ISSN 0888-7543
Volume 98, Issue 3, September 2011
ELSEVIER
GENOMICS
Wavelet analysis of human DNA
Gene conversions in growth hormone genes
SINE, LINE, and epigenome in methylation susceptibility
C–T variant in a miRNA target site of BCL2
Available online at www.sciencedirect.com
ScienceDirect

Genomics 98 (2011) 194–201

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Assembly of the antifreeze glycoprotein/trypsinogen-like protease genomic locus in the Antarctic toothfish *Dissostichus mawsoni* (Norman)☆

Jessie Nicodemus-Johnson [a,1], Stephen Silic [b], Laura Ghigliotti [c], Eva Pisano [c], C.-H. Christina Cheng [b,*]

[a] Department of Molecular and Integrative Physiology, University of Illinois, Urbana-Champaign, IL 61801, USA
[b] Department of Animal Biology, University of Illinois, Urbana-Champaign, IL 61801, USA
[c] Department of Biology, University of Genoa, Genoa, Italy

ARTICLE INFO

Article history:
Received 6 May 2011
Accepted 2 June 2011
Available online 12 June 2011

Keywords:
Antarctic notothenioids
Dissostichus mawsoni
BAC library
AFGP
Chimeric gene
Repetitive sequence assembly
Gene family evolution

ABSTRACT

To investigate the genomic architecture underlying the quintessential adaptive phenotype, antifreeze glycoprotein (AFGP) that enables Antarctic notothenioid survival in the frigid Southern Ocean, we isolated the *AFGP* genomic locus from a bacterial artificial chromosome library for *Dissostichus mawsoni*. Through extensive shotgun sequencing of pertinent clones and sequence assembly verifications, we reconstructed the highly repetitive *AFGP* genomic locus. The locus comprises two haplotypes of different lengths (363.6 kbp and 467.4 kbp) containing tandem *AFGP*, two *TLP* (trypsinogen-like protease), and surprisingly three chimeric *AFGP*/*TLP*, one of which was previously hypothesized to be a *TLP*-to-*AFGP* evolutionary intermediate. The ~100 kbp haplotype length variation results from different *AFGP* copy number, suggesting substantial dynamism existed in the evolutionary history of the AFGP gene family. This study provided the data for fine resolution sequence analyses that would yield insight into the molecular mechanisms of notothenioid AFGP gene family evolution driven by Southern Ocean glaciation.

1. Introduction

Antarctic notothenioids comprise ~100 related species of cold-adapted teleost fishes isolated within the frigid Southern Ocean by the Antarctic Circumpolar Current (ACC) [1]. The onset of the ACC at ~34–30 mya [2], and the subsequent cooling of Antarctic waters reaching freezing temperatures at ~14–12 mya [3] was accompanied by mass extinction of temperate-water species and the adaptive diversification of the Antarctic notothenioids [1,4]. Remarkable genomic and transcriptomic changes [5] and evolutionary adaptations [6–8] underlie the ecological success of Antarctic notothenioids in their freezing habitats. The most notable adaptive phenotype is antifreeze glycoprotein (AFGP) [9,10] that protects the fish from freezing death. High circulating concentrations (10–35 mg/mL) of AFGPs [10,11] non-colligatively lower the freezing point of the hyposmotic blood and body fluids of notothenioids below that of seawater by binding to ice crystals that enter the fish and inhibiting ice expansion, thereby protecting the fish from freezing death [10].

AFGPs are synthesized as long polyprotein precursors comprising as many as 46 AFGP molecules, each consisting of varying numbers of repeats of the tripeptide Ala(Pro)-Ala-Thr, linked in a series by conserved three-residue sequences (mostly Leu-Asn-Phe) that are post-translationally cleaved [12,13]. The encoding genes consist of two exons; exon 1 encodes a signal peptide and exon 2 encodes the long polyprotein precursor. The primordial AFGP gene evolved from a trypsinogen-like protease (TLP) gene, presumably through an ancestral chimeric *AFGP/TLP* intermediate [13,14]. Extant Antarctic notothenioids have large AFGP gene families [15] indicating extensive gene duplications occurred in response to strong selection from Antarctic sea-level glaciation. Although the molecular origin and mechanism of AFGP evolution have been deduced [13,14], other aspects of the evolutionary process remain unclear. These include the unknown origin of the three-residue linker sequence in the AFGP polyprotein, why a putative evolutionary intermediate—the chimeric AFGP/TLP gene persists in extant notothenioid genomes, and the molecular mechanism of AFGP gene family expansion. Also, the range of AFGP size heterogeneity [10,16] could not be fully accounted for by the few AFGP genes previously sequenced [12,13]. To address these questions requires characterizations of the entire *AFGP/TLP* genomic locus to gain a comprehensive view of the genomic architecture associated with the evolution and maintenance of the adaptive AFGP genotype and phenotype in Antarctic notothenioids.

A Bacterial Artificial Chromosome (BAC) library is useful for isolation of targeted genomic regions for sequencing and characterization [17]. We have constructed a BAC library for the giant Antarctic notothenioid

☆ Sequence data have been deposited with GenBank under accession numbers HQ447059 and HQ447060.

* Corresponding author at: University of Illinois, Urbana-Champaign, Department of Animal Biology, 515 Morrill Hall, 505 S. Goodwin Ave., Urbana, IL 61801, USA. Fax: +1 217 244 4565.

E-mail address: c-cheng@illinois.edu (C.-H.C. Cheng).

[1] Present address: Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.06.002

fish *Dissostichus mawsoni*, estimated at about 7× genome coverage. In this study, we isolated the BAC clones derived from the AFGP/TLP genomic locus, sequenced the shotgun libraries of the minimal set of overlapping clones spanning the locus and reconstructed the locus sequence and structure. AFGP polyprotein coding sequences are highly repetitive (9-nt repeats) resembling lengthy strings of simple sequence repeats [12,13] not amenable to assembly from short sequences from next generation sequencing methods [18–20] despite their large throughput. We used Sanger sequencing [21] for its much longer read lengths, which in conjunction with paired-end sequence matching of shotgun subclones to sequence assemblies has been shown to be more effective in assembling repetitive DNA elements [22]. Here we detail the reconstruction of the physical map and sequence of the highly repetitive *AFGP/TLP* genomic locus of the Antarctic notothenioid *D. mawsoni* (commonly known as Antarctic toothfish).

2. Results and discussion

2.1. FPC minimal tiling path and chromosomal localization of AFGP/TLP locus

We isolated 86 putative *AFGP* and/or *TLP* positive BAC clones from the *D. mawsoni* BAC library and verified that 70 are true positive by Southern blot of *Not*I digests of these clones (SI Fig. 1). Fifty six clones hybridized to both AFGP and TLP gene probes, 13 to the AFGP probe only, and one to the TLP probe only (data not shown). To assess the spatial relationship of these clones, we analyzed their *Hind*III digest fingerprints using FingerPrinted Contigs (FPC) analysis [23,24]. FPC assembled the 70 BAC clones into 2 contig groups (Fig. 1), and the remaining six clones were ungrouped singletons. Contig group 1 encompasses all 56 dual *AFGP* and *TLP* positive clones, with a minimal tiling path (MTP) (i.e. the smallest number of overlapping clones spanning the contig group) of four clones—DmBAC42, 10, 64 and 39 (Fig. 1A). Contig group 2 encompasses eight of the 13 *AFGP*-positive only clones, and is represented by a single MTP clone, DmBAC78 (Fig. 1B). The six ungrouped singletons consisted of five *AFGP*-positive only (DmBAC74, 75, 79, 80 and 85) and one *TLP*-positive only (DmBAC35) clones.

FPC prediction of clone overlap relies on sufficient numbers of shared restriction fragments in the overlapping region, and may fail to predict overlap for small overlaps. To determine if the FPC groupings are spatially distinct in the genome, we performed fluorescence *in situ* hybridization (FISH) of *D. mawsoni* metaphase chromosomes, using DmBAC64 of contig 1 (dual *AFGP/TLP* positive), DmBAC78 of contig 2 (*AFGP*-positive only), and the singleton DmBAC35 (*TLP*-positive only) as probes. All three BAC clone probes hybridized to the same region in a single pair of chromosomes (Fig. 2), indicating there is a single *AFGP/TLP* genomic region in *D. mawsoni*. The localization of the *AFGP*-containing BAC probes (DmBAC78 and DmBAC64) to a single chromosomal site also indicates that the five *AFGP*-containing singletons must also belong to the same genomic location. These five BAC clones have relatively small inserts (one at ~23 kbp and four at ~55 kbp) resulting in fewer *Hind*III fragments (4 to 12, versus >20 for contig group clones) (data not shown), which likely produced an inadequate number of shared bands for FPC to predict their overlap with the other *AFGP*-positive clones. The *TLP*-positive only singleton DmBAC35 was found in subsequent BAC clone sequence alignment (see later sections) to overlap with MTP clone DmBAC42, consistent with the chromosomal FISH results of a single *AFGP/TLP* genomic locus. We will refer to the six BAC clones (DmBAC42, 10, 64, 39, 78 and 35) as the minimal tiling path (MTP) clones of the locus from hereon.

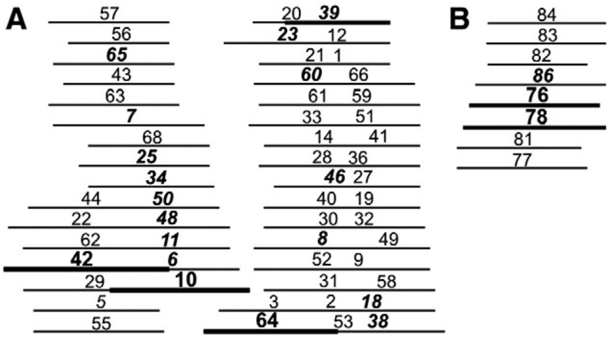

Fig. 1. FPC contig assemblies of *Hind*III fingerprints of 70 *AFGP/TLP*-positive clones from *D. mawsoni* BAC library. (A) Contig group 1 encompasses 56 BAC clones that hybridized to both *AFGP* and *TLP* probes. (B) Contig group 2 encompasses eight of 13 BAC clones that hybridized to *AFGP* probe only. The remaining six BAC clones were ungrouped singletons. Bolded lines and clone numbers indicate minimal tiling path (MTP) clone/s for each contig group. Bolded italic clone numbers are BAC clones that were paired-end sequenced to corroborate BAC clone order and presence of two distinct *AFGP* haplotypes (see text).

2.2. Sequencing and assembly strategy of AFGP/TLP MTP BAC clones

The insert sizes of the MTP BAC clones ranged from ~124 kbp to ~161 kbp based on pulsed field electrophoresis of *Not*I digests of the plasmid DNA (SI Fig. 2). We constructed a shotgun (1–5 kbp inserts)

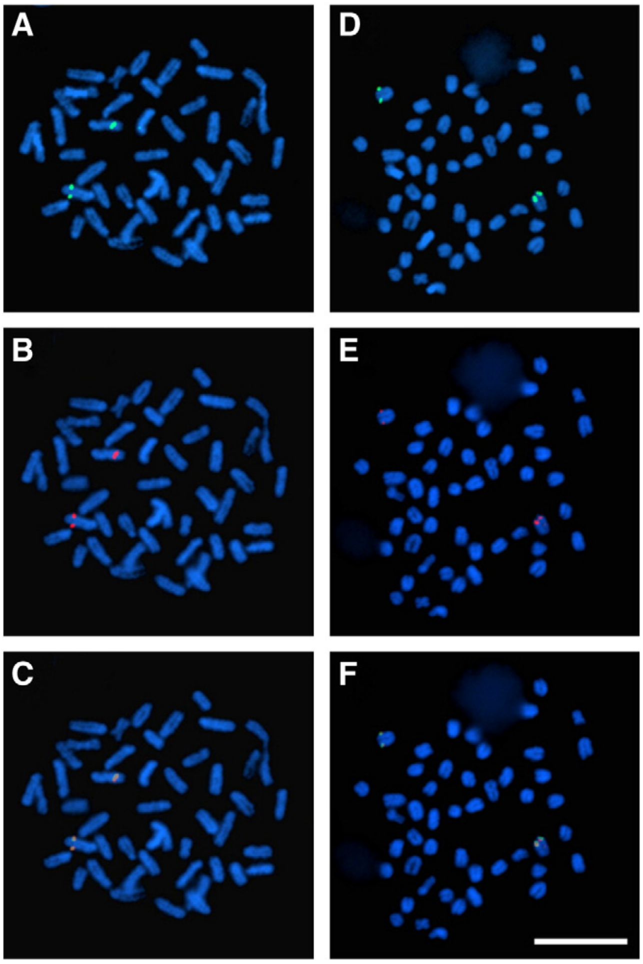

Fig. 2. FISH (fluorescence in situ hybridization) on *D. mawsoni* chromosomes using MTP BAC clones as probes. (A–C) Same chromosome spread sequentially hybridized with probes from (A) DmBAC64 (FPC contig group 1) and (B) DmBAC78 (FPC contig group 2); (C) superposition of images A and B. (D–F) Same chromosome spread sequentially hybridized with probes from (D) DmBAC78 (FPC contig group 2) and (E) DmBAC35 (singleton TLP-positive only); (F) superposition of images D and E. Within the distance resolution of chromosome FISH, AFGP and TLP genes are contained with a single chromosomal region. Scale bar = 10 μm.

library for each MTP clone, and sequenced them to 8–12× BAC insert coverage. Initial database BLASTN and TBLASTX searches using shotgun sequences as queries revealed that the *D. mawsoni AFGP/TLP* locus contains multiple copies of AFGP, TLP and chimeric AFGP/TLP genes, as well as two other types of trypsinogen genes. The tandemly repeated genes and the plethora of highly repetitive AFGP coding sequences (cds) greatly complicated sequence alignment. We thus sequenced the paired ends of the shotgun libraries (~3.5× BAC insert coverage) to aid in shotgun sequence alignments. We also constructed a larger-insert (5–30 kbp) shotgun library for the AFGP gene-rich MTP clones DmBAC42 and 78, and paired-end sequenced them to provide longer-distance anchors in ordering 1–5 kbp shotgun sequence contigs. A fully contiguous sequence for each BAC clone insert was not obtained due to intermittent presence of secondary structures that could not be sequenced through. Thus each assembled BAC clone insert is composed of an ordered series of sequence contigs ranging from 1.4 to 83 kbp separated by gaps of mostly ~100 bp to ~2 kbp (Fig. 3).

2.3. Physical maps of MTP clones DmBAC10, 64, 39 and 35

DmBAC10 and DmBAC64 sequence alignments revealed a large overlap (92.1 kbp) between the two clones, thus we present a collinear physical map of the two clones (Fig. 3A). DmBAC10 assembly produced 10 contigs and 9 gaps, spanning ~160 kbp. It contains two chimeric AFGP/TLP genes (*10C1*, *10C2*), one intact and seven trypsinogen3 pseudogenes, five AFGP genes (*10A$_8$-1*, *10A$_9$*, *10A$_{11}$*, *10A$_7$*, *10A$_8$-2*), and two AFGP pseudogenes (*10A$_6$-1* and *10A$_6$-2*) (the subscripted number in AFGP gene names refers to the number of tripeptide repeats in the last AFGP molecule of the encoded polyprotein precursor, which is one of the most distinctive feature among otherwise very similar AFGP genes) (Fig. 3A). The two AFGP pseudogenes are 5′ truncated, thus missing exon1 (signal peptide) and the majority of intron1. The trypsinogen3 gene annotation was based on high sequence similarity to trypsinogen3 of the Japanese flounder *Paralichthys olivaceus* (AB029752.2) in the database. The seven trypsinogen3 pseudogenes correspond to exon4, intron4, and exon5 of the intact 5-exon trypsinogen3 gene in DmBAC10 (solid blue arrow, Fig. 3A).

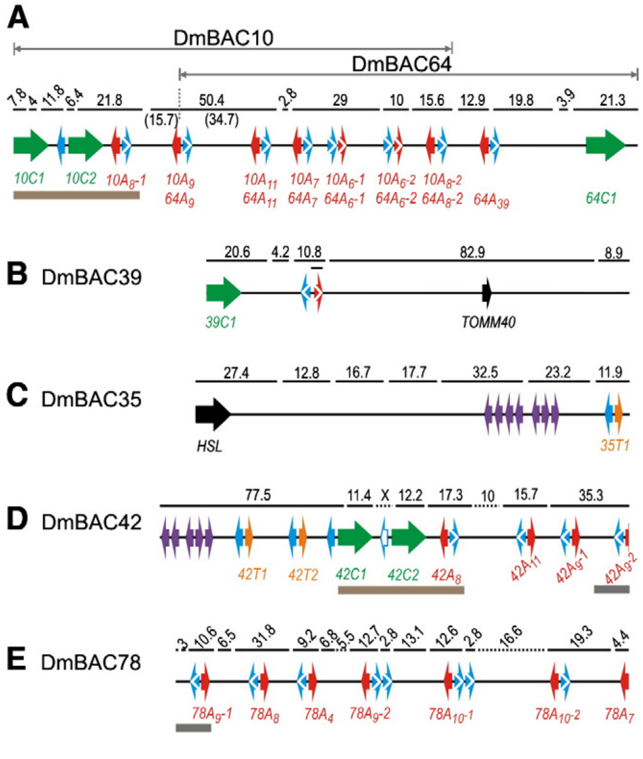

Fig. 3. (A–E) Physical maps of the six MTP BAC clones of *D. mawsoni AFGP/TLP* genomic locus, with BAC clone name as indicated. The series of black line segments with length in kbp indicated represents the ordered series of assembled shotgun sequence contigs of the clone. Space between two contigs represents a gap of ~100 bp–2 kbp unless stated otherwise. Dashed lines are major gaps of unknown (X) or indicated size in kbp. Genes and their positions in a clone are shown as the series of colored arrows below the sequence contig lines. Arrows point in the sense direction of the gene. Solid color arrows are intact genes, fragmented arrows are pseudogenes, and unfilled arrows are genes inferred from occurrence in overlapping clones. Gene color codes: green—chimeric *AFGP/TLP*, red—*AFGP*, blue—*trypsinogen3*, purple—*trypsinogen1*, orange—*TLP*, black—unrelated protein genes, and corresponding gene names are in the same color. In *AFGP/TLP* locus gene names, *C* is abbreviation for chimeric *AFGP/TLP*, *T* for *TLP*, and *A* for *AFGP*. The number before the letter is the clone number. The subscripted number in *AFGP* gene names refer to the number of tripeptide repeats in the last AFGP molecule of the polyprotein encoded by that gene. The pair of color bars (brown in A and D; grey in D and E) indicated the regions of similarity or overlap between the two clones (see text).

The DmBAC64 shotgun sequence assembly produced 9 contigs and 8 gaps, spanning 150 kbp (Fig. 3A). The overlapping region between DmBAC10 and DmBAC64 contain identical genes, which are four *AFGP* (*64A$_9$*, *64$_{11}$*, *64A$_7$*, and *64A$_8$*), two AFGP pseudogenes (*64A$_6$-1* and *64A$_6$-2*), and six trypsinogen3 pseudogenes. A fifth *AFGP*, *64A$_{39}$*, an additional trypsinogen3 pseudogene, and a chimeric *AFGP/TLP*, *64C1* occur in DmBAC64 beyond the overlap region (Fig. 3A). The upstream end of DmBAC64 ends in intron 1 of *64A$_9$*, thus *64A$_9$* is a partial gene in this BAC clone; however it is clearly an intact gene because its counterpart *10A$_9$* in DmBAC10 is a complete gene. From here on, a partial gene refers to an intact gene with part of its sequence occurring in another BAC clone due to the cloning process, distinct from a truncated, pseudogene.

DmBAC39 shotgun sequence assembly produced 6 contigs and 5 gaps, spanning 127 kbp (Fig. 3B). One ~2.1 kbp contig containing a long run of AFGP tripeptide cds without 3-residue linkers could not be mapped with certainty but likely occurs near the 10.8 kbp contig containing a similar AFGP tripeptide cds (depicted as a short line below the 10.8 kbp contig line in Fig. 3B). DmBAC 39 contains a partial chimeric AFGP/TLP gene *39C1* at one end, one trypsinogen3 pseudogene truncated at exon1 and exon5, two AFGP coding fragments (one each in the 2.1 kbp and 10.8 kbp contig), and a gene similar in sequence to *TOMM40* (translocase of outer mitochondrial membrane 40) of *Danio rerio* (BC053295) (Fig. 3B).

The DmBAC35 shotgun assembly has 7 contigs and 6 gaps, spanning 142 kbp (Fig. 3C). The upstream end contains a partial gene highly similar in sequence to hormone sensitive lipase (*HSL*), and the downstream end contains seven tandem 6-exon trypsinogen genes similar in sequence to trypsinogen1 of teleost fishes in the database, one 5-exon trypsinogen3, and one TLP gene *35T1* (Fig. 3C). The presence of signature genes of the locus (*AFGP*, *TLP*, and/or chimeric *AFGP/TLP*) in DmBAC35 and DmBAC39, flanked upstream and downstream respectively by apparently unrelated protein genes, indicates these two clones most likely represent opposite ends of the *AFGP/TLP* genomic locus.

The total sequence contig length from shotgun sequence assemblies of these four MTP BAC clones (Fig. 3) closely approximate their insert size estimated from *Not*I digest of the BAC plasmid DNA (SI Fig. 1; also in Fig. 4), indicating our overall sequence assemblies were reliable.

2.4. AFGP gene-rich MTP clones DmBAC42 and DmBAC78

The remaining two MTP clones, DmBAC42 and DmBAC78 yielded an abundance of shotgun sequences containing repetitive AFGP cds and highly similar intergenic sequences that were difficult to reliably align, resulting in many small contigs. To order these small contigs, we utilized extensive paired-end sequence matching of the 1–5 kbp subclones to the sequence assemblies, and of the larger insert (5–30 kbp) shotgun subclones we constructed for these two BAC clones to provide longer distance anchors. The assembled DmBAC42 consisted

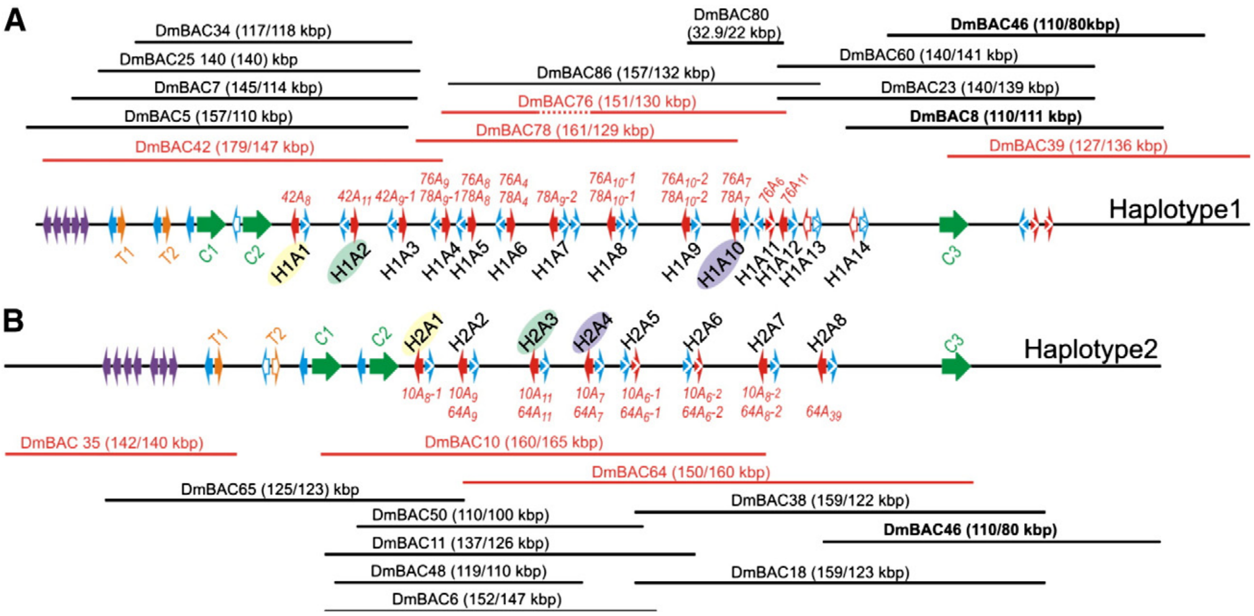

Fig. 4. Reconstruction of the two haplotypes of the AFGP/TLP genomic locus in *D. mawsoni*. (A) Haplotype 1 (467.4 kbp). (B) Haplotype 2 (363.6 kbp). The codes for genes (color, arrow format, sense direction, gene names) are the same as in Fig. 3. The two haplotypes are conserved in gene complement in the upstream and downstream portions, but is polymorphic in the central *AFGP*-populated region in copy number and sequences resulting in the length difference. Haplotype AFGP gene names are in numerical order, *H1A1* and *H2A1*, and so on. The names of MTP BAC clones belonging to each haplotype are in red, and the names of sister clones corroborating the segregation of the MTP clones in two haplotypes are in black. Clones with bolded labels could map to either haplotype. The lines below BAC clone names represent the span of the clones and their positions in the haplotype schematics. The two bracketed numbers are clone lengths based on the distance between the positions of the paired-BAC end sequences in the haplotype consensus sequence and the *Not*I estimated insert size, in that order. The lengths of clones and haplotypes are not to scale, but clone overlap and positions of clones with respect to the haplotype and gene content are exact. The three pairs of AFGP genes identified by Neighbor-Joining phylogenetic analysis to be allelic (see Fig. 5 and text) are indicated by same color oval highlight.

of six contigs and 5 gaps (Fig. 3D), and DmBAC78 consisted of 14 contigs and 13 gaps (Fig. 3E). Gaps are mostly 100 bp–2 kbp, but two additional sizable gaps (10 kbp, and one of undetermined size) in DmBAC42, and two (5.5 kbp and 16.6 kbp) in DmBAC78 (Figs. 3D and E) persisted. DmBAC42 contained five trypsinogen1 genes, four trypsinogen3 genes, two *TLP* (*42T1* and *42T2*), four trypsinogen3 pseudogenes, two chimeric AFGP/TLP genes (*42C1* and *42C2*), and four AFGP genes (*$42A_8$*, *$42A_{11}$*, *$42A_9$-1*, and *$42A_9$-2* (partial gene)) (Fig. 3D). DmBAC78 contained seven AFGP genes (*$78A_9$-1*, *$78A_8$*, *$78A_4$*, *$78A_9$-2*, *$78A_{10}$-1*, *$78A_{10}$-2*, and *$78A_7$* (partial gene)), and eight trypsinogen3 pseudogenes (Fig. 3E).

The reconstructed sequence lengths of DmBAC42 (179.4 kbp) and DmBAC78 (161.4 kbp) exceeded their BAC insert length (~147 kbp and ~129 kbp respectively) estimated from *Not*I digest by ~32 kbp. We have performed multiple assemblies for both clones, and similar consensus sequences emerged each time. The ~32 kbp discrepancy for both BAC clones indicates that some sequence misalignments remained, likely due to alignment uncertainties stemming from the abundance of highly repetitive AFGP cds in these two AFGP-gene rich clones, which could not be fully resolved by the current alignment methods.

2.5. Reconstruction of MTP BAC clone order by shared sequence identity

We next aligned the assembled MTP clone consensus sequences to establish clone overlap and clone order in the *AFGP/TLP* locus. Overlapping regions were assessed on the criterion that they share >90% nt (nucleotide) identity. By this criterion, DmBAC39, DmBAC64, and DmBAC10 overlap each other (SI Fig. 3), in the order predicted by FPC (Fig. 1A). MTP clone DmBAC35, a singleton in FPC analysis, overlapped with DmBAC42 by 40.5 kbp, corroborating the chromosome FISH mapping of DmBAC35 to the single *AFGP/TLP* genomic locus (Fig. 2). The DmBAC35/DmBAC42 overlap region spans seven genes—one TLP (*35T1/42T1*), one trypsinogen3, through five of the seven trypsinogen1 genes (Figs. 3C and D).

The mid-portion of DmBAC42, encompassing *42C1*, *42C2*, *$42A_8$* contains the same gene content and similar sequences to the 5′ end of DmBAC10, encompassing *10C1*, *10C2*, *$10A_8$-1* (indicated by brown bar in Figs. 3D and A respectively), which is likely the cause of FPC prediction of overlap between these two clones (Fig. 1A). However, beyond this similar segment, the 3′ remainder of DmBAC42 (containing *$42A_{11}$*, *$42A_9$-1* and *$42A_9$-2*; Fig. 3D) diverged in sequence from the corresponding *AFGP*-populated segment in DmBAC10 (and the DmBAC64 segment that overlaps with DmBAC10) (Fig. 3A), thus DmBAC42 and DmBAC10/DmBAC64 could not be overlapping clones. Instead, we found the 3′ end of DmBAC42 shares 12.1 kbp of a near-identical sequence with the 5′ end of DmBAC78 inclusive of one common AFGP gene with 100% nt identity (indicated by grey bar in Figs. 3D and E respectively). FPC had placed DmBAC42 and DmBAC78 in two separate contig groups (Fig. 1), very likely a result of their relatively small overlap, which contains only three *Hind*III sites and thus only two shared bands (5.8 and 3.9 kbp), insufficient for FPC assessment of clone overlap.

The chimeric *AFGP/TLP* of DmBAC42 and DmBAC10 (*42C1* and *10C1*, *42C2* and *10C2*), and *AFGP* (*$42A_8$* and *$10A_8$-1*) in their similar segments (brown bar region, Figs. 3D and A) are not identical in sequence, thus these gene pairs are most likely alleles on the two homologous chromosomes [25]. The large sequence variations in the downstream *AFGP*-populated regions of DmBAC42 and DmBAC10 (Figs. 3D and A), which continue further downstream between DmBAC78 and DmBAC10/DmBAC64 (Figs. 3E and A) led us to hypothesize that there are two distinct haplotypes in the *AFGP*-populated region of the *AFGP/TLP* locus, differing in AFGP gene sequences and copy number, with DmBAC42/DmBAC78 representing one haplotype and DmBAC10/DmBAC64 the other (Fig. 4; details of haplotypes in following sections). FPC had apparently assembled clones from both haplotypes in contig group 1 (Fig. 1A), and clones containing *AFGP* variations specific to one of the haplotypes in contig group 2 (Fig. 1B).

2.6. Additional MTP BAC clone DmBAC76

To extend beyond DmBAC78 (Fig. 3E, SI Fig. 3) and identify clones within the putative variable *AFGP* haplotype region, we sequenced the BAC ends of sister clones of DmBAC78 in FPC contig 2, and aligned them within the DmBAC78 shotgun consensus. One end of DmBAC76 matched within DmBAC78, while the other end did not, which presumably would contain additional downstream sequence. We constructed and paired-end sequenced a 1–5 kbp shotgun subclone library for DmBAC76. Due to the large FPC-predicted overlap with DmBAC78 (Fig. 1B) and difficulties in aligning repetitive *AFGP* sequences, DmBAC76 was partially assembled to determine gene content only. DmBAC76 contains seven AFGP genes ($76A_9$ (partial gene), $76A_8$, $76A_4$, $76A_{10}$-*1*, $76A_{10}$-*2*, $76A_7$, $76A_{11}$ (partial gene)), one AFGP pseudogene ($76A_6$), and seven trypsinogen3 pseudogenes (Fig. 4A). DmBAC76 is missing the counterpart of DmBAC78 AFGP gene $78A_9$-*2* (Fig. 4A), most likely a result of sequence misalignment due to difficulties in aligning the highly repetitive *AFGP* cds that exclusively populated these two clones. Regardless, the other seven AFGP genes and their flanking sequences common to the two clones share 100% nt identity, providing confidence that they originate from the same region in one chromosome.

2.7. BAC clone paired-end sequence matching corroborates the presence of two haplotypes with variable AFGP gene copy number

To verify our hypothesis that the locus is comprised of two distinct haplotypes, we first grouped the MTP clones based on sequence identity between shared AFGP, TLP, and/or chimeric AFGP/TLP genes. Clones belonging to the same haplotype should share 100% gene sequence identity, while clones from a separate haplotype would exhibit allelic variations. Genes in the overlapping region of DmBAC 10 and 64, DmBAC 42 and 78, and DmBAC 78 and 76 show 100% nt identity, indicating the two clones in each pair originated from the same region in one chromosome. In contrast, genes in overlapping regions of DmBAC 42 and 35, and DmBAC 64 and 39 contain SNPs (single nucleotide polymorphism), and thus the two clones in each pair belong to allelic regions in separate chromosomes. Based on these assignments, haplotype 1 contains MTP DmBAC 42, 78, 76 and 39 (Fig. 4A) and haplotype 2 contains MTP DmBAC 35, 10 and 64 (Fig. 4B) in that order. These two sets of haplotype clones, while collectively spanning the entire *AFGP/TLP* genomic region, do not cover the full length of their respective haplotype. We thus obtained paired-end sequences of a number of other BAC clones from both FPC contig groups (Fig. 1A; SI Fig. 4) and aligned them within each haplotype consensus sequence assembled thus far. We assigned clones to a given haplotype on the criteria of 100% sequence match of its BAC end sequences within the consensus sequence of that haplotype, correct orientation of their 5′ and 3′ directions, and the physical distance of the paired-ends on the haplotype consensus sequence approximating the BAC insert length estimated from *Not*I digest. If the sequence match is less than 100% identical, the clone is allelic and belongs to the opposite haplotype. By these criteria, we identified bridging and redundant BAC clones for each set of haplotype MTP clones consistent with a structure of two distinct haplotypes (Fig. 4), verified the presence of similar sequence flanking the apparent variable *AFGP*-populated region on both haplotypes, and determined the length of each haplotype, described as follows.

In haplotype 1, the 5′ end MTP clone DmBAC42 ends right ahead of five trypsinogen1 genes (Figs. 4A and 3D). The corroborating clone DmBAC5 spans further, with its 5′ end matching its allelic site ahead of the seven trypsinogen1 genes in haplotype 2 clone DmBAC35. This establishes that allelic sequences are present in the two haplotypes at the 5′ end of the locus. The gap between MTP clone DmBAC76 in the central *AFGP*-populated region and the 3′ end MTP clone DmBAC39 was bridged by DmBAC23 and DmBAC60 (Fig. 4A). In haplotype 2, DmBAC65 linked MTP clones DmBAC35 and DmBAC10 at the 5′ portion (Fig. 4B). At the 3′ portion, the corroborating clones DmBAC 18 and 38 span further downstream of MTP clone DmBAC64 of the central *AFGP*-populated region, with their 3′ ends matching their allelic sites within DmBAC39 in haplotype 1. This establishes that allelic sequences are present in the two haplotypes at the 3′ end portion of the locus. The haplotype locations of these corroborating clones agree with their relative positions assigned by FPC in contig group 1 (Fig. 1A). Thus, we have accounted for the full length of each haplotype and the presence of the full complement of the gene families (AFGP, TLP, chimeric AFG/TLP and trypsinogens) in both haplotypes.

The physical distance between the paired-end sequences matched within the haplotype consensus sequence agreed well with BAC insert length estimated from *Not*I digest for many of the corroborating clones (Fig. 4; SI Fig. 4). We note however, length difference between the two estimates due to the uncertainties in the sequence assemblies particularly in *AFGP*-populated region covered by DmBAC42 and DmBAC78 in haplotype 1 would be carried over in the corroborating clones that overlap with them. The limitation of the accuracy of *Not*I insert length estimates also likely contributed to the size disagreements for those clones.

Overall, our locus sequence alignments coupled with BAC clone paired-end sequence matching to the alignments support the presence of two distinct *AFGP/TLP* locus haplotypes containing variable *AFGP* copy numbers, which we estimated to be approximately 467.4 kbp (haplotype 1) and 363.6 kbp (haplotype 2).

2.8. Phylogenetic support for two AFGP/TLP locus haplotypes

We further evaluated the presence of two *AFGP/TLP* locus haplotypes by analyzing the evolutionary relationship of 22 AFGP genes from all the MTP BAC clones using Neighbor-Joining phylogenetic analysis. Our expectation was that redundant, identical genes would occupy the same branch in the phylogenetic tree, while allelic genes from different haplotypes would cluster closely but not as closely as identical genes. We used *AFGP* exon 1 (signal peptide cds) and intron1 sequences only, excluding exon 2 (AFGP polyprotein cds) due to its variable lengths and uncertainty in aligning homologous sites in repetitive sequences. Five partial AFGP genes ($42A_9$-*2*, $64A_9$, $76A_9$, $76A_{11}$ and $78A_7$) occurring at BAC clone ends lacked either exon 1 or part of intron 1sequence, and $76A_4$ with an incomplete intron 1, were excluded. The unrooted Neighbor-Joining tree (Fig. 5) shows that five pairs of AFGP genes from overlapping BAC clones determined to be identical genes based on sequence indeed occupy the same branch. Same branch occupancy of DmBAC76/78 *AFGP*, and of DmBAC10/64 *AFGP* also supports each clone pair to be overlapping clones. AFPG genes of DmBAC76/78 and DmBAC10/64 occupy different branches (Fig. 5), which supports their segregation in two separate haplotypes. Three sets of AFGP genes cluster closely and are apparently allelic genes. (Fig. 5; also highlighted in Fig. 4). The Neighbor-Joining inferred allelic relationship of these gene pairs agrees with our assignment of the respective BAC clones that contain them to opposite haplotypes described in the last section. The remaining AFGP genes lack alleles in the opposite haplotype (Figs. 4 and 5), indicating substantial polymorphism exists in AFGP gene families in *D. mawsoni* populations.

To summarize, we have constructed a BAC library for the Antarctic notothenioid fish *D. mawsoni*, isolated, sequenced and reconstructed the genomic region containing the AFGP gene family crucial for notothenioid survival in the freezing Southern Ocean. *AFGPs* and paralogs of its *TLP* ancestor reside in a single genomic locus in *D. mawsoni*. The locus comprises two distinct haplotypes polymorphic in gene *AFGP* copy number, 14 and 8 for haplotype 1 and 2 respectively, resulting in a ~100 kbp difference in haplotype lengths (467.4 kbp and 363.6 kbp respectively) (Fig. 4). Unexpectedly and intriguingly,

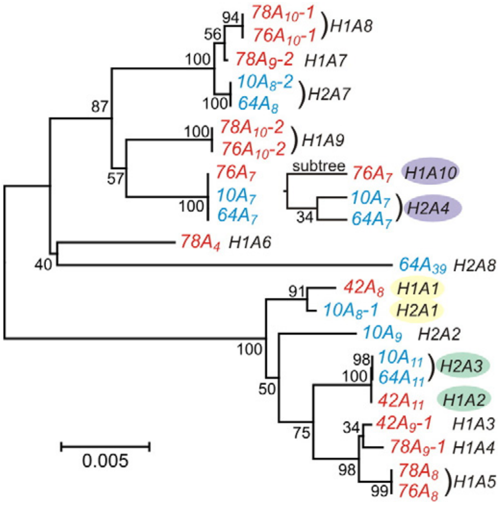

Fig. 5. Unrooted Neighbor-Joining phylogenetic tree of 22 AFGP genes from all MTP BAC clones comprising the *AFGP/TLP* genomic locus. The subtree expands the topology of the relevant clade whose branch lengths were not resolved with the scale of the parent tree. Node supports from 1000 bootstrap replicates are as indicated. AFGP genes belonging to haplotype 1 and 2 are given in red and blue respectively. Identical genes represented twice due to redundant BAC clones from the same location in a haplotype are connected by a black half bracket. Haplotype AFGP gene names (in black) are the same as here and in Fig. 4. Only three pairs of AFGP genes are allelic, which are highlighted here and in Fig. 4.

there are two additional chimeric AFGP/TLP genes besides the one we previously discovered [14]. The evolutionary significance of multiple chimeric genes, and how it may relate to our prior hypothesis of the chimeric gene being an evolutionary intermediate [14] needs further investigation. The presence of two TLP genes and multiple intact and truncated trypsinogen3 genes within the locus, as well as seven trypsinogen1 genes as immediate neighbors corroborate the evolutionary origin of AFGP genes in a trypsinogen locus. The tandem array of *AFGP* and their highly repetitive cds, plus considerable heterozygosity (two distinct haplotypes), epitomize the bioinformatics challenges of assembling and reconstructing repetitive genomic sequence. While four additional AFGP-positive only BAC clones (~55 kbp each) remained unmapped, we believe they would largely serve to increase the final AFGP gene count only, without altering the overall locus organization determined in this study. The sequence data and the locus structure from this study will allow us to examine in detail the molecular and/or recombination mechanism of the AFGP gene family expansion, address the remaining unknown aspects of the *TLP*-to-*AFGP* evolutionary process as well as the genic basis for the AFGP protein heterogeneity, to be reported elsewhere. Lastly, the *D. mawsoni* BAC library we constructed for this study, estimated at 7× genome coverage, adds to the emerging Antarctic notothenioid genomic resources [26] for genome-enabled studies in Antarctic notothenioid fish as a model of cold-extreme adaptation [5,27], and as models for several human genetic diseases proposed recently [28].

3. Materials and methods

3.1. BAC library construction and screening

D. mawsoni specimens were caught with a baited vertical line in McMurdo Sound, Antarctica. Red blood cells from a single specimen were washed with physiological buffer, embedded in 1% low melting agarose and cast in 80 µL block molds (Bio-Rad), lysed *in situ* and stored in a preservation buffer before returning to the USA. Library construction followed published protocols [29,30]. Briefly partially *Eco*RI digested DNA was resolved by pulsed field electrophoresis (CHEF Mapper XA, BioRad), and 75–150 kbp fragments were ligated into pCC1BAC vector (Epicentre) and transformed into *E. coli* DH10B-T1 (Invitrogen). Colonies were robotically picked and cultured in 384-well plates to produce the archive library. A duplicate library was macroarrayed on nylon hybridization filters and evaluated for quality and depth of coverage (details to be reported elsewhere). The library filters were screened sequentially for clones containing AFGP and TLP genes by Southern hybridization using ^{32}P-labeled AFGP cds and TLP cDNA probes respectively. The hybridized filters were scanned using the STORM PhosphoImager (Molecular Dynamics), and hybridized BAC clones were identified to their plate addresses in the archived library.

3.2. FPC analysis and construction of a minimal tiling path

About 2 µg BAC plasmid DNA was digested with *Hind*III and electrophoresed on a 1% agarose gel with Genomic DNA Marker II (Fermentas) in every fifth lane at 4 °C for 18 h. The gel was stained with Sybr green, and the *Hind*III fingerprint was scanned with STORM PhosphoImager (Molecular Dynamics). Band calling of the digital fingerprints was performed using IMAGE3 (www.sanger.ac.uk/resources/software/image/). The band data were then analyzed in FingerPrinted Contigs (FPC) v.7 (www.agcol.arizona.edu/software/fpc/) which clustered clones into contig groups based on their probability of coincidence scores [23,24], and predicted the minimal tiling path (MTP).

3.3. Chromosomal localization of AFGP/TLP locus by FISH

Metaphase chromosomes were prepared from the head kidney cells of *D. mawsoni*, and chromosome fluorescence *in situ* hybridizations (FISH) were performed following published protocol for Antarctic notothenioid fish [31]. BAC plasmid DNA of selected MTP clones was biotin-14-dCTP labeled (BioPrime Labeling System, Invitrogen) and hybridized to denatured chromosomes on slides at 37 °C overnight. The hybridized probe was detected by incubation with streptavidin-Cy3 (Amersham Biosciences) or streptavidin-Alexa 488 (Molecular Probes), and the chromosomes were counter stained with DAPI (Vector). For hybridization with a second BAC DNA probe, the hybridized chromosomes were stripped of the first probe, and rehybridized with the second probe using the same procedures except for shorter chromosome denaturation times. Chromosomal spreads were examined with an Olympus BX61 epifluorescence microscope, and hybridization signals were captured with a Sensys (Photometrics) CCD camera and processed with the software Genus (Applied Imaging).

3.4. Shotgun library construction, plasmid DNA preparation and sequencing

Plasmid DNA of the six MTP BAC clones, and other FPC contig group BAC clones selected for BAC end sequencing, was transformed into TransforMax EPI300 *E. coli* (Epicentre), which was then cultured and induced to replicate the plasmid to high copy number following the manufacturer's instructions. BAC plasmid DNA was isolated by alkaline lysis miniprep. Shotgun subclone (1–5 kbp) sequencing libraries were constructed for the MTP BAC clones by cloning size-selected fragments into the pCR4Blunt-TOPO vector using the TOPO shotgun subcloning kit (Invitrogen). A 5–30 kbp shotgun library was made for two MTP clones using the pJAZZ-OK Blunt vector and BigEasy subcloning system (Lucigen). Shotgun subclone plasmid DNA was prepared in 96-well plates by alkaline lysis miniprep and sequenced in 96-well format. The pCR4Blunt-TOPO subclone plasmids were sequenced with T3 and T7, pJAZZ-OK plasmids with SL1 and NZ1 (Lucigen), and BAC clones with pCC1/pEpiFOS-Forward (Epicentre) or T7 and RP3 (5′-ACACTTTATGCTTCCGGCTCGTATGT-3′) using Big Dye v.3 Terminator Cycle Sequencing kit (Applied Biosystems) and run on

ABI3730*xl* sequence analyzer (Applied Biosystems) at the Keck Center for Comparative and Functional Genomics, University of Illinois.

3.5. Shotgun sequence assembly and gene annotation of individual MTP BAC clones

Shotgun sequence files were edited, aligned, and analyzed in Sequencher 4.5 (Gene Codes). Accuracy of subclone sequence alignments, gap sizes, and sequence contig order were determined by paired-end sequence matching of shotgun subclones (1–5 kbp and/or 5–30 kbp) that spanned the gaps. Accuracy of the locations of paired-end sequences in turn was assessed on the agreement between the physical distance of the ends on the sequence contigs they linked and the insert size estimated from restriction digest, and correct orientation of their 5′ and 3′ directions. The insert size of paired-end sequenced 1–5 kbp subclones was determined by comparing *Eco*RI excised subclone insert to a 1 kbp ladder (Invitrogen) on gel electrophoresis. The 5–30 kbp subclones insert sizes were determined by comparing *Not*I excised insert to low range pulsed field standards (New England Biolabs) on pulsed field gel electrophoresis (CHEF Mapper XA, Bio-Rad). Presence and annotation of genes in shotgun subclone assembly consensus sequences were based on BLASTN and/or BLASTX hits in the database. Protein coding sequence and intron–exon junctions of genes were delineated manually with the aid of Protein Translator in www.justbio.com.

3.6. Assembly of the AFGP/TLP genomic locus consensus sequence from MTP BAC clone shotgun sequence assemblies

The consensus sequences of MTP BAC shotgun clone assemblies were analyzed for nucleotide identity of shared sequences between BAC clones to determine which are overlapping clones. The alignment of two clones identified by shared high sequence identity (>90%), were further assessed by the more stringent criterion that overlapping clone from one chromosome share ≥99% gene sequence identities in overlap regions. This separated the seven MTP BAC clones in two hypothesized haplotypes. The two haplotype structure was verified by matching BAC end sequences of sister *AFGP* and/or *TLP* positive BAC clones to the haplotype consensus sequence assembled from the two sets of haplotype MTP clones. Correct placement of the paired-end sequences of a corroborating clone was by the criteria that BAC end sequences share ≥99% nucleotide identity with their matching sites and their 5′ and 3′ directions are oriented correctly within the assembly, and at a distance approximating the insert size of the clone. To estimate BAC clone insert length, *Not*I digested BAC plasmid DNA and low range pulsed field gel standards were resolved by pulsed field gel electrophoresis. A regression line was generated for the PF standard sizes versus migration distance measured from a digital gel image, and the size/s of *Not*I excised insert band/s were calculated by entering the insert band migration distance in the regression equation.

3.7. Phylogenetic analysis

Phylogenetic relatedness of AFGP genes from all BAC clones were determined using the AFGP signal peptide (exon 1) and intron 1 sequences. The highly repetitive exon 2 sequence (AFGP polyprotein cds) was excluded because they cannot be aligned with high confidence. Sequence alignment and Neighbor-Joining analysis were performed using MEGA v4 [32]. Analysis was run under pairwise deletion and maximum composite likelihood parameters, and node support was evaluated with 1000 bootstrap replicates.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.06.002.

Acknowledgment

We are deeply grateful to Chris Amemiya for making his laboratory facilities at the Benaroya Research Institute, Seattle available as well as providing guidance for the construction of the BAC library. The research was supported by National Science Foundation grants OPP 0231006 and OPP 0636696 to C-HCC.

References

[1] J.T. Eastman, Antarctic Fish Biology: Evolution in a Unique Environment, Academic, San Diego, 1993.

[2] R. Livermore, A. Nankivell, G. Eagles, P. Morris, Paleogene opening of Drake Passage, Earth Planet. Sci. Lett. 236 (2005) 459–470.

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

[4] J.T. Eastman, The nature of diversity of Antarctic fishes, Polar Biol. 28 (2005) 93–107.

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

[6] H.W. Detrich III, S.K. Parker, R.C.J. Williams, E. Nogales, K.H. Downing, Cold adaptation of microtubule assembly and dynamics. Structural interpretation of primary sequence changes present in the alpha- and beta-tubulins of antarctic fishes, J. Biol. Chem. 275 (2000) 37038–37047.

[7] P.A. Fields, G.N. Somero, Hot spots in cold adaptation: localized increases in conformational flexibility in lactate dehydrogenase A4 orthologs of Antarctic notothenioid fishes, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 11476–11481.

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

[9] A.L. DeVries, Glycoproteins as biological antifreeze agents in Antarctic fishes, Science 172 (1971) 1152–1155.

[10] A.L. DeVries, C.-H.C. Cheng, Antifreeze proteins and organismal freezing avoidance in polar fishes, in: A.P. Farrell, J.F. Steffensen (Eds.), The Physiology of Polar Fishes, Elsevier Academic Press, San Diego, 2005, pp. 155–201.

[11] Y. Jin, A.L. DeVries, Antifreeze glycoprotein levels in Antarctic notothenioid fishes inhabiting different thermal environments and the effect of warm acclimation, Comp. Biochem. Physiol. B 144 (2006) 290–300.

[12] K.C. Hsiao, C.-H.C. Cheng, I.E. Fernandes, H.W. Detrich III, A.L. DeVries, 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 (1990) 9265–9269.

[13] L. Chen, A.L. DeVries, C.-H.C. Cheng, Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 3811–3816.

[14] C.-H.C. Cheng, L. Chen, Evolution of an antifreeze glycoprotein, Nature 401 (1999) 443–444.

[15] C.-H.C. Cheng, H.W. Detrich III, Molecular ecophysiology of Antarctic notothenioid fishes, Phil. Trans. R. Soc. Lond. B Biol. Sci. 362 (2007) 2215–2232.

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

[17] C.T. Amemiya, T.P. Powers, S.J. Prohaska, J. Grimwood, J. Schmutz, M. Dickson, T. Miyake, M.A. Schoenborn, R.M. Myers, F.H. Ruddle, P.F. Stadler, HOX clusters of Latimeria: complete characterization provides further evidence for slow evolution of the coelacanth genome, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 3622–3627.

[18] M. Ronaghi, Pyrosequencing sheds light on DNA sequencing, Genome Res. 11 (2001) 3–11.

[19] M.L. Metzker, Emerging technologies in DNA sequencing, Genome Res. 15 (2005) 1767–1776.

[20] S.M. Goldberg, 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, J.C. Venter, 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 (2006) 11240–11245.

[21] J. Zimmermann, H. Voss, C. Schwager, J. Stegemann, W. Ansorge, Automated Sanger dideoxy sequencing reaction protocol, FEBS Lett. 233 (1988) 432–436.

[22] T. Wicker, E. Schlagenhauf, A. Graner, T.J. Close, B. Keller, N. Stein, 454 sequencing put to the test using the complex genome of barley, BMC Genomics 7 (2006) 275.

[23] C. Soderlund, S. Humphray, A. Dunham, L. French, Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 10 (2000) 1772–1787.

[24] C. Soderlund, I. Longden, R. Mott, FPC: a system for building contigs from restriction fingerprinted clones, Comp. Appl. Biosci. 13 (1997) 523–535.

[25] M.J. Smith, Evolutionary genetics, 1998 Oxford.

[26] H.W. Detrich III, A. Stuart, M. Schoenborn, S.K. Parker, B.A. Methe, C.T. Amemiya, Genome enablement of the notothenioidei: genome size estimates from 11 species and BAC libraries from 2 representative taxa, J. Exp. Zool. B Mol. Dev. Evol. 314 (2010) 369–381.

[27] N.R. Council, Frontiers in Polar Biology in the Genomic Era, National Academy of Sciences, Washington DC, 2003.

[28] R.C. Albertson, W. Cresko, H.W. Detrich III, J.H. Postlethwait, Evolutionary mutant models for human disease, Trends Genet. 25 (2009) 74–81.

[29] C.T. Amemiya, T. Ota, G.W. Litman, Construction of P1 artificial chromosome (PAC) libraries from lower vertebrates, in: E. Lai, B. Birren (Eds.), Analysis of Nonmammalian Genomes, Academic Press, San Diego, CA, 1996, pp. 223–256.

[30] R.J. Parchem, F. Poulin, A.B. Stuart, C.T. Amemiya, N.H. Patel, BAC library for the amphipod crustacean, *Parhyale hawaiensis*, Genomics 95 (2010) 261–267.

[31] L. Ghigliotti, F. Mazzei, C. Ozouf-Costaz, C. Bonillo, R. Williams, C.-H.C. Cheng, E. Pisano, The two giant sister species of the Southern Ocean, *Dissostichus eleginoides* and *Dissostichus mawsoni*, differ in karyotype and chromosomal pattern of ribosomal RNA genes, Polar Biol. 30 (2006) 625–634.

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