene and correspond to their position within haplotype 1 from left to right (ex. A1 is the first AFGP gene encountered when proceeding from left to right). Outlined genes are genes predicted to be present by PCR analysis of DmBAC clones spanning that region. AFGP/TLP positive BAC clones tiling path determined by paired end matching of BAC clone end sequences within the shotgun subclone sequence reconstruction (DmBAC 42, 78, and 39) are shown below the alignment. Minimal tiling path clones (42, 78, and 39) and all clones with shotgun subclone sequence data are shown in red. Clones which may be present in either haplotype are shown in bold. DmBAC76 dashed line depicts the gene not detected in overlapping regions between 76 and 78. DmBAC clone predicted size according to *Not*I enzymatic digestion and PFG electrophoretic analysis is indicated in parentheses. BAC clone size determined by shotgun subclone sequence reconstruction (42, 78, 76, 39) or position of BAC clone end sequences within corresponding subclone sequence reconstruction is indicated next to each BAC clone name. Clones with large variations between *Not*I and shotgun subclone predicted size are shown in blue and also include DmBAC 78 and 76.

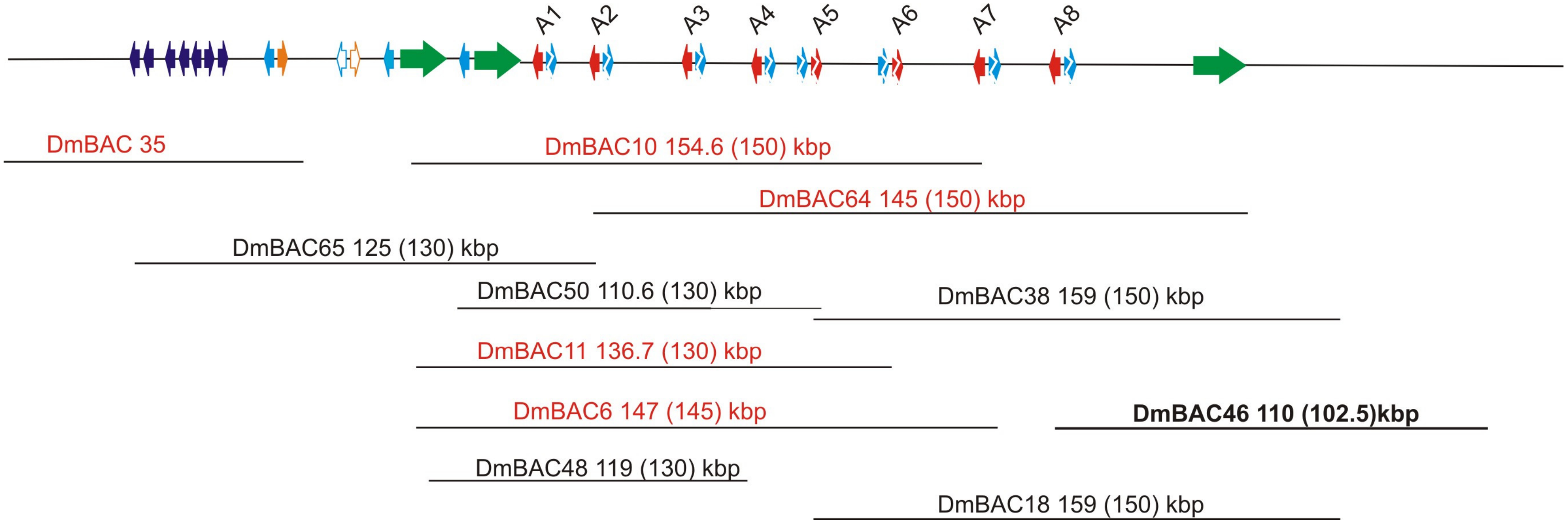


Figure 11. AFGP/TLP locus haplotype 2. Genes are represented by arrows in a 5' to 3' directionality. Pseudogenes are represented as fragmented arrows. Outlined genes are predicted to be present by shared sequence identity with haplotype 1. Color variations in arrows represent different genes, and correspond as follows: purple group I trypsinogens, blue arrows- group III trypsinogens, orange- TLP, green-chimeric, red AFGP. AFGP gene names are indicated above each gene and correspond to their position within haplotype 2 from left to right (ex. A1 is the first AFGP gene encountered when proceeding from left to right). AFGP/TLP positive BAC clones tiling path determined by paired end matching of BAC clone end sequences within the shotgun subclone sequence reconstruction (DmBAC 35, 10, 64, and 39) are shown below the alignment. Minimal tiling path clones (35, 10, 64, 39) and all clones with shotgun subclone sequence data are shown in red. Clones unable to be assigned to a specific haplotype are shown in bold. BAC clone *Not*I enzymatic digestion predicted size is indicated in parentheses. BAC clone size determined by shotgun subclone sequence reconstruction (35, 10, 64, 39) or position of BAC clone end sequences within the haplotype 2 shotgun subclone sequence reconstruction is indicated next to each BAC clone name.

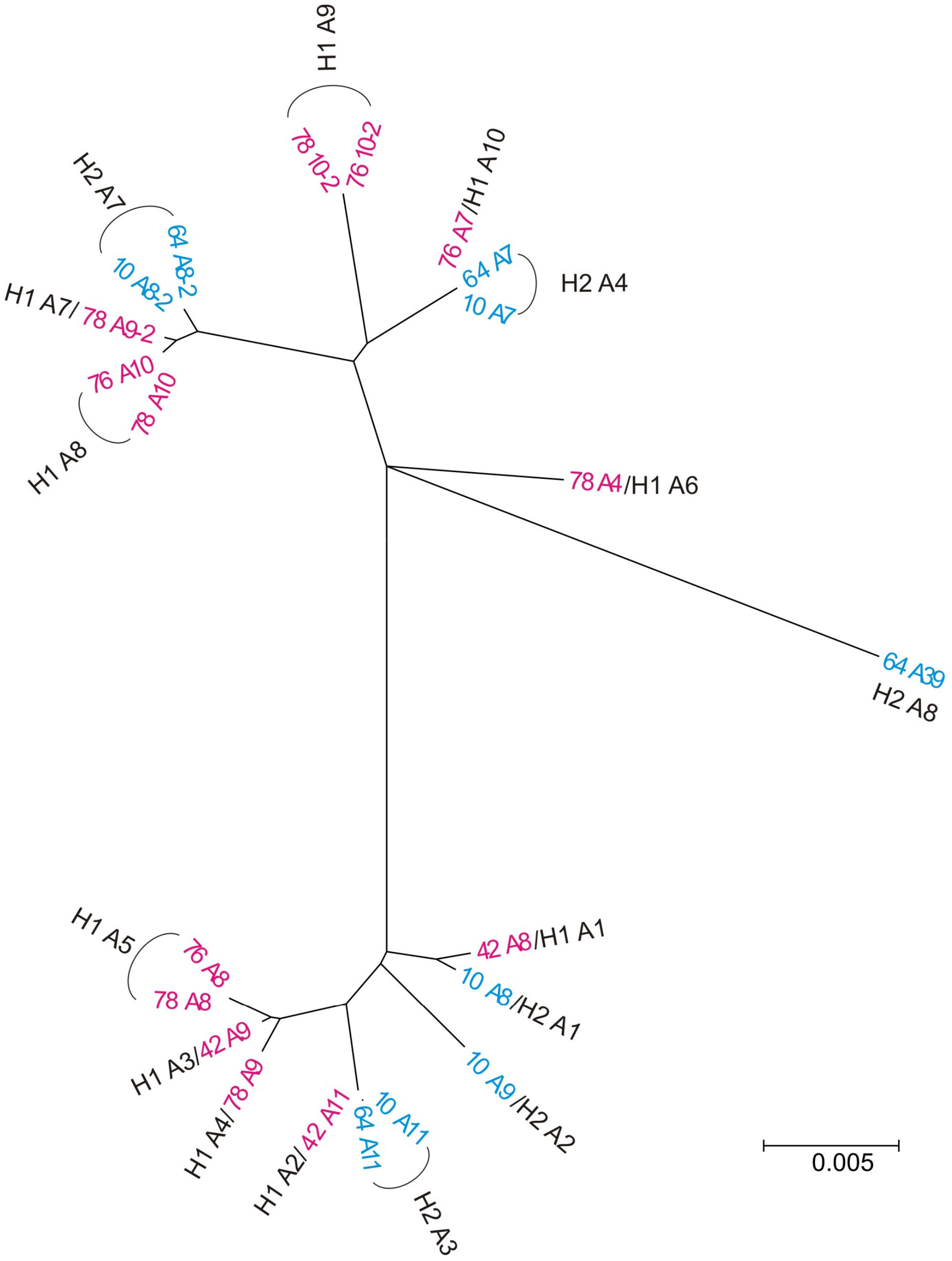


Figure 12. Neighbor-joining phylogenetic analysis of all minimal tiling path DmBAC AFGP genes. Gene sequence obtained from BAC clone alignments were aligned in the clustal function of MEGAv4. The consensus tree represents one run of 1000 bootstrap replicates, where nodes have bootstrap values >50. AFGP genes within each BAC clone are named according to the BAC clone in which they are found and the last mature AFGP peptide encoded in exon 2. Genes within haplotype 1 are colored pink, and haplotype 2 are blue. Homologous genes represented twice in the sequence assembly due to redundant BAC clone overlap are connected by a black line. The black text next to each BAC clone name shows the AFGP clone names in the haplotypic consensus alignment after omission of redundant genes. H1 corresponds to haplotype 1. H2 refers to haplotype 2.

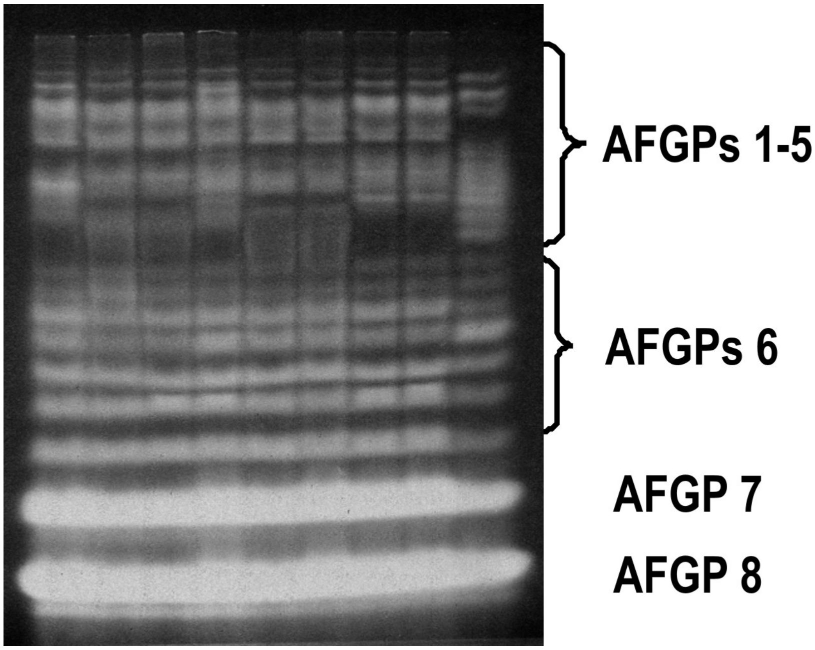


Figure 13. **Non-denaturing gel of purified serum AFGP.** Each lane of the gel shows purified serum from different *Dissostichus mawsoni* specimens. AFGP serum content is composed predominantly of AFGPs 7 and 8 (four and five Ala-Ala-Thr repeats). The size and amount of larger mature AFGP peptides varies within *D. mawsoni*.

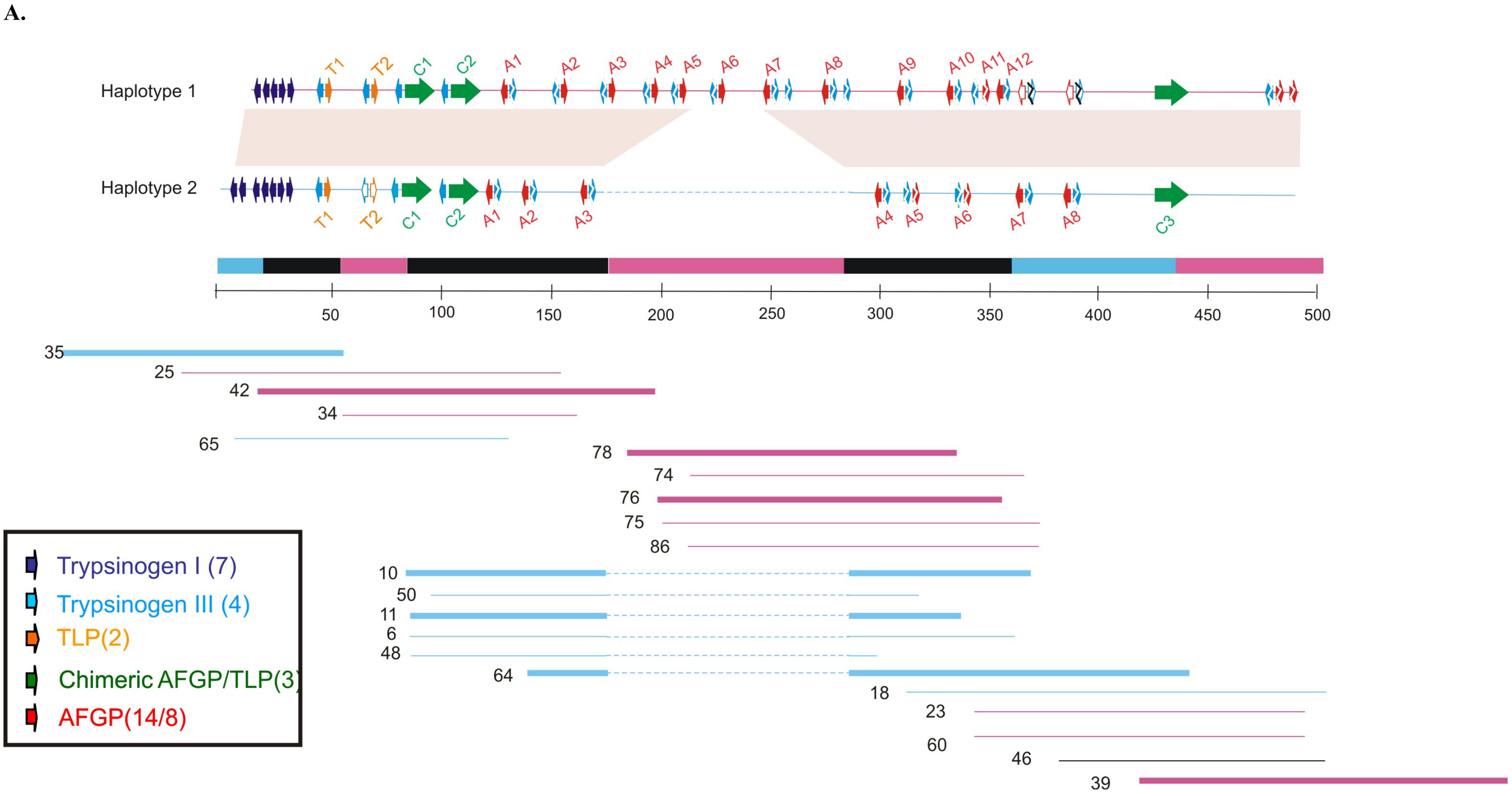


Figure 14A (see figure legend on following page)

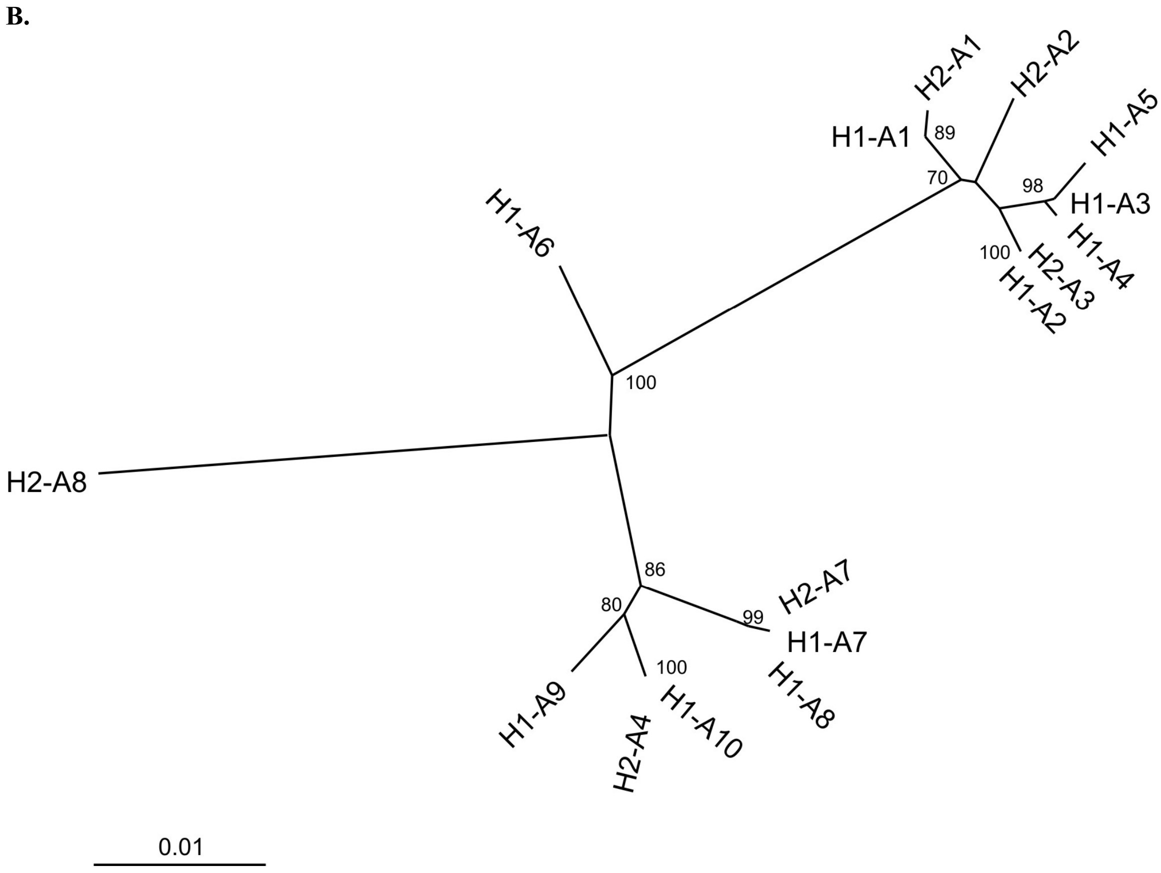


Figure 14. Linear representation of the *Dissostichus mawsoni* AFGP/TLP genomic locus. A. A cartoon representation of the AFGP/TLP locus. Genes are represented by arrows pointing toward the 3' end of the gene. Solid arrows indicate genes verified to be present within a haplotype by genomic sequencing, outline arrows indicate genes inferred to be present by PCR analysis of bacterial artificial chromosome (BAC) clones within that region. Pseudogenes are indicated by broken arrows. Each gene type is color coded according to the key. Gene naming corresponds to gene type TLP (T), chimeric (C), and AFGP (A) and the order in which they are encountered in the locus proceeding from left to right. The locus consists of five distinct gene types, three trypsinogen as well as AFGP and AFGP/TLP chimeric genes. The names and positions of BAC clones used to generate the contiguous locus sequence are indicated below the alignment (bold lines). BAC clone positions identified by BAC end sequence pair and matching are also indicated (non-bolded lines). The AFGP/TLP locus is represented by two haplotypes in the *D. mawsoni* genome, termed haplotype 1 (pink-500 kbp) and haplotype 2 (blue-400 kbp). The gap created by size variation between the two haplotypes is indicated by a dashed line in haplotype 2 BAC clones. A consensus bar located below the alignment indicates from which haplotype the genomic sequence was obtained. Black consensus regions represent genomic sequence from both haplotypes. Homologous regions are indicated by shading connecting the two haplotypes. Within haplotypes gene content remains fairly stable with the exception of the AFGP genes which vary (14-haplotype 1, 8-haplotype 2). **B.** AFGP gene homology was inferred by maximum likelihood using the algorithm of PhyML. The tree was constructed using signal peptide (exon 1) and intron 1 sequence shared by all AFGP genes.

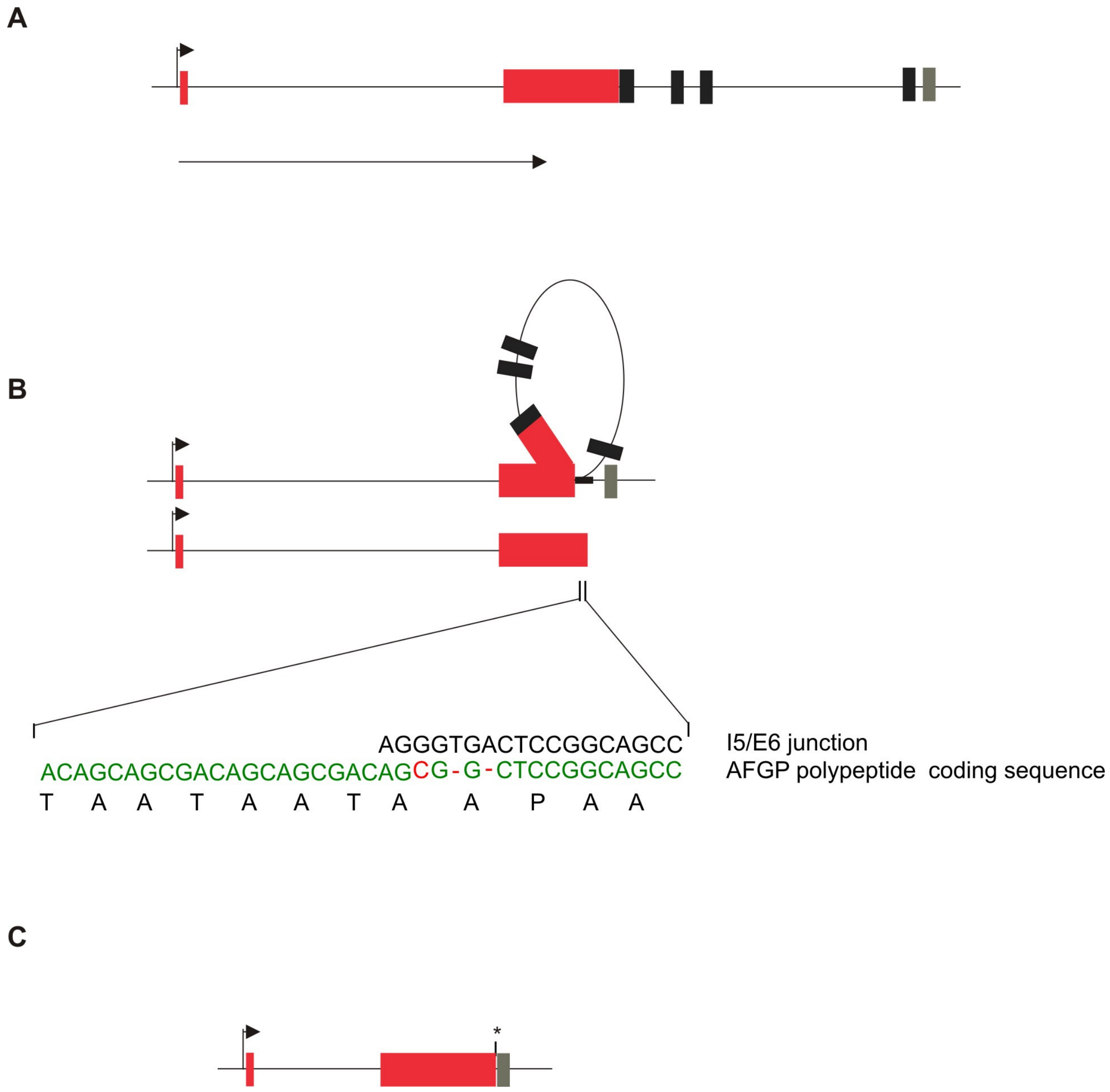

A
B
AGGGTGACTCCGGCAGCC
ACAGCAGCGACAGCAGCGACAGCG-G-CTCCGGCAGCC
T A A T A A T A A P A A
I5/E6 junction
AFGP polypeptide coding sequence
C
*

Figure 15. AFGP gene genesis. A. Cartoon depiction of the AFGP/TLP chimeric gene. Chimeric exons shared with the AFGP gene are shown as red boxes. Black boxes represent exons unique to chimeric gene, while the gray box is an exon in the chimeric gene and an exonic remnant in the AFGP gene. Introns and flanking sequence are shown as lines. The direction of polymerase replication during genome duplication is indicated by an arrow below. **B.** Replication of the template strand (top) results in nascent strand synthesis (bottom) within the replication bubble. A polymerase stall on the highly repetitive exon 2 coding sequence, results in intron 5 (I5)/exon 6 (E6) sequence (template strand) interaction with synthesized exon 2 sequence (nascent strand) within the replication bubble. The interaction of chimeric I5/E6 nucleotide sequence (black) with a hypothetical AFGP exon 2 nucleotide sequence (green translated below) is also shown. Red lettering within the AFGP sequence indicate mismatches in the alignment. **C.** Resumption of template synthesis after the I5/E6 interaction results in deletion of a portion of exon 2 through intron 5 in the chimeric gene, and incorporation of a stop codon, giving rise to the present AFGP gene structure.

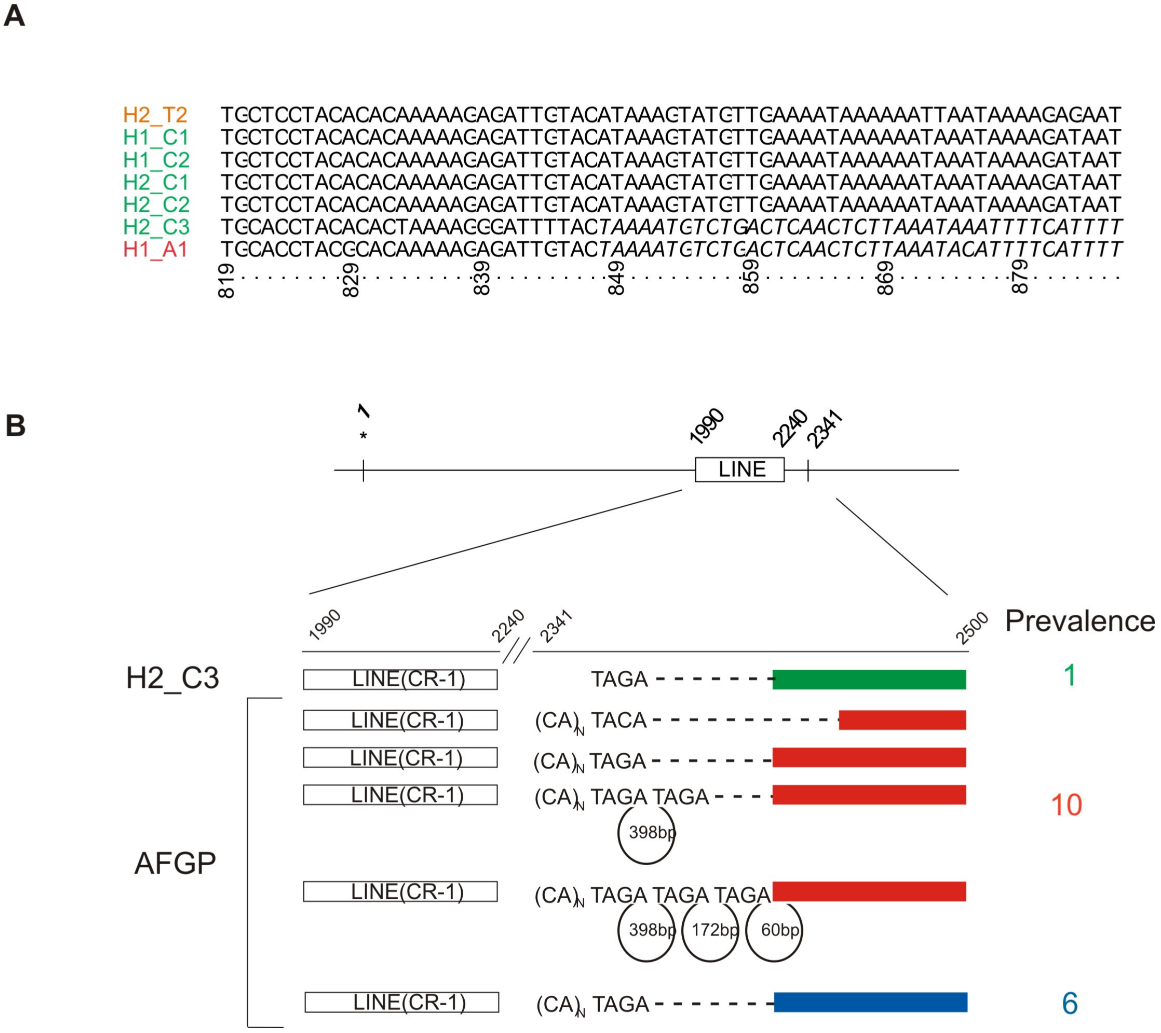

A
H2_T2 TGCTCCTACACACAAAAAGAGATTGTACATAAAGTATGTTGAAAATAAAAAATTAATAAAAGAGAAT
H1_C1 TGCTCCTACACACAAAAAGAGATTGTACATAAAGTATGTTGAAAATAAAAAATAAATAAAAGATAAT
H1_C2 TGCTCCTACACACAAAAAGAGATTGTACATAAAGTATGTTGAAAATAAAAAATAAATAAAAGATAAT
H2_C1 TGCTCCTACACACAAAAAGAGATTGTACATAAAGTATGTTGAAAATAAAAAATAAATAAAAGATAAT
H2_C2 TGCTCCTACACACAAAAAGAGATTGTACATAAAGTATGTTGAAAATAAAAAATAAATAAAAGATAAT
H2_C3 TGCACCTACACACTAAAAGGGATTTTAC*TAAAATGTCTGACTCAACTCTTAAATAAATTTTCATTTT*
H1_A1 TGCACCTACGCACAAAAAGAGATTGTAC*TAAAATGTCTGACTCAACTCTTAAATACATTTTCATTTT*
819 829 839 849 859 869 879
B
1
*
1990
2240
2341
LINE
1990
2240
2341
2500
Prevalence
H2_C3
LINE(CR-1)
TAGA
1
AFGP
LINE(CR-1)
(CA)N TACA
LINE(CR-1)
(CA)N TAGA
LINE(CR-1)
(CA)N TAGA TAGA
398bp
10
LINE(CR-1)
(CA)N TAGA TAGA TAGA
398bp
172bp
60bp
LINE(CR-1)
(CA)N TAGA
6

Figure 16. *Dissostichus mawsoni* breakpoint analysis. A. Alignment of AFGP (H1_A1), TLP (H2_T2), and chimeric (H1_C1, H1_C2, H2_C1, H2_C2, H2_C3) breakpoint junctions from haplotype 1 (H1) and haplotype 2 (H2) obtained from sequencing of the *Dissostichus mawsoni* AFGP/TLP genomic locus. The numbering below the alignment corresponds downstream sequence of all three gene types. Sequence numbering begins at the AFGP/TLP chimeric,TLP, and AFGP E6 remnant stop codon. Downstream sequence is included for all chimeric genes, one TLP, and one AFGP gene. H1_A1 AFGP gene downstream sequence is representative of all AFGP sequence at this position. The AFGP/H2_C3 breakpoint from all other AFGP/TLP chimeric and the TLP downstream sequence is 846 nt downstream of the AFGP/TLP chimeric and TLP stop codon. The breakpoint junction identified by variation in sequence identity between aligned sequences is indicated by italics. **B.** A cartoon depiction of the H2_C3 and AFGP downstream sequence. The asterisk indicates the stop codon of the H2_C3 chimeric gene. A detailed cartoon alignment of all variations of AFGP/H2_C3 breakpoint region is shown below. Numbering corresponds to H2_C3 downstream sequence. Dashes within the alignment represent gaps created by the sequence alignment. Variations in colored boxes (green-chimeric, red and blue-AFGP) indicate breakpoints junctions between alignments and subsequent variations in sequence identity (sequence identity < 50%). The number of genes containing each breakpoint type is indicated to the right. The H2_C3 breakpoint is located 100 bp downstream of a line remnant. Rearrangement, most likely involving this line remnant deposited a CA_N stretch in the AFGP downstream sequence. Repeated breakage at the CA_N stretch (two of which can be observed) resulted in increased AFGP gene dosage.

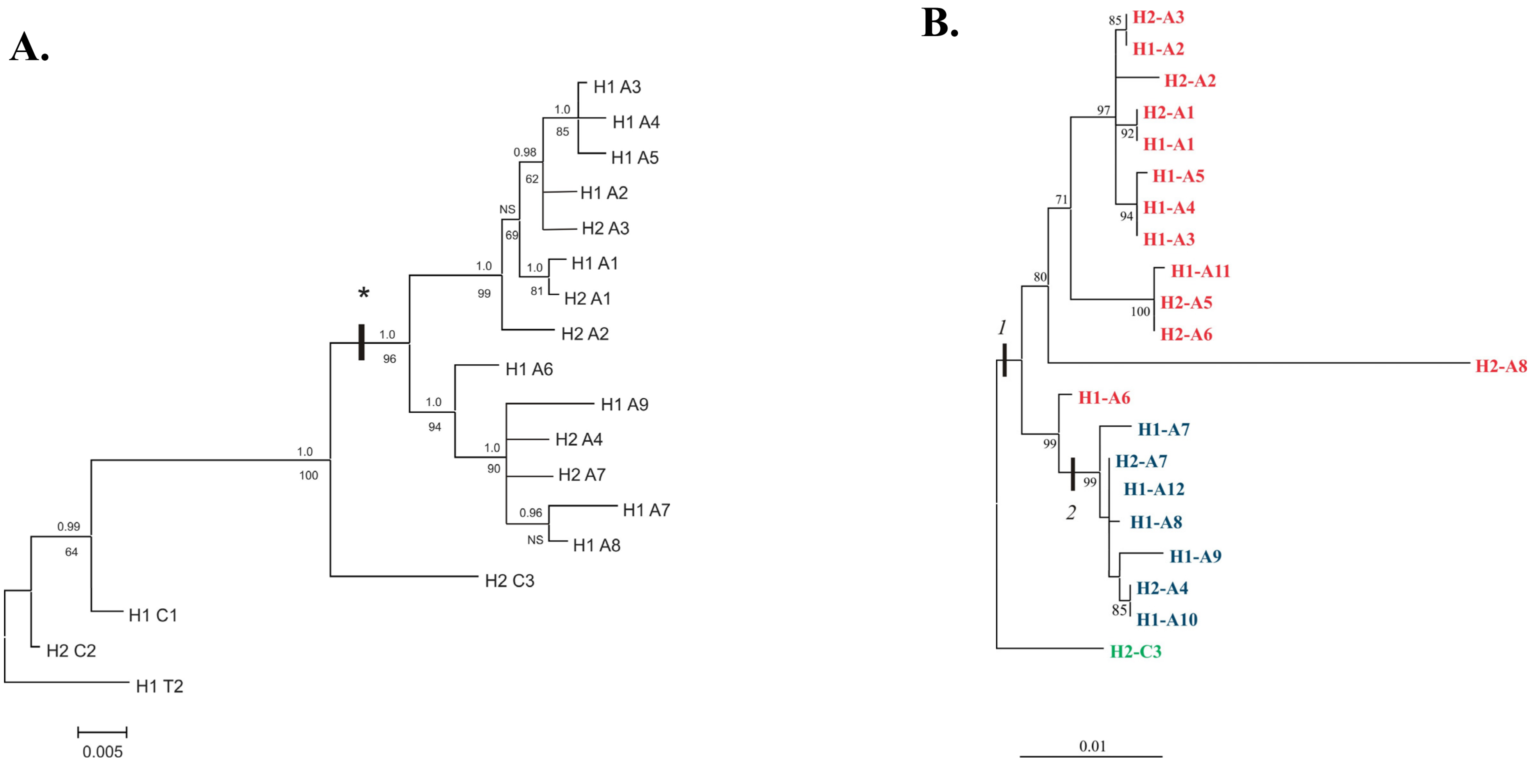


Figure 17. AFGP gene family phylogeny. A. Maximum likelihood and Bayesian trees indicating phylogenetic relatedness of AFGP (H1_A or H2_A), TLP (H1_T2), and chimeric (H1_C1, H2_C2, and H2_C3) genes. Phylogenetic inference was inferred using concatenated nucleotide sequence of signal peptide, intron1, exon 6 and 3' flanking sequence. The tree was rooted with H1_T2. Asterisk shows AFGP gene genesis. Bayesian posterior probabilities are indicated above the branch while maximum likelihood bootstrap replicates are below. NS= not supported. **B.** Maximum likelihood tree rooted with the AFGP/TLP chimeric gene that shares a most recent common ancestry with the AFGP lineage. Inference was drawn from the 1.9 kbp downstream sequence shared between all genes in the tree. Asterisk shows AFGP gene genesis.

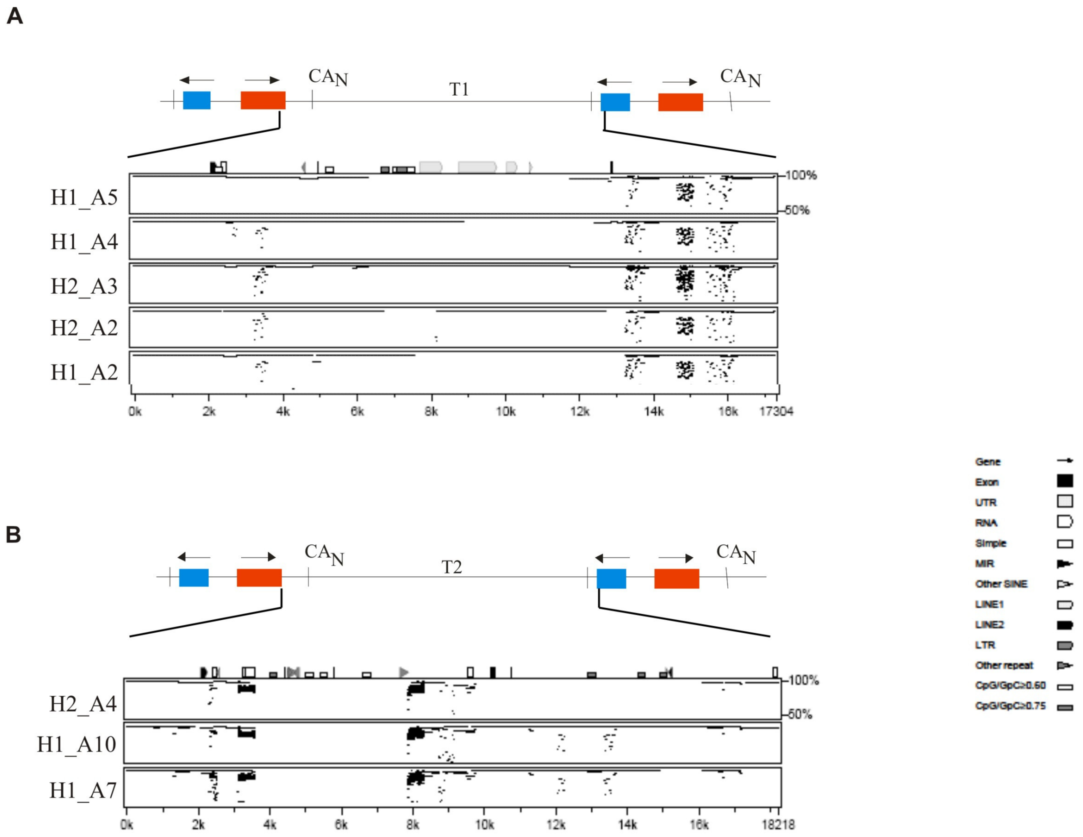


Figure 18. AFGP genes are present within two types of flanking sequence. A cartoon depiction of AFGP genes. AFGP Type 1 (A) and Type 2 (B) breakpoint segmental duplication alignments and output were made using multi PIP maker under the default conditions. Blue boxes correspond to trypsinogen 3 pseudogenes and red boxes correspond to AFGP genes. Arrows point toward the 3' end of the genes. The alignment for each duplicated module was aligned according to a reference (H1_A3 for Type 1 and H1_A8 for Type 2). Shared sequence identity is indicated by a horizontal line in each segment. The height of the line indicates percent sequence identity between the reference and each individual segment according to the scale bar on the right. Motifs identified by PIP maker are present within the reference. Regions of shared sequence identity between reference and subject suggest the presence of said motifs in the subject sequence. Icons representing motifs identified by PIP maker can be found in the key. Segmental duplicates share > 95% similarity within duplicates. Most gaps are the product of insertions or deletions, however some gaps are artificial resulting from gaps in contig alignment.

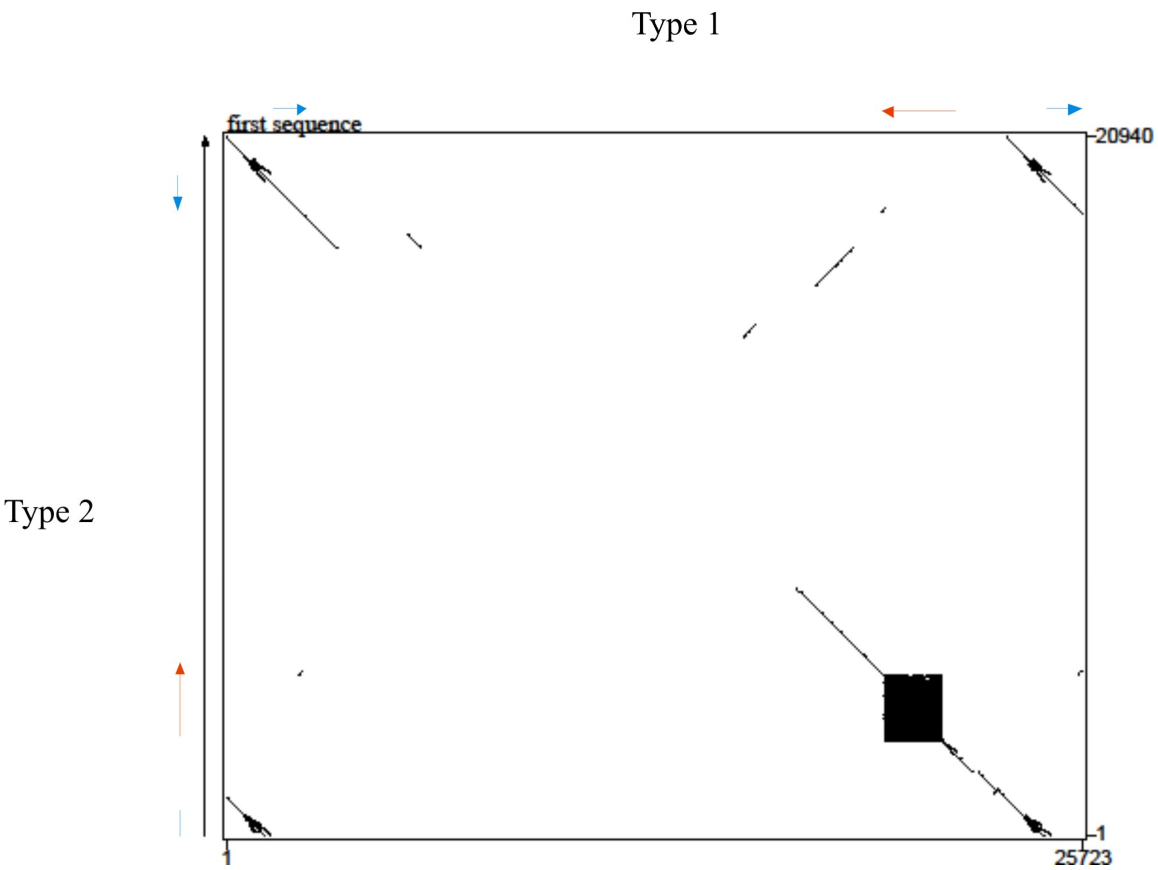


Figure 19. Type 1 and Type 2 AFGP gene containing modules contain very little sequence identity. AFGP Type 1 and Type 2 module alignments and dot plot outputs were obtained from PIP maker. Red arrows indicated AFGP gene location and blue arrows indicated trypsinogen pseudogene location. The Type 1 reference sequence corresponded to H1_A3 while the Type 2 reference sequence corresponded to the H1_A10 AFGP gene. Both sequences include downstream sequence from the stop codon of the AFGP gene until the stop codon of the proceeding trypsinogen pseudogene. Gaps in the dot plot alignment were created by low sequence identity (<50%) between the Type 1 and Type 2 AFGP gene containing modules.

H1 A5: **4-PALNY-4-AALIF-4- PALIF-4-PALIF-5-AAFNF-4-PAFIF-4-PALNF-4-PALNF-4-AALNF-4-AAFNF-4-PALIF-4-PALIF-4-PALIF-4-PALNF-4-AALNF-4-PALNF-4-PALMF-4-AAFNF-4-PALIF-4-PALIF-4-PALIF-AAFNF-4-PALIF-4-AALNF-4-AALNF-4-AAFNF-4-AAFNF-4-PALMF-4-PALIF-4-PACNF-4-PALIF-4-PALIF-8**

H2 A2: **6-PALNY-4-AALIF-4-PALIF-4-PALNF-4-PALIF-5-AAFNF-4-PAFIF-3-PALNF-4-AALNF-4-PALNF-4-PALNF-4-AALNF-4-PALNF-5-AALNF-4-PALNF-5-AALNF-4-PALNF-5-AALNF-4-PALNF-5-PALNF-4-PALNF-4-AALNF-4-PALNF-5-PALNF-5-PALNF-5-PALNF-5-PALNL-5-AALNF-4-PALNF-5-AALNF-4-PALNF-5-PALNF-5-AALNF-4-AALNF-4-AALNF-4-PALNF-7-PALNF-5-PALNF-4-AALNF-4-AALNF-4-PALNF-5-AALNF-4-PALNF-5-AALNF-4-PALNF-5-AALNF-4-PALNF-4-PALMF-4-PALNL-4-PALMF-4-PALIF-4-AAFNF-4-PALIF-4-PALIF-4-AAFNF-4-PALIF-4-AALNF-4-PALNF-4-PALMF-4-PALIF-3-AALIF-4-PALIF-4-PALNL-4-PALMF-4-PALIF-4-AAFNF-4-PALIF-4-PALIF-9**

H1A1: **6-PALNY-4-AALIF-4-PALIF-4-PALIF-4-PALNF-4-PALIF-4-AAFNF-4-PAFIF-4-PALNF-5-AALNF-4-PALNF-4-PALMF-4-PALNL-4-PALMF-4-PALIF-4-AAFNF-4-PALIF-4-AALIF-4-AAFNF-4-PALIF-4-AALNF-4-PALNF-4-PALMF-6-PALNL-4-PALMF-4-PALIF-4-AAFNF-4-PALIF-4-AALIF-4-AAFNF-4-PALIF-4-AALNF-4-AALNF-4-PALIF-8**

Figure 20. Numerical representation of select AFGP coding sequence. AFGP coding sequences consists of series of (Ala-Ala-Thr)$_N$ separated by a three residue linker sequence. Ala-Ala-Thr repeats are represented numerically. Linker sequence is colored, while the linker sequence associated two residue Ala/Pro-Ala remnant are in black to further illustrate block duplication. Areas of block repeat expansion are underlined to illustrate expansion between genes (H1 A5 and H2 A2) and within a gene (H1 A1).

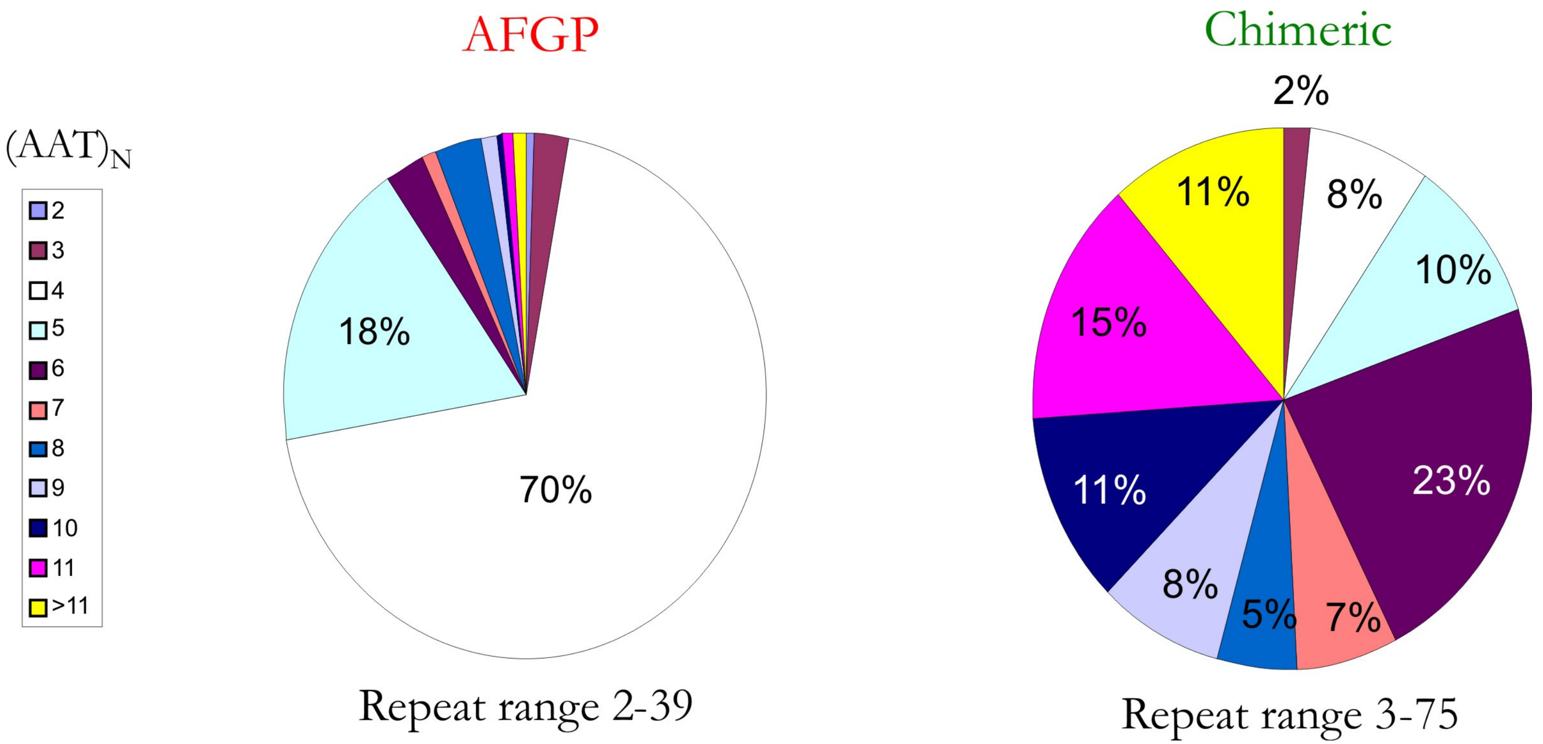


Figure 21. AFGP and chimeric polypeptide gene abundance. Polypeptides encoded within AFGP and chimeric exon 2 are displayed as a percent of total polypetides encoded per gene type. Polypeptide sizes are color coded according to the key. AFGP and chimeric exon 2 encodes polypeptides ranging in size from 2 to 78 $(AAT)_{N}$. AFGP peptides contain primarily 4 and 5 repeat, while chimeric genes contain larger polypeptides.

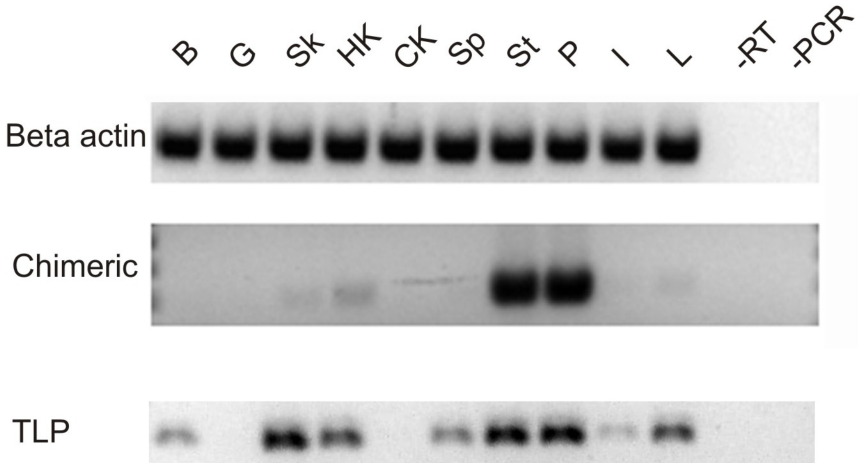


Figure 22. Chimeric transcript tissue distribution pattern. The tissue distribution of chimeric, TLP and chimeric, and beta actin RT-cDNA transcript are shown as an inverted image, to allow better visualization of expression pattern across *Dissostichus mawsoni* tissues tested. The expression patterns shown are representative of three independent samples. –RT and –PCR refer to negative controls for RT and PCR reactions, respectively. B- Brain, G-Gill, Sk-Skin, HK-Head Kidney, CK-Caudal Kidney, Sp-Spleen, St-Stomach, P-Pancreas, I-Intestine, L-Liver

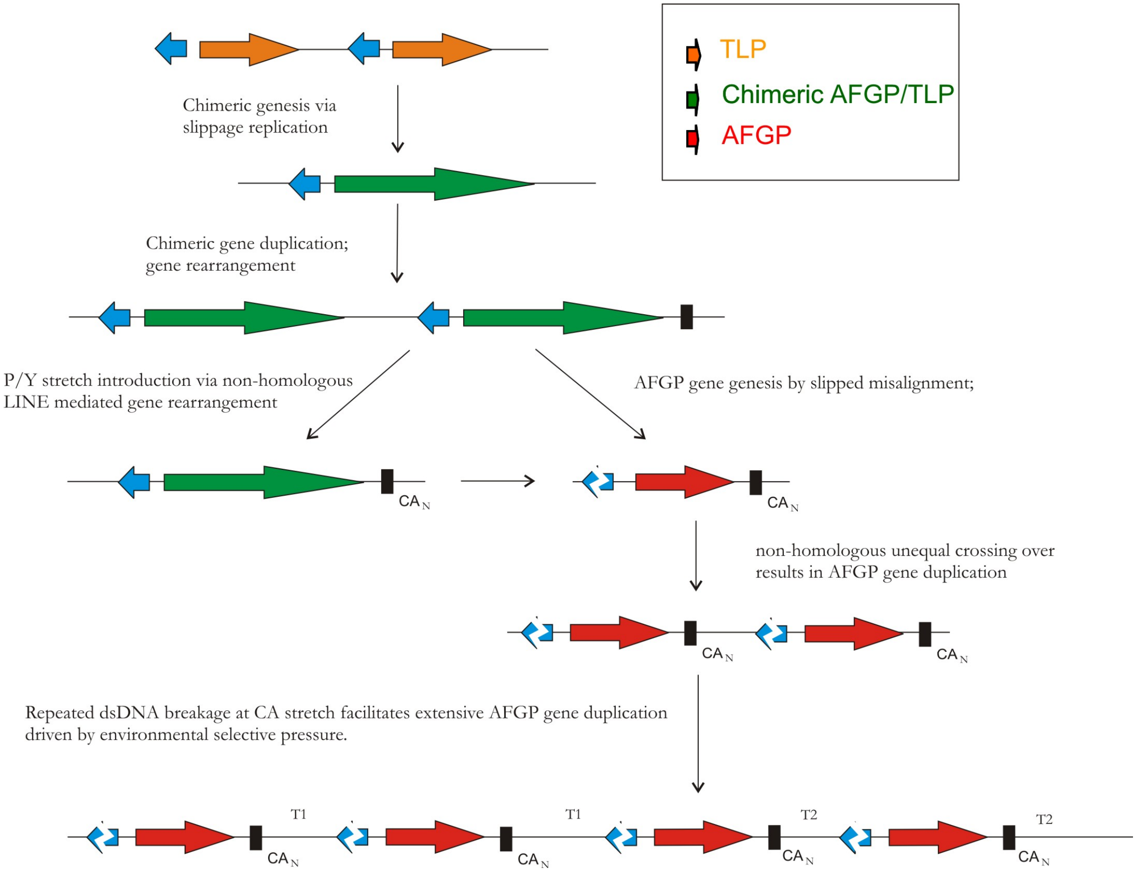

TLP
Chimeric AFGP/TLP
AFGP
Chimeric genesis via slippage replication
Chimeric gene duplication; gene rearrangement
P/Y stretch introduction via non-homologous LINE mediated gene rearrangement
AFGP gene genesis by slipped misalignment;
CA N
non-homologous unequal crossing over results in AFGP gene duplication
Repeated dsDNA breakage at CA stretch facilitates extensive AFGP gene duplication driven by environmental selective pressure.
T1
T1
T2
T2

Figure 23. Summary of AFGP/TLP locus expansion. Arrows represent genes. Genes are color coded according to the key. Type 1 (T1) and Type 2 (T2) AFGP gene flanking sequences are indicated as such.

```
                 *  20  (16)  *  40  *  60 (57)  *  80  *  (87) 100  (102) *  120  *  140 (136)  (142)  *
Dmaw_1a   : MMSLVFILLIG--AAFATEEDKIVGGKECTPYSMPHQVSLNSGYHFCGGSLVNENWVVSAAHCYKS--RVEVHLGEHNLRVKEGNEQYISSSRVIRHPNYNSYNIDNDIMLIKLSKPATLNQYVQAVALPSSCAPAGTMCTVSGWGSTQS : 146
Dmaw_1b   : MMSLVFILLIG--AAFATEEDKIVGGKECTPYSMPHQVSLNSGYHFCGGSLVNENWVVSAAHCYKS--RVEVHLGEHNLRVKEGNEQYISSSRVIRHPNYNSYNIDNDIMLIKLSKPATLNQYVQAVALPSSCAPAGTMCTVSGWGSTQS : 146
Dmaw_1c   : MRSLVFVLLIG--AAFATEEDKIVGGKECTPYSMPHQVSLNSGYHFCGGSLVNADWVVSAAHCYKT--RVEVQLGEHNFRVTEGNEQYISSSRVIRHPNYNSYNIDNDIMLIKLSKPATLNQYVQPVALPSSCAPAGTMCTVSGWGSTMS : 146
Dmaw_1d   : MKSLVFVLLIG--AAFAAEDDKIVGGYECTPHSQPHQVSLNIGYHFCGGSLVNENWVVSAAHCYKS--RIEVRMGEHHIGVTEGNEQFISSLSVITHPYYDRYSLTNDIMLIKLSKPATLNQYVQPVALPSSCAPAGTMCTLSGWGNTMS : 146
Dmaw_1e   : MMSLVFILLIG--AAFATEEDKIVGGKECAPYSMPYQVSLNSGYHFCGGSLVNENWVVSAAHCYKS--RVEVRMGEHNIRVTEGNEQFISSSRVIRHPNYSSYNIDNDIMLIKLSKPATLNQYVKPVALPRSCAPAGTMCTVSGWGSTQS : 146
Dmaw_1f   : MMSLVFILLIG--AAFATEEDKIVGGKECTPYSMPHQVSLNSGYHFCGGSLVNENWVVSAAHCYKS--RVEVRMGEHHIRVTEGNEQFISSSRVIRHPNYNSYNIDNDIMLIKLSKPATLNQYVKPVALPRSCAPAGTMCTVSGWGSTQS : 146
Dmaw_1g   : MKSLVFALLLG--AVFATEDDKIVGGQECVPHSQPHQVSLNSGYHFCGGTLVNENWVVSAAHCYNK---MDIVLGDHNRWFMDGNEQIISAERVIPHPNYESWLVNNDIMLIKLSQPATLNKYVQPVALPSSCAPAGTMCTVSGWGVTMS : 145
Dmaw_3a   : MIGLIVLALLG--AAAPMD-DKIVGGFQCTAHSQPWQVSINLGYHYCGGSLINDQWIVSAAHCWQNPYSLIAILGDNHIWMNEGTEQFMSVDAIYWHQSYDYQTLDYDIMLMKLAHPVTVNQYVKPVALPKACPAAGDMCMVSGWGNIYT : 147
Dmaw_3b   : MIGLIVLALLG--AAAPMD-DKIVGGFQCTAHSQPWQVSINLGYHYCGGSLINDQWIVSAAHCWQNPYSLIAILGDNHIWMNEGTEQFMSVDAIYWHQSYDYQTMDYDIMLMKLAHPVTVNQYVKPVALPKACPAAGDMCMVSGWGNIYT : 147
Dmaw_TLP1 : MTLLALLLLIGAAAAVPREDGRIIGGYECSPHSRPYMASLNYGYHFCGGVLINNQWVLSVAHCWYNPYSMQVILGDHNLRVFEGTEQLMKTNTIIWHPSYDYQTLDFDIMLIKLYHPVEVTEAVAPIPLPTSCPYGGLSCSVSGWGIAKL : 150
Dmaw_TLP2 : MTLLALLLLIG-AAAVPREDGRIIGGYECSPHSRPYMASLNYGYHFCGGVLINNQWVLSVAHCWYNPYSMQVILGDHNLRVFEGTEQLMKTNTIIWHPSYDYQTLDFDIMLIKLYHPVEVTEAVAPIPLPTSCPYGGLSCSVSGWGIAKL : 149

                 160  *  180  (182)  *  (189)  (192) 200  (195)  *  (217) 220  (227) *  (230)  240
Dmaw_1a   : S-SADK-NKLQCLNIPILSDRDCDNSYPGQITDAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAQRDNPGVYAKVCLFNDWLETTMANY- : 242
Dmaw_1b   : S-SADK-NKLQCLNIPILSDRDCDNSYPGQITDAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAQRDNPGVYAKVCLFNDWLETTMANY- : 242
Dmaw_1c   : S-TADK-NKLQCLNIPILSDRDCDNSYPGMITDSMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAQKDNPGVYTKVCLFNNWLETTMASY- : 242
Dmaw_1d   : S-TADG-NRLQCVAVPILSYEDCDNSYPGMTDNTMFCAGYLEGGKDSCQGDSGGPVVCDGELQGVVSWGYGCAEKNHPGVYSKVCVQTEWLHNTMATY- : 242
Dmaw_1e   : S-TADG-NKLQCLNIPILSDRDCDNSYPGMITDAMFCAGYLEGGKDSCQGDSGGPVVCNGELQGVVSWGYGCAERDNPGVYAKVCLFNDWLETTMASY- : 242
Dmaw_1f   : S-SADG-NKLQCLNIPILSDRDCDNSYPGQITDAMFCAGYLEGGKDSCQ-YSGGPVVCNGELQGVVSWGYGCAQRDNPGVYAKVCLFNDWLETTMANY- : 241
Dmaw_1g   : S------GELQCLNIPILSRENCDNSYPDMITDAMFCAGDLEGGKDSCQGGSGGPVVCDGELQGVVSWGFGCAEKNQPGVYAKTCIFTDWLQSTMASY- : 237
Dmaw_3a   : D-QVFNPFYLQCVEVPILSHKDCDGSYPGMITDRMVCAGYLEGGKDSCQGDSGGPLVCNGELQGVVSWGQGCAQPNYPGVYTKVCSLMPWINDILSTYS : 245
Dmaw_3b   : D-QVFNPFYLQCVEVPILSHKDCDGSYPGMITDRMVCAGYLEGGKDACQGDSGGPLVCNGELQGVVSWGQGCAQPNYPGVYTKVCSLMPWINDILSTYS : 245
Dmaw_TLP1 : GGEAYMPTLLQCLNVPIVDQQVCENTYPGLISTRMVCAGYMEGGKDACNGDSGSPLVCDGEVQGLVSWGQGCAEPNYPGVYVKLCEFHSWFEEVLAANP : 249
Dmaw_TLP2 : GGEAYMPILLQCLNVPIVDQQVCENAYPGMISTRMVCAGYMEGGKDACNGDSGSPLVCDGEVQGLVSWGQGCAEPNYPVVYVKLCEFHSWFEEVLAANP : 248
```

Figure 24. Clustal X alignment of *Dissostichus mawsoni* trypsinogen amino acid sequence. Amino acid sequence was translated from genomic sequence obtained from *D. mawsoni* AFGP/TLP genomic locus construction. Sequences were aligned in Clustal X under default conditions. Residues within the catalytic triad are darkly shaded, and conserved cysteine residues are bolded. Residues known to affect binding in the specificity pocket are lightly shaded. The chymotrypsin numbering system is in parenthesis.

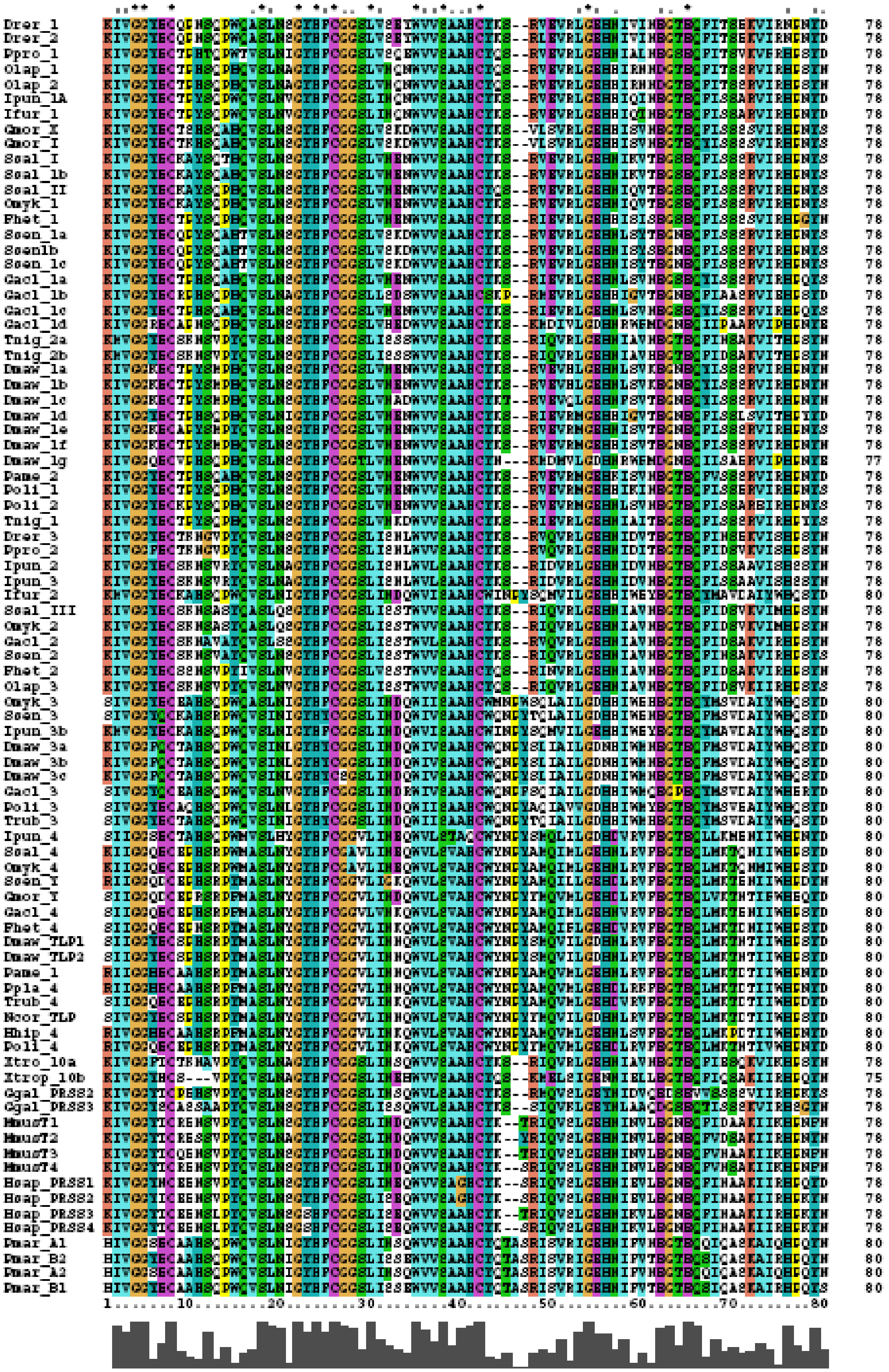

Drer_1 78
Drer_2 78
Ppro_1 78
Olap_1 78
Olap_2 78
Ipun_1A 78
Ifur_1 78
Gmor_X 78
Gmor_I 78
Ssal_I 78
Ssal_1b 78
Ssal_II 78
Omyk_1 78
Fhet_1 78
Ssen_1a 78
Ssen1b 78
Ssen_1c 78
Gacl_1a 78
Gacl_1b 78
Gacl_1c 78
Gacl_1d 78
Tnig_2a 78
Tnig_2b 78
Dmaw_1a 78
Dmaw_1b 78
Dmaw_1c 78
Dmaw_1d 78
Dmaw_1e 78
Dmaw_1f 78
Dmaw_1g 77
Pame_2 78
Poli_1 78
Poli_2 78
Tnig_1 78
Drer_3 78
Ppro_2 78
Ipun_2 78
Ipun_3 78
Ifur_2 80
Ssal_III 78
Omyk_2 78
Gacl_2 78
Ssen_2 78
Fhet_2 78
Olap_3 78
Omyk_3 80
Ssen_3 80
Ipun_3b 80
Dmaw_3a 80
Dmaw_3b 80
Dmaw_3c 80
Gacl_3 80
Poli_3 80
Trub_3 80
Ipun_4 80
Ssal_4 80
Omyk_4 80
Ssen_Y 80
Gmor_Y 80
Gacl_4 80
Fhet_4 80
Dmaw_TLP1 80
Dmaw_TLP2 80
Pame_1 80
Ppla_4 80
Trub_4 80
Ncor_TLP 80
Hhip_4 80
Poli_4 80
Xtro_10a 78
Xtrop_10b 75
Ggal_PRSS2 78
Ggal_PRSS3 78
MmusT1 78
MmusT2 78
MmusT3 78
MmusT4 78
Hsap_PRSS1 78
Hsap_PRSS2 78
Hsap_PRSS3 78
Hsap_PRSS4 78
Pmar_A1 80
Pmar_B2 80
Pmar_A2 80
Pmar_B1 80
1.......10........20........30........40........50........60........70........80

Figure 25

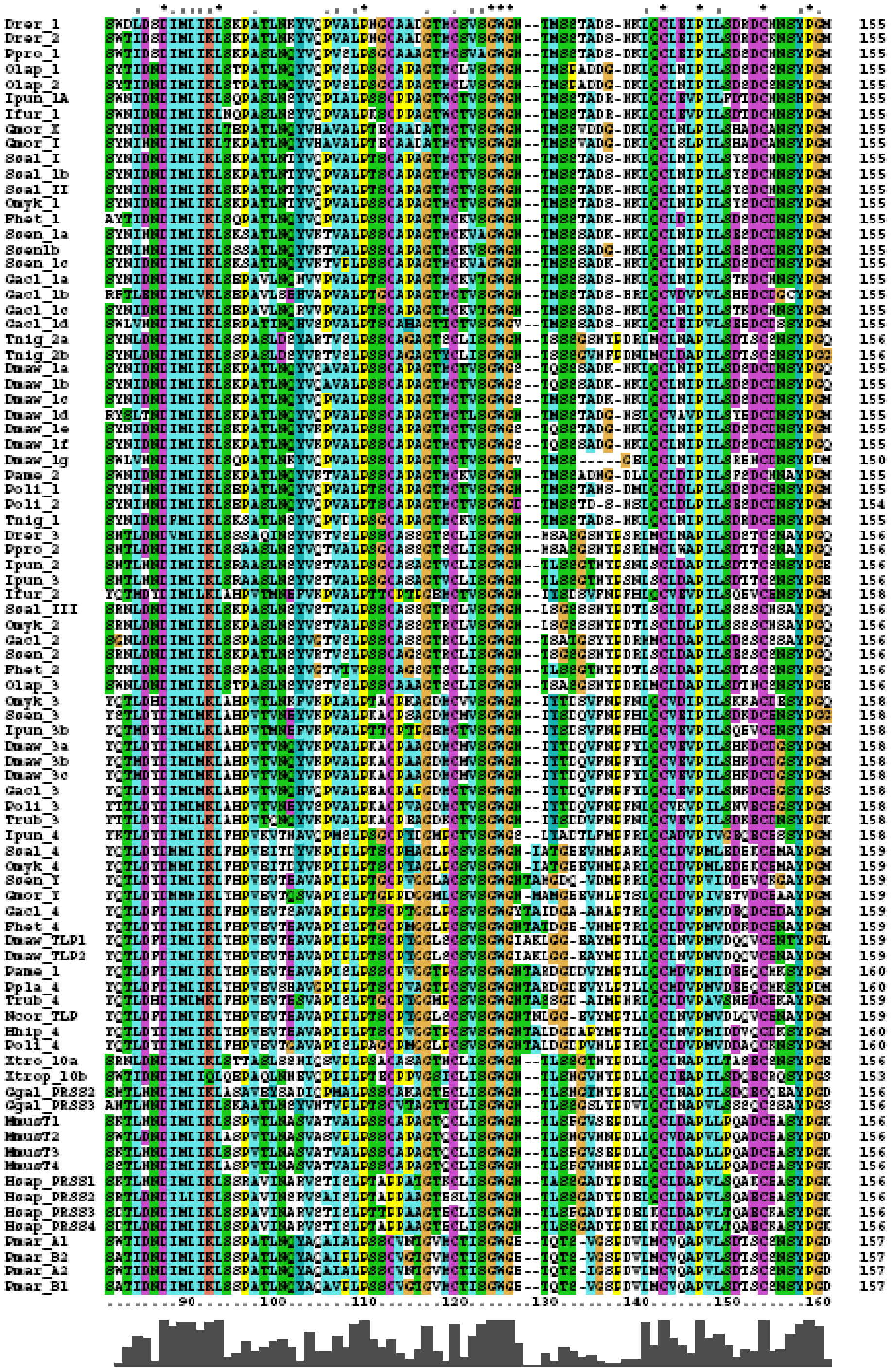

Drer_1 155
Drer_2 155
Ppro_1 155
Olap_1 155
Olap_2 155
Ipun_1A 155
Ifur_1 155
Gmor_X 155
Gmor_I 155
Ssal_I 155
Ssal_1b 155
Ssal_II 155
Omyk_1 155
Fhet_1 155
Ssen_1a 155
Ssen1b 155
Ssen_1c 155
Gacl_1a 155
Gacl_1b 155
Gacl_1c 155
Gacl_1d 155
Tnig_2a 156
Tnig_2b 156
Dmaw_1a 155
Dmaw_1b 155
Dmaw_1c 155
Dmaw_1d 155
Dmaw_1e 155
Dmaw_1f 155
Dmaw_1g 150
Pame_2 155
Poli_1 155
Poli_2 154
Tnig_1 155
Drer_3 156
Ppro_2 156
Ipun_2 156
Ipun_3 156
Ifur_2 158
Ssal_III 156
Omyk_2 156
Gacl_2 156
Ssen_2 156
Fhet_2 156
Olap_3 156
Omyk_3 158
Ssen_3 158
Ipun_3b 158
Dmaw_3a 158
Dmaw_3b 158
Dmaw_3c 158
Gacl_3 158
Poli_3 158
Trub_3 158
Ipun_4 158
Ssal_4 159
Omyk_4 159
Ssen_T 159
Gmor_Y 159
Gacl_4 159
Fhet_4 159
Dmaw_TLP1 159
Dmaw_TLP2 159
Pame_1 160
Ppla_4 160
Trub_4 159
Ncor_TLP 159
Hhip_4 160
Poli_4 160
Xtro_10a 156
Xtrop_10b 153
Ggal_PRSS2 156
Ggal_PRSS3 156
MmusT1 156
MmusT2 156
MmusT3 156
MmusT4 156
Hsap_PRSS1 156
Hsap_PRSS2 156
Hsap_PRSS3 156
Hsap_PRSS4 156
Pmar_A1 157
Pmar_B2 157
Pmar_A2 157
Pmar_B1 157
90
100
110
120
130
140
150
160

Figure 25.

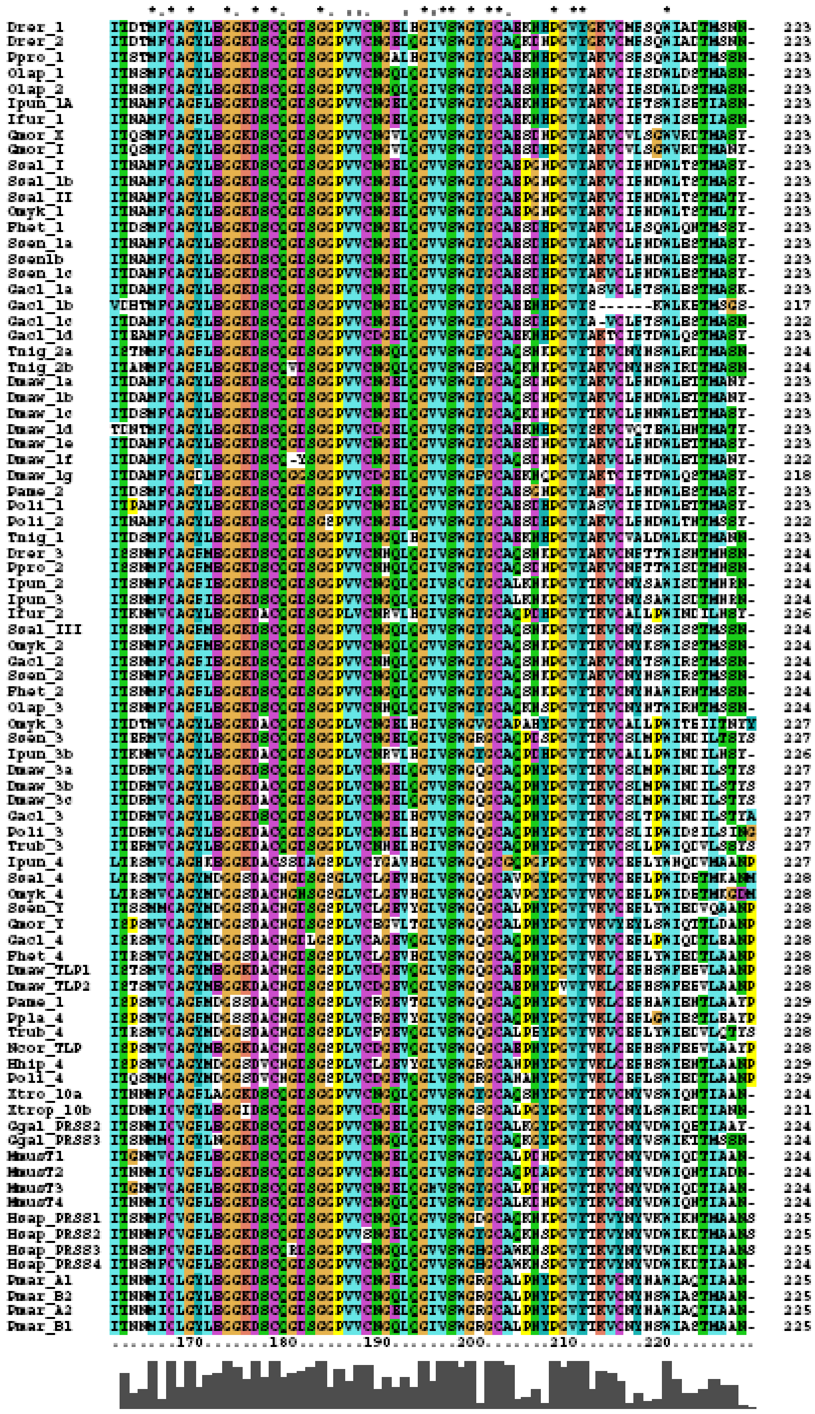


Figure 25. **Amino acid alignment of mature trypsinogen sequences.** Sequences are translations of cDNA obtained from NCBI BLAST screens of EST libraries or spliced genomic sequence (*Dissostichus mawsoni*). Alignment was performed using default settings in Clustal X. Numbering corresponds to mature trypsinogen amino acid sequence.

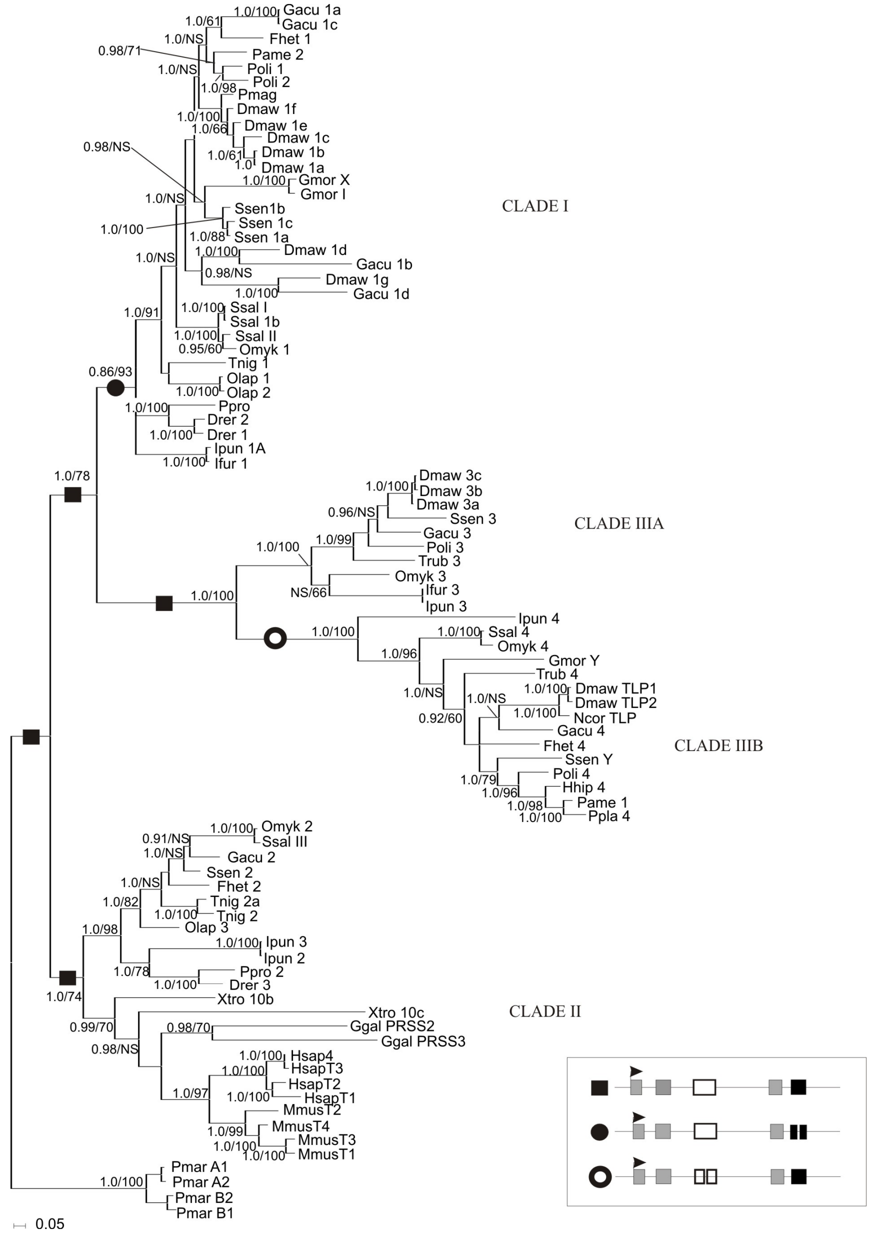


Figure 26. Trypsinogen phylogeny resulting from Bayesian analysis of cDNA nucleotide sequence, corresponding to the mature peptide. Sequences were aligned in Clustal X using default parameters. Nucleotide alignments were constrained by the amino acid sequence alignment. The tree is rooted with *Petromyzon marinus*. Numbers at nodes represent Bayesian posterior probability (top) and maximum likelihood values (bottom), nodes not supported are indicated as NS. Trypsinogens previously were previously classified into three groups, group I, group III, and group III(Roach et al. 1997; Roach 2002). All teleost group I trypsinogens are located within our clade I. Group III trypsins are located in our clade IIIA and IIIB. Our clade II is composed of teleost group II and tetrapod group I and II trypsins. The figure key corresponds to representative intron-exon motif of ancestral trypsinogen genes and present-day trypsinogen clades. Present-day trypsinogen motifs were obtained from NCBI, ENSMBL genome sequencing projects and the *Dissostichus mawsoni* trypsinogen loci alignment (SI Table 1). Ancestral motifs were determined from the current trypsinogen clade motifs, via the most parsimonious gain/loss of introns.

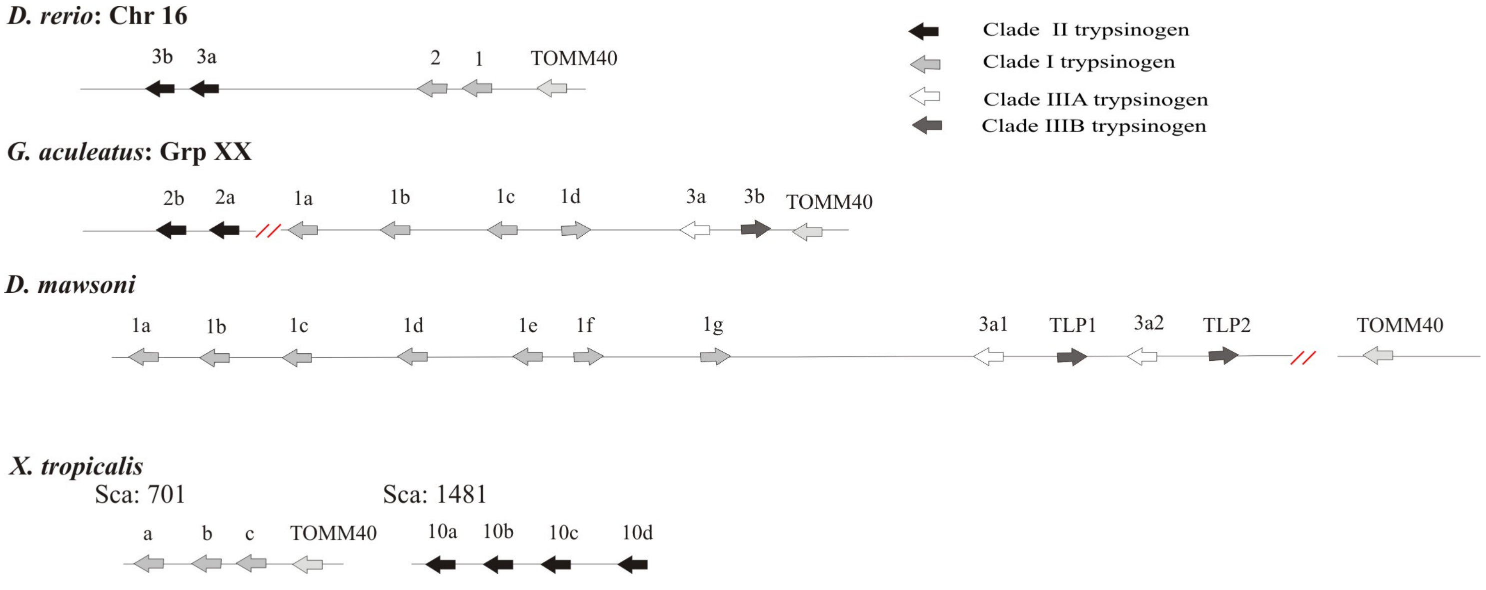


Figure 27. Digestive trypsinogen syntenic loci. Digestive trypsinogen syntenic loci of *Dissostichus mawsoni, Gasterosteaus aculeatus*, *Danio rerio*, and *Xenopus tropicalis*. Loci of *D. rerio* and *G. aculeatus* were obtained from ENSEMBL genome projects. Genes are indicated by arrows pointing toward the 3' end of the gene. Trypsinogen gene types correspond to the key. Genomes were screened for clade IIIB/Clade II trypsinogen gene presence and gene presence was determined by the SNAP/SCAN program in ENSEMBL. *D. mawsoni* trypsinogen locus and Translocase of outer mitochondrial membrane 40 (TOMM40) were found during AFGP/TLP locus assembly. AFGP gene evolved from a TLP precursor, thus the gap in *D. mawsoni* represents AFGP gene containing regions. Gene presence and orientation was determined by BLASTing segments of the locus alignment, locus spans 70 kbp. Gene sequence identity and orientation were determined manually. ENSMBL accession numbers for *G. aculeatus, D. rerio,* and *X. tropicalis* can be found in Table 5.

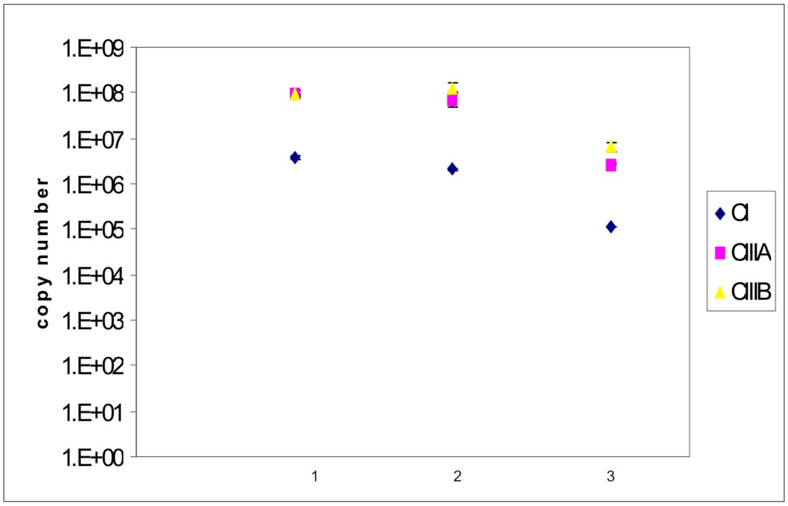


Figure 28. Absolute quantitative PCR detection of *Dissostichus mawsoni* trypsinogen transcript abundance. Pancreatic reverse transcribed RNA transcript abundance is displayed as copy number relative to a standard curve for each trypsinogen gene type, produced from known quantities of template. Clade transcript abundance was determined for three individuals in triplicate. Pancreas is not a discrete organ in teleosts, thus individual pancreatic samples are unable to be directly compared. Clade transcript values are displayed for each individual separately ± SEM indicating the overall trend observed across the three individuals tested. P values were determined using the student's T-test in Excel.

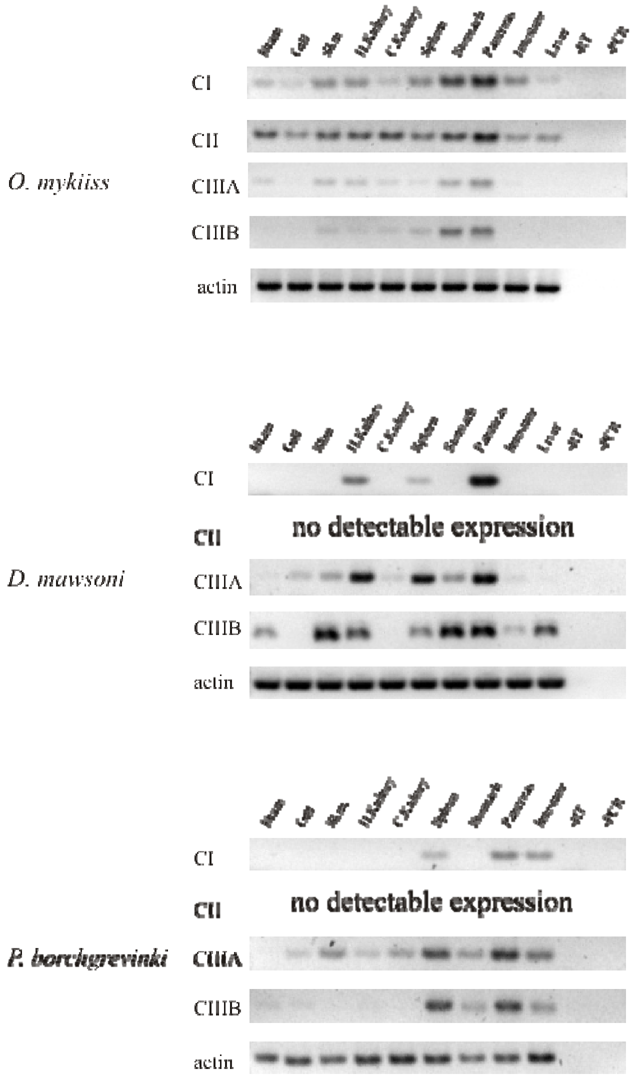


Figure 29. Multiple tissue expression analysis. An inverted digital image of trypsinogen tissue expression patterns in *Onchorynchus mykiss*, *Dissostichus mawsoni*, and *Pagothenia borchgrevinki* was used to enhance band visualization. The expression patterns shown are representative of 3 independent samples of each tissue (*D.mawsoni* and *O. mykiss*). *P. borchgrevinki* tissue distribution patterns are representative of an individual sample. Polymerase chain reaction (PCR) was performed on reverse transcribed (RT) RNA extracted from 10 different tissues in *D. mawsoni* and *O. mykiss*, as well as 9 different tissues in *P. borchgrevinki*. PCR primers specific for clades I, II, IIIA, and IIIB were used to amplify each trypsinogen clade. -RT and -PCR refer to negative controls for the RT and PCR reactions. H. kidney and C. kidney refer to head and caudal kidney samples.

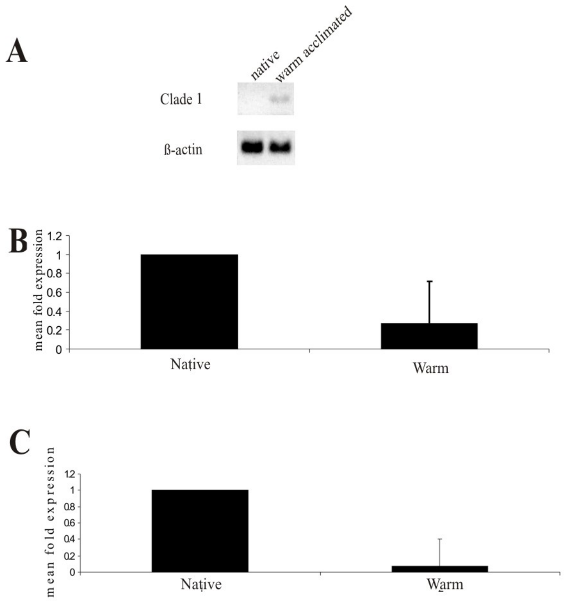


Figure 30. Relative quantitative PCR anterior stomach trypsinogen expression levels. *Pagothenia borchgrevinki* native (-1.86°C) vs warm-acclimated (4°C) trypsinogen cold-responsiveness. **A.** A digital image of clade I transcript presence, detected by polymerase chain reaction, in native and warm- acclimated *P. borchgrevinki* is present. Clade I transcript presence is displayed as an inverted digital image to allow for visualization of faint bands. The image shown is representative of three individual samples. Clade IIIA (**B**) and TLP/IIIB (**C**) transcript abundance was determined by Real-time relative quantitative PCR analysis of native and warm-acclimated anterior stomach reverse transcribed RNA. Quantitative PCR Ct values were normalized against beta-actin, and analyzed using the delta delta Ct method. Values are displayed as the mean fold transcript expression change versus native ± SEM. Clade 3A transcript was significantly lower in warm-acclimated as compared to native *P. borchgrevinki* (N=3, $p < 0.05$). Clade TLP/IIIB transcript was significantly lower in warm-acclimated as compared to native *P. borchgrevinki* (N=3, $p < 0.001$).

TABLES

Table 1. *Dissostichus mawsoni* bacterial artificial chromosome (BAC) clones not present within the AFGP/TLP locus and their putative gene presence determined by Southern Blot screening of the *D. mawsoni* BAC library.

DmBAC	BAC library Screen
4	trace TLP
13	trace TLP
15	trace TLP
16	trace TLP
17	trace TLP
20	AFGP
24	trace TLP
26	trace TLP
37	trace TLP
45	trace TLP
47	trace TLP
54	trace AFGP/TLP
67	trace AFGP/TLP
69	trace TLP
70	trace TLP
71	trace TLP
72	trace TLP
73	trace TLP
74	AFGP
75	AFGP
79	AFGP
80	AFGP
85	AFGP

Table 2. AFGP, TLP, and AFGP/TLP chimeric gene similarity.

	percent identity (nt)	E2 coding region
TLP	99	N/A
Chimeric	94-98	$(P/AAT)_{117\text{-}164}$
AFGP	84-99	$(P/AAT)_{143\text{-}403}$

Table 3. Primers used in gene expression analysis.

Target	Primers	Amplicon (bp)
Omyk clade I	5’- CTACAAGTCCCGTGTGGAG-3’ (F) 5’- ATGTCATTGTCGATGTTGTAG-3’ (R)	136
Omyk clade II	5’- GACACTCTGAGGTGCCTGGATCTC-3’(F) 5’- GGAGTCTCCCTGGCAAGAGTCCTT-3’ (R)	130
Omyk clade IIIA	5’- ATTAATGAGGGAGCCGCCACAGAA-3’ (F) 5’- TTGGCCTGATTGTACTCACACTCCTG-3’ (R)	152
Omyk clade IIIB	5’- ATCTGGCATCCCAGCTATG-3’ (F) 5’- GCATGTTTACCTCCTCTCC-3’ (R)	187
Omyk β-actin	5’- TGATAACGGCTCCGGTATGTGCAA -3’(F) *5’*- ATCCCAACCATCACTCCCTGATGT-3’ (R)	164
Dmaw clade I	5’- CTACAAGTCGTAAGTGGAGG-3’ (F) 5’- ATGTCATTGTCGATGTTGTAG-3’ (R)	131
Dmaw clade IIIA	5’- CACTGCTGGCAAAACCCTTATTCACTGAT-3’ (F) 5’- TGGTAGTCGTAGCTCTGATGCCAGTA-3’ (R)	124
Dmaw clade IIIB	5’- ATCTGGCATCCCAGCTATG-3’ (F) 5’- GCATGTAAGCCTCTCCTCC-3’(R)	186
Dmaw β-actin	5’- TGTGACCAACTGGGATGACA-3’ (F) 5’- GGGGTGTTGAAGGTCTCGAA-3’ (R)	164

Table 4. Chimeric and AFGP exon 2 linker sequence residue variability is displayed as a percentage of the total number of linker sequences variants within each gene type.

Linker	AFGP	Chimeric
LIF	41.09	37.70
LFF	0.00	59.02
FNF	18.69	
LNY	5.82	
LMF	8.64	
LNF	18.17	
CNF	0.88	
LNL	5.29	
FNL	0.53	
LIL	0.71	
other	0.00	3.28

Table 5. Accession number for trypsinogen nucleotide sequences displaying trypsinogen groupings (I, II, or III) by Roach (1997 and 2002) nomenclature with respect to the our phylogenetic clade groupings (I, II, IIIA, and IIIB).

Species name	proposed	Clade	Group	Accession
	gene name			
number				
D.rerio	1,2	I	I	NM_199605
	3a	II	II	BC092664.1
D.mawsoni	1a,b,c,d,e,f	I	I	
	3a,b	IIIA	III	
	TLP1,2	IIIB	III	
F. heteroclitus	2	II	II	EV457191.1
	4	IIIB	III	EV457297.1
G.aculeatus	1	I	I	CD507358
	2a	II	II	DT972686
	3	IIIA	III	CD499912.1
	4	IIIB	III	DW651310
G.gallus	PRSS2	II		NM_205384
	PRSS3	II		NM_205385
G.morhua	I,X	I	I	X76887.1,
				X76886.1
	Y	IIIB	III	FC075319.1
H. hippoglossus	4	IIIB	III	EU412432.1
H.sapiens	T1	II	I	NM_002769,
	T2			AAL14244
	T3			AL358573
	T4	II	I	AL358573
I.punctatus	1A	I	I	BM438246.1
	2_2	II	II	CK407942,
	3			CK407525
	3b	IIIA	III	CK422715.1
	4	IIIB	III	CK422095.1
I.furicatus	6	I	I	CK409598.1
	3	IIIA	III	CK408802.1
M.musculus	T1	II	I	NM_011645,
	T3			NM_053243,
	T4			NM_009430
N.coriiceps	4	IIIB	III	AF134323.1
O.lapites	1	I	I	AB272106
	2			NM_001104900
	3	II	II	BJ896704.1
O. mykiss	1	I	I	CU071179.1

Table 5. **continued**

	3	IIIA	III	BX077582.2
	4	IIIB	III	BX075844.2
P.americanus	1	IIIB	III	AF012462.1
	2	II		AF012463.1
	3	IIIA	III	AF012464.1
P.marinus	A1	-		AF011352.1,
	A2			AF011898
	B1	-		AF011901.1,
	B2			AF011900
P.olivaceus	1	I	I	AB029750.1
	2	II	II	EU007656.1
	3	IIIA	III	AB029752.2
	4	IIIB	III	FE042314.1
P.platessa	4	IIIB	III	X56744.1
P.promelas	1	I	I	DT305640.1
	2	II	II	DT084603.1
S.salar	I	I	I	X70075.1
	b			X70071.1
	II			X70073.1
	III	II	II	X70074.1
	4	IIIB	III	DW563871.1
S.senegalensis	1a	I	I	AB359189
	1b			AB359190
	1c			AB359191
	2	II	II	AB359192
	3	IIIA	III	AB359193
	Y	IIIB	III	AB359194
T.nigroviridis	1	I	I	CR646532
	2a	II	II	CA845335.1
	2b			CR647041
T.rubripes	3	IIIA	III	AY661445
	4	IIIB	III	AY661446
X.tropicalis	10b	I	II	AAH87759
	10c	I	II	AAH87759

Roach, J. C., K. Wang, L. Gan, and L. Hood. 1997. The molecular evolution of the vertebrate trypsinogens. J Mol Evol 45: 640-52.

Roach, J.C. 2002. A clade of trypsins found in cold-adapted fish. Proteins 47: 31-44

Table 6: Predicted Isoelectric points of select trypsins

Gene	Clade	pI
Omyk 1	I	5.4
Dmaw 1a		5.2
Dmaw 1b		5.2
Dmaw 1c		5.2
Dmaw 1d		5.2
Dmaw 1e		4.9
Dmaw 1f		5.4
Ssen 1a		5.8
Ssen 1b		5.4
Ssen 1c		5.4
Ssal I		5.9 [1]
Ssal 1b		5.9
Ssal II		5.5 [1]
Tnig 1		5.2 [1]
Drer 1		4.9
Drer 2		4.9
Ipun 1A		5.6
Omyk 2	II	8.0
Ssen 2		8.6
Ssal III		8.1 [1]
Tnig 2a		7.0
Ipun 3		6.4
Drer 3		7.0
Xtro 10b	II	7.0
Xtro 10c		4.3
Ggal PRSS2		4.6
Ggal PRSS3		8.4
Hsap T1		7.5 [1]
Hsap T2		5.0 [1]
Hsap T3		6.8 [1]
Omyk 3	IIIA	4.9
Dmaw 3a		4.7
Dmaw 3b		4.7
Dmaw 3c		4.7
Ssen 3		4.7
Trub 3		4.7

Table 6 continued

Ipun 3a		4.9
Omyk 4	IIIB	4.7
Dmaw TLP1		4.4[2]
Dmaw TLP2		4.6
Ssen Y		4.5
Trub 4		4.5
Ipun 4		4.9

[1] **Roach, J. C., K. Wang, L. Gan, and L. Hood. 1997.** The molecular evolution of the vertebrate trypsinogens. J Mol Evol 45: 640-52.
[2] **Spilliaert, R., and A. Gudmundsdottir. 1999.** Atlantic Cod Trypsin Y-Member of a Novel Trypsin Group. Mar Biotechnol (NY) 1: 598-607.