Marine Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

Method paper

Reconstruction of the repetitive antifreeze glycoprotein genomic loci in the cold-water gadids *Boreogadus saida* and *Microgadus tomcod*

Xuan Zhuang[a,b,*], Katherine R. Murphy[a], Laura Ghigliotti[c], Eva Pisano[c], C.-H. Christina Cheng[a,*]

[a] Department of Animal Biology, University of Illinois at Urbana – Champaign, 515 Morrill Hall, Urbana, IL 61801, USA
[b] Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA
[c] Institute of Marine Sciences (ISMAR), National Research Council (CNR), Genoa 16149, Italy

ARTICLE INFO

Keywords:
AFGP-bearing cods
BAC library construction
Sequencing strategies for repetitive sequences
AFGP sequence assembly
chromosomal FISH

ABSTRACT

Antifreeze glycoproteins (AFGPs) are a novel evolutionary innovation in members of the northern cod fish family (Gadidae), crucial in preventing death from inoculative freezing by environmental ice in their frigid Arctic and sub-Arctic habitats. However, the genomic origin and molecular mechanism of evolution of this novel life-saving adaptive genetic trait remained to be definitively determined. To this end, we constructed large insert genomic DNA BAC (bacterial artificial chromosome) libraries for two AFGP-bearing gadids, the high-Arctic polar cod *Boreogadus saida* and the cold-temperate Atlantic tomcod *Microgadus tomcod,* to isolate and sequence their *AFGP* genomic regions for fine resolution evolutionary analyses. The BAC library construction encountered poor cloning efficiency initially, which we resolved by pretreating the agarose-embedded erythrocyte DNA with a cationic detergent, a method that may be of general use to BAC cloning for teleost species and/or where erythrocytes are the source of input DNA. The polar cod BAC library encompassed 92,160 clones with an average insert size of 94.7 kbp, and the Atlantic tomcod library contained 73,728 clones with an average insert size of 89.6 kbp. The genome sizes of *B. saida* and *M. tomcod* were estimated by cell flow cytometry to be 836 Mbp and 645 Mbp respectively, thus their BAC libraries have approximately 10- and 9.7-fold genome coverage respectively. The inclusiveness and depth of coverage were empirically confirmed by screening the libraries with three housekeeping genes. The BAC clones that mapped to the *AFGP* genomic loci of the two gadids were then isolated by screening the BAC libraries with gadid AFGP gene probes. Eight minimal tiling path (MTP) clones were identified for *B. saida*, sequenced, and assembled. The *B. saida AFGP* locus reconstruction produced both haplotypes, and the locus comprises three distinct AFGP gene clusters, containing a total of 16 AFGP genes and spanning a combined distance of 512 kbp. The *M. tomcod AFGP* locus is much smaller at approximately 80 kbp, and contains only three AFGP genes. Fluorescent *in situ* hybridization with an AFGP gene probe showed the *AFGP* locus in both species occupies a single chromosomal location. The large *AFGP* locus with its high gene dosage in *B. saida* is consistent with its chronically freezing high Arctic habitats, while the small gene family in *M. tomcod* correlates with its milder habitats in lower latitudes. The results from this study provided the data for fine resolution sequence analyses that would yield insight into the molecular mechanisms and history of gadid AFGP gene evolution driven by northern hemisphere glaciation.

1. Introduction

Gadids (family Gadidae, order Gadiformes) are a major group of cold water fishes that primarily inhabit waters of the northern oceans. They are believed to have originated from the North Atlantic Ocean and expanded into Arctic and circum-Arctic regions over evolutionary time (Cohen et al., 1990; Howes, 1991; Coulson et al., 2006). Under strong selection from their freezing ice-laden environments, several Arctic and northern gadid fishes have evolved blood-borne antifreeze glycoproteins (AFGPs). AFGPs bind to environmental ice crystals that enter these fishes and inhibit their growth, thereby preventing inoculative freezing of the hyposmotic body fluids that otherwise would lead to inevitable death (DeVries, 1983; Cheng, 1998). How this novel, life-saving gadid AFGP evolved has remained an intriguing question as there are no homologous sequences in the databases to infer ancestry or mechanism. Our previous work (Zhuang, 2014) suggested that the gadid AFGP gene evolved from non-protein-coding DNA, which is corroborated by a recent report using whole genome shotgun sequencing (Baalsrud et al., 2017). However, the detailed evolutionary process and the underlying molecular mechanism remain undetermined. Additionally, in gadid

* Corresponding authors at: Department of Animal Biology, University of Illinois at Urbana – Champaign, Urbana, IL 61801, USA.
E-mail addresses: xzhuang3@uchicago.edu (X. Zhuang), c-cheng@illinois.edu (C.-H.C. Cheng).

https://doi.org/10.1016/j.margen.2018.02.003
Received 6 February 2018; Received in revised form 21 February 2018; Accepted 21 February 2018

phylogeny (Teletchea et al., 2006) species known to have AFGPs nest with non-AFGP bearing species in two separate subclades, which raises the alternative hypotheses of separate emergences of the AFGP trait versus a single origin with multiple losses in extant non-AFGP bearing lineages. Regardless, the evolution of AFGPs may be potentially associated with the northern hemisphere's Quaternary glacial advances and retreats that produce temporally and spatially separate freezing selection pressures on closely related gadids. Investigating the evolutionary mechanism and history of gadid *AFGP* requires detailed characterization of the AFGP genomic region. Therefore, in this study we constructed large-insert DNA bacterial artificial chromosome (BAC) libraries for two gadid species, and isolated and sequenced their AFGP loci for fine-scale sequence analyses.

Two AFGP-bearing gadids that represent separate clades in the gadid phylogeny (Teletchea et al., 2006) were selected for this study - the polar cod *Boreogadus saida* and the Atlantic tomcod *Microgadus tomcod*. *B. saida* is native to high and circum-Arctic latitudes, living in the extreme cold and icy Arctic seas (Craig et al., 1982; Gradinger and Bluhm, 2004), and expresses high levels of AFGPs for protection against freezing (Denstad et al., 1987; Cheng et al., 2006). *M. tomcod* inhabits lower latitudes along the cool temperate Atlantic coasts of North America, where the risk of freezing is limited to winter and therefore its AFGP levels show seasonal variation (Reisman et al., 1987). The sequencing and reconstruction of their AFGP genomic loci and neighboring regions enable fine-scale sequence analyses that will allow us to decipher the origin and molecular mechanisms of gadid AFGP gene evolution, and whether AFGP emerged separately in these two species.

Gadid AFGPs are composed of Thr-Ala/Pro-Ala repeats in the protein backbone, and are encoded as polyprotein precursors consisting of multiple AFGP molecules linked in tandem. The encoding genes thus consist of long runs (1–3 kbp) of repetitive 9-nucleotide (codons for the tripeptide repeats) resembling simple sequence repeats (Chen et al., 1997). The repetitiveness is further magnified by the duplication of AFGP polyprotein genes in some cold water gadids (e.g. *B. saida*), forming large multigene families under selection from low temperatures (Zhuang et al., 2012). Our previous study indicated that it is extremely difficult to completely assemble such a repetitive genomic locus from short read shotgun sequencing (Zhuang et al., 2012). Thus isolation of the pertinent smaller, targeted genomic regions and sequencing with longer read lengths are needed to reconstruct the repetitive AFGP genomic region.

A BAC library, which represents the entire genome of an organism without artificial amplification or rearrangements, provides large DNA insert (up to 300 kbp) clones physically separated in an addressable format. The BAC to BAC approach can create a crude physical map of the whole genome. As such, BAC libraries continue to be utilized to-date to verify positional accuracies of contigs and scaffolds in whole genome sequence assemblies through using BAC end sequences as mile posts, and whole BAC clones of known gene content in FISH (fluorescent *in situ* hybridization) to metaphase chromosomes. BAC libraries are also ideal for isolating targeted genomic regions of interest for fine mapping and analyses, isolation of unknown genes without reference sequences, and development of molecular markers.

BAC cloning is known to be technically challenging in that any number of factors could lead to limited insert sizes and cloning efficiencies. We indeed encountered a peculiar difficulty in achieving a sufficient yield of recombinant clones in the construction of the BAC libraries for our two cold water gadids. We therefore include in this report our resolution for this difficulty that may benefit BAC library construction efforts of teleost species in general.

2. Materials and methods

2.1. Specimens and flow cytometric determination of genome size

Specimens of *B. saida* were collected from the coastal waters of NE Greenland (75°N, 20°W) by trawling from the R/V Jan Mayen (renamed Helmer Hanssen) during participation in a TUNU research conducted by the UiT Arctic University of Norway during a TUNU cruise to those waters. Specimens of *M. tomcod* were caught with traps from the inshore waters of Shinnecock Bay, Long Island, New York (41°N, 72°W) with kind assistance from Dr. Howard Reisman. Fish were anesthetized and heparinized blood was obtained with needle and syringe from the caudal blood vessel. All fish handling followed University of Illinois IACUC approved protocol. The red blood cells (RBCs) were gently spun down and washed with teleost PBS (phosphate buffered saline, 420 mOsm, pH 7.8), resuspended with fresh teleost PBS at the same original blood volume, mixed with an equal volume of a preservation solution (320 mM sucrose, 40 mM sodium citrate, and 5% DMSO, pH 7.5) and flash frozen in liquid nitrogen for transport to the University of Illinois. An aliquot of the frozen cells was thawed and nuclei density was estimated using a hemocytometer. Approximately 0.5–1 × 10^6 cells per sample were fixed in 70% ethanol for 30 min on ice. Fixed nuclei were centrifuged at 1000 rpm for 7 min and stained in 0.5 ml 50 μg/ml Propidium Iodide containing 100 μg/ml RNase A, then incubated at room temperature for 30 min. Flow cytometric measurements for genome size (C value) were performed using a FACSCanto cytometer (BD Biosciences). Chicken erythrocyte nuclei (BioSure) (2C = 2.5 pg) were used as an internal standard and reference. Seven *B. saida* individuals and five *M. tomcod* individuals were examined, and each sample consisted of four replicates. Calculation of genome size in bp was based on genome size (bp) = $(0.978 \times 10^9) \times$ DNA content (pg) (Dolezel et al., 2003).

2.2. Preparation of DNA agarose plugs from blood samples

Aliquots of buffer-washed RBCs of known concentration (determined with a hemocytometer) from a single individual were embedded in 1% low melting point agarose plugs using BioRad plug molds (1 cm × 0.5 cm × 0.75 cm) to prevent shearing of high molecular weight (HMW) genomic DNA following Miyake and Amemiya (2004). Each plug contained an appropriate number of RBCs to provide roughly 30 μg of DNA, estimated using the genome size (1C value). The embedded RBCs were lysed exhaustively *in situ* within the agarose plugs using a 1% LDS lysis buffer (1% lithium dodecyl sulfate, 10 mM Tris-HCl pH 8.0, 100 mM EDTA, pH 8.0) and preserved in a 20% NDS solution (0.2% N-laurylsarcosyl, 2 mM Tris-HCl, 100 mM EDTA, pH 9.0).

2.3. CTAB treatment of DNA agarose plugs

Ten DNA plugs for each species were incubated with 20 ml CTAB extraction buffer (2% cetyltrimethylammonium bromide, 100 mM Tris-HCl, 20 mM EDTA, 1.4 M sodium chloride, pH 8.0) (Teknova) along with 2 ml 20% SDS (sodium dodecyl sulfate) and 0.4 ml proteinase K (10 mg/ml) at 50 °C overnight. CTAB is a detergent commonly used in plant DNA extraction. An additional proteinase K treatment was performed if the plugs had obvious adherent fibrous mucopolysaccharide or glycoprotein precipitates by incubating each of them in 755 μl reaction buffer containing 300 μg proteinase K, 1% N-laurylsarcosyl, 50 mM Tris-HCl pH 8.0, 100 mM EDTA pH 8.0 at 37 °C for 1 h. The plugs were then treated with 0.1 mM PMSF (phenylmethylsulfonyl fluoride) in 0.5× TE (5 mM Tris-HCl, 0.5 mM EDTA, pH 8.0) at 4 °C for 1 h to inactivate the proteinase K. All plugs were equilibrated in 0.5× TE followed by 0.5× TBE (45 mM Tris-borate, 1 mM EDTA, pH 8), each for 1 h at 4 °C before the next step.

2.4. BAC library construction

BAC library construction was performed with HMW DNA plugs from a single individual of each species, following the protocol described by Miyake and Amemiya (2004) with modification. The plugs were first pre-run (4 V/cm, field angle 120°, switch times 5 s, for 10 h) on a 1%

pulsed field agarose gel (PFG) on a CHEF Mapper® XA (Bio-Rad) electrophoresis system to eliminate small-size DNA fragments. One pre-run plug was subdivided into six equal sections and used in a series of partial digestion conditions to determine the optimal amounts of the competing enzymes *Eco*RI and *Eco*RI methyl transferase (in *Eco*RI/*Eco*RI methylase buffer with 0.5 mg/ml bovine serum albumin and 2.6 mM spermidine solution) that would maximize the production of 100–300 kbp DNA fragments. The optimal enzyme combination was found to be 10U *Eco*RI and 40U *Eco*RI methyl transferase. A preparative partial digestion of twelve half plugs was then carried out using this enzyme combination. DNA fragments in the digested plugs were resolved by size with PFG electrophoresis (5 V/cm, switch times 10 s to 60 s, included angle 120°) in 0.5 × TBE at 14 °C for 30 h. Consecutive 0.5 cm gel slices containing DNA fragment size fractions in approximately 50kbp increments were excised for the 50 kbp to 200 kbp range without staining of the DNA with ethidium bromide or exposure to UV irradiation. The DNA fragment size was determined by running a small sliver from each gel slice on an analytical PF gel. DNA fragments from the target size range were recovered from the gel slice by electro-elution and dialyzed, and the final DNA concentration was estimated by qualitative comparison to the staining intensity of known amounts of *Hind*III digested lambda DNA ladder on a 0.8% agarose gel (Fig. 1A and B). Approximately 125 ng of DNA was ligated to 12.5 ng *Eco*RI-predigested BAC vector pCC1BAC (Epicentre) and desalted. The ligated product was then electroporated into competent *E. coli* strain DH10B (Invitrogen) using a BTX ECM 630 electroporator (Harvard Apparatus). A small volume (100 μl) of transformed cells was plated on LB (Luria-Bertani) agar containing 12.5 mg/l chloramphenicol plus IPTG and X-gal for white/blue selection, and a number of white colonies were randomly used to verify cloning efficiency. DNA from these putative recombinant clones was prepared by alkaline lysis, digested with *Not*I (excises the insert), and run on a 1% PFG (6 V/cm, switch times 10 s to 40 s, included angle 120°) for 15 h to estimate the percentage of true recombinants and insert sizes. More ligations and transformations were repeated until sufficient numbers of recombinant clones were obtained to provide an estimated 10 × genome coverage for each library. Clones were picked and archived in 384-well LB/10% glycerol plates with a robotics workstation and then stored at −80 °C.

Archived BAC clones from each library were printed in duplicate as high-density macroarrays on nylon hybridization membrane filters. Each filter had six sectors, and each sector contained 2592 double-spotted clones from eight 384-well plates. Printed filters were placed clone-side up on LB agar containing 12.5 mg/l chloramphenicol and incubated at 37 °C overnight. The colonies on the filters were lysed in lysis buffer (2 × SSC (0.3 M NaCl, 30 mM Na citrate), 5% SDS) at room temperature for 3 min and then microwaved at maximum power for 2 min until completely dry. The dry filters were treated with a 10 μg/ml proteinase K solution at 37 °C for 3 h, briefly rinsed in 2 × SSC, UV crosslinked using a Stratalinker (Stratagene) while the filter was still damp, and air-dried overnight.

2.5. BAC clone insert size analysis

To estimate both the distribution of insert sizes and the average insert size of the final BAC libraries, we isolated the plasmid DNA from randomly selected clones from each 384-well plate of each library using alkaline lysis in a 96-well format (Sambrook and Russell, 2001). BAC DNA was digested with the restriction enzyme *Not*I and electrophoretically separated on 1% agarose PFG for 12 h. A regression line was generated for the PFG standard sizes versus migration distance measured from a digital gel image, and the sizes of the *Not*I-excised insert bands were calculated from their migration distances based on the regression equation.

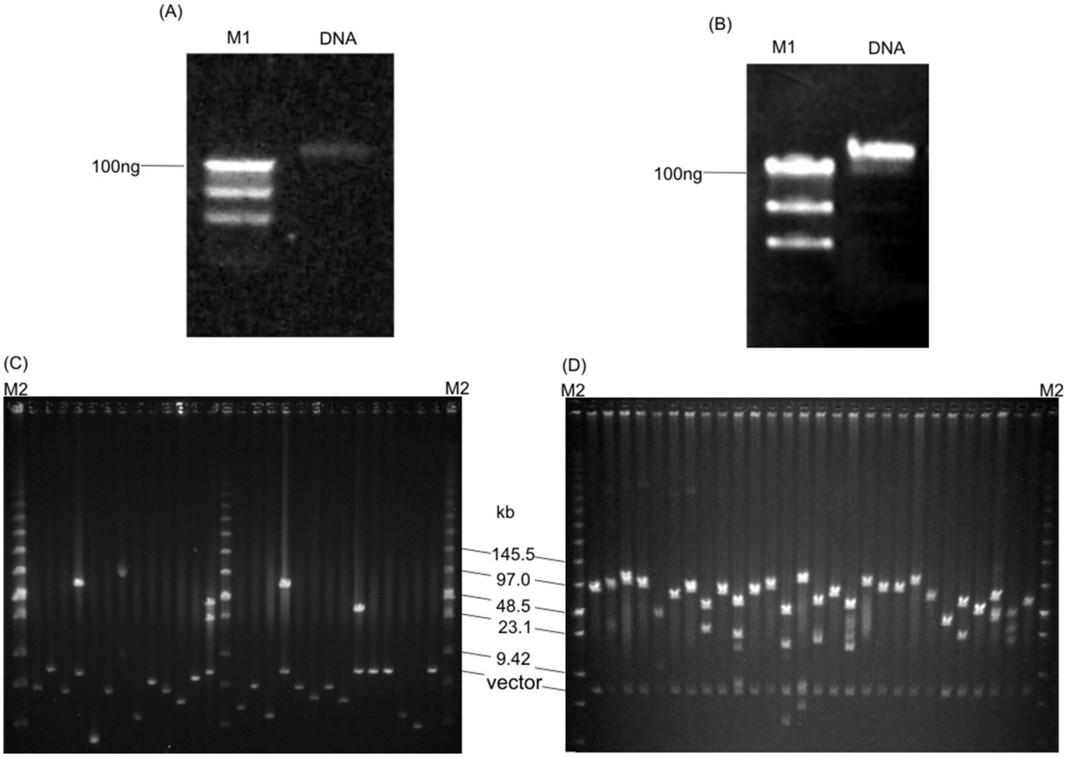

Fig. 1. Comparison of the *Boreogadus saida* BAC library construction with and without CTAB/NaCl and proteinase K treatment for DNA agarose plugs. (A) Recovered DNA (5% of total) from *Eco*RI partial digestion of DNA plugs without CTAB/NaCl and proteinase K pre-treatment; (B) Recovered DNA (5% of total) from *Eco*RI partial digestion of DNA plugs with CTAB/NaCl and proteinase K pre-treatment; (C) *Not*I digestion of randomly selected BAC clones constructed using un-pretreated DNA plugs; (D) *Not*I digestion of randomly selected BAC clones constructed using pre-treated DNA plugs. M1 is *Hind*III-cut lambda DNA. M2 is Low Range PFG marker (New England BioLabs).

Table 1
Genome coverage estimation of *B. saida* and *M. tomcod* BAC libraries by screening with housekeeping genes.

Gene	Source of organism	Primers (5′ → 3′)	Fragment length	GenBank accession no.	No. of positive clones		Gene No. in genome	Estimated genome coverage	
					B. saida	*M. tomcod*		*B. saida*	*M. tomcod*
Zic 1 (zic family member I)	*B. saida*	zic1_F (Li et al., 2007) GGACGCAGGACCGCARTAYC zic1_R (Li et al., 2007) CTGTGTGTGTCCTTTTGTGRATYTT	873 bp	MG879369	12	6	1	12	6
Myh 6 (myosin heavy polypeptide 6)	*M. tomcod*	myh6_F (Li et al., 2007) CATMTTYTCCATCTCAGATAATGC myh6_R (Li et al., 2007) ATTCTCACCACCATCCAGTTGAA	855 bp	MG879370	10	8	1	10	8
β-actin	*B. saida*	cod_β-actin_F[a] CAGAAGGACAGCTACGTTGGTGA cod_β-actin_R[a] TACTCCTGCTTGCTGATCCACATCT	1436 bp	MG879371	58	65	6–9	6–10	7–11

[a] Designed in this study based on the conserved sites in teleost β-actin genes in GenBank.

2.6. BAC library screening

To assess the quality and genome coverage of the BAC libraries, we screened the libraries with three housekeeping genes - *Zic1* (zic family member I), *Myh6* (myosin heavy polypeptide 6), and β-*actin* (a member of the actin family) (GenBank accession No. in Table 1). We PCR amplified species-specific fragments of these three genes from the genomic DNA of the gadids, which were recovered from their HMW DNA agarose plugs using β-agarase I (New England BioLabs). PCR primers sequences for each gene fragment are given in Table 1. PCR amplifications were carried out in reaction volumes of 50 μl containing approximately 200 ng of genomic DNA, 0.2 mM dNTPs, 0.2 μM each primer, 2.0 mM MgCl$_2$, 5.0 μl 10× reaction buffer, and 2 U *Taq* polymerase, and using the following cycling parameters: 94 °C initial denaturation for 3 min, 35 cycles of 94 °C denaturation for 55 s, 55 °C annealing for 55 s, and 72 °C elongation for 1 min, and a final extension at 72 °C for 7 min. PCR products were purified, ligated to the pGemT-easy vector (Promega), and transformed into competent *E. coli* strain XL1blue. Insert sequences were verified by sequencing selected recombinant clones. Zic1 and β-actin gene fragments from *B. saida* and the Myh6 gene fragment from *M. tomcod* were used for probe synthesis. As the orthologous gene fragments of *B. saida* and *M. tomcod* share higher than 90% sequence similarity, the same probes could be used for both species. The inserts of the plasmid DNA of these genes were excised and gel purified (Qiagen), and then were labeled with α^{32}P-dATP primed with random heptamers.

To identify the *AFGP*-containing clones in the BAC libraries of *B. saida* and *M. tomcod*, we screened the libraries with α^{32}P-dATP labeled probes derived from their respective species-specific AFGP gene that we cloned (*B. saida AFGP*, Chen et al., 1997; *M. tomcod AFGP*, unpublished). We utilized only the tripeptide repetitive coding sequence in the gene as a template for the probe synthesis.

BAC library macroarray filters were pre-hybridized in PerfectHyb solution (Sigma) supplemented with 100 μg/ml of denatured salmon sperm DNA overnight at 55 °C. Hybridizations with ^{32}P–labeled DNA probes in PerfectHyb were performed overnight at 55 °C; filters were then washed in the low stringency solution 2× SSC/0.1% SDS followed by the high stringency solution 0.1× SSC/0.5% SDS, from room temperature up to 55 °C. Hybridized filters were autoradiographed on a phosphor storage screen (GE Health Sciences), which was then scanned with a Molecular Dynamics STORM® 860 phosphoImager (GE Healthcare) to detect hybridized clones.

Restriction digestions of all putative positive clones containing *AFGPs* or its homologs from the libraries were performed with *Not*I, separated by PFG electrophoresis, vacuum-transferred to nylon membranes, and hybridized with the *AFGP* probe to verify that they were true positives.

2.7. FPC analysis and construction of a minimal tiling path

BAC plasmid DNA was isolated from the 101 *B. saida AFGP*-positive clones and digested with *Eco*RI and *Hind*III. The use of *Eco*RI excises the BAC vector from insert DNA, while *Hind*III produces a larger number of restriction fragments for subsequent analysis, thus both endonucleases were used in the digest. Approximately 1 μg BAC DNA was digested and electrophoresed on a 0.8% agarose gel in 1× TBE with a Wide Range Analytical DNA Marker (Promega) in every fifth lane at 2 V/cm for 18 h. The gel was stained with SYBR green I Nucleic Acid Gel Stain (Lonza) and scanned with a STORM® 860 PhosphoImager (GE Healthcare). The mobility patterns of DNA fragments in the gel were edited with the digital fingerprint band calling program IMAGE V3.10 (www.sanger.ac.uk/resources/software/image/). The band data were then analyzed in the program FPC V9.4 (Soderlund et al., 2000) (www.agcol.arizona.edu/software/fpc/), which clustered clones based on shared banding patterns into contig groups to determine the spatial overlap of these clones and predict the minimal tiling path (MTP).

2.8. Chromosomal localization of the AFGP locus by FISH (fluorescence in situ hybridization)

Metaphase chromosomes of *B. saida* and *M. tomcod* were prepared from mitotic head kidney cells using standard cytogenetics protocols optimized for polar fishes (Ghigliotti et al., 2015). Chromosomes were hybridized to an AFGP gene probe to localize the *AFGP* genomic locus in each species. Probes were derived from a plasmid clone containing a characterized *B. saida* AFGP gene (Chen et al., 1997) by nick translation labeling with biotin-16-dUTP (Roche Applied Science). The labeled probes were purified by ethanol precipitation and dissolved in hybridization buffer (50% formamide/2× SSC, 40 mM KH2PO4, 10% dextran sulfate) to yield a final concentration of 10 ng/μl.

FISH was performed according to standard procedures for fish (Bonillo et al., 2015). Briefly, the chromosomes were denatured by heating at 66 °C for 1 min in 70% formamide/2× SSC (pH 7), dehydrated in a cold ethanol series, and air-dried. The probes were denatured by heating at 75 °C for 10 min, applied to chromosomal spreads (15 μl per slide), and incubated overnight in a moist chamber at 37 °C. Post-hybridization washes were performed at 43 °C - twice in 50% (v/v) formamide/2× SSC, twice in 2× SSC, and once in 4× SSC Tween-20, for 5 min each. The bound probe was detected by incubation with streptavidin-Cy3 (Amersham Biosciences). Chromosomes were counterstained in 0.3 g/ml DAPI/2× SSC and mounted in a standard antifade solution (Vector). 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).

2.9. Next generation sequencing of B. saida AFGP-positive MTP clones and sequence assembly

The *B. saida AFGP*-positive MTP clones (comprised of eight clones) were sequenced using the Roche-454 GS-FLX Titanium platform to obtain longer reads. A 3kbp paired-end library was constructed for each BAC clone, and additional shotgun sequencing libraries were constructed using the Nextera kit (Epicentre) for the four MTP clones from the two biggest FPC contig groups – Bs52, 85, 93, and 94. All libraries were sequenced to 50–100× clone insert coverage at the University of Illinois Roy J. Carver Biotechnology Center. Short-read sequences were assembled using Roche GS *De Novo* Assembler (Newbler) V2.6, and contig scaffold(s) for each BAC clone were constructed using paired-end information. Assembly parameters of Newbler were optimized for the repetitive AFGP tripeptide coding sequences to maximize the N50 value, to minimize the number and length of gaps between contigs, and to get the total length of assembled sequences close to the length of the corresponding BAC insert estimated by FPC and from the electrophoretic mobility of *Not*I excised insert from BAC clone plasmid DNA. Accuracy of the individual BAC sequence assembly was assessed based on the agreement between the predicted restriction map from the assembled sequence and the *Hind*III + *Eco*RI fingerprinting pattern of the clone, as well as the information of the AFGP positive bands in the Southern blot of the fingerprinted gel. Some gaps between sequence contigs, caused by simple sequence repeats, could not be closed. Others were closed by sequencing gap regions that were amplified by using sequence-specific primers to contig ends. The individual BAC insert sequences were further assembled based on shared sequence identity (100%) in the overlapping regions between BAC clones. BAC end sequencing of twenty additional *AFGP*-positive clones (not in the MTP set) was used to verify the correctness of the AFGP locus assembly. Correct alignment of these BAC end sequences was accessed by shared 100% sequence identity at their matching sites, their correct 5′ and 3′ sequence orientations, and the paired-ends distance approximating the insert size of the clone.

2.10. Shotgun sequencing and assembly of M. tomcod AFGP locus

Screening of the *M. tomcod* BAC library produced a single *AFGP*-positive clone, thus we chose to construct a shotgun plasmid sequencing library for sequencing using traditional Sanger dideoxy chain termination chemistry. The DNA of the *AFGP*-positive BAC clone was first transformed into EPI300 *E.coli* (Epicentre), which was then induced to replicate the BAC plasmid to high copy number. BAC plasmid DNA was sheared into 2–5 kbp fragments and subcloned into the pCR4Blunt-TOPO vector (Invitrogen), and the shotgun library of recombinant clones was archived in 96-well plates. Plasmid DNA of shotgun clones was prepared and sequenced in 96-well format to approximately 10× BAC insert coverage using Big Dye v.3 Terminator Cycle Sequencing chemistry (Applied Biosystems) and run on an ABI3730xl sequence analyzer (Applied Biosystems) at the University of Illinois Roy J. Carver Biotechnology Center. Sequencing results were first manually proofread and edited using ChromasPro v.1.42 (Technelysium) and then assembled using Sequencher v.4.7 (Gene Codes Corp.). The insert sizes of sequenced subclones were determined by comparing the *Eco*RI excised subclone insert to a 1 kbp DNA ladder (Invitrogen) on gel electrophoresis. Assembly accuracy was assessed on the agreement between the actual insert sizes of the sequenced subclones and the distance of their paired-ends in the assembly. The contig order and gap sizes were determined by the orientation and predicted distances of the paired-end sequences; primer walking PCR amplification and sequencing were used to close gaps or to extend the sequence read of a given subclone.

3. Results and discussion

3.1. Genome size estimations

The genome sizes of *B. saida* and *M. tomcod* were relatively small, at 836 Mbp (2C = 1.71 ± 0.17 pg; n = 7) and 645 Mbp (2C = 1.32 ± 0.02 pg, n = 5) respectively. The genome size of *B. saida* is consistent with older estimates of 866 Mbp (2C = 1.77 pg) by densitometry analysis of digital images of Feulgen stained blood smears (Hardie and Hebert, 2003). The estimated genome sizes of the two gadids were subsequently used for determining the genome coverage of their BAC libraries.

3.2. BAC library and macroarray construction

The deep-coverage large DNA insert BAC library we constructed for *B. saida* comprised 92,160 clones, which were archived into 240,384-well plates and printed on five macroarray filters. The *M. tomcod* library comprised 73,728 clones, which were archived into 192,384-well plates and printed on four macroarray filters.

Our previous experience with BAC library construction has shown that HMW DNA preparation, fragment size selection, ligations and transformations must be performed with great care to obtain optimal BAC libraries. Small DNA fragments resulting from inadvertent mechanical shearing can outcompete the large DNA fragments in the ligation and transformation and thus would reduce the insert sizes of the library. Even with careful attention to eliminate small DNA fragments, it appeared to be inadequate in our initial attempts at constructing high-quality BAC library of these two gadids. We found that library construction following an existing protocol (Miyake and Amemiya, 2004), which has been successful applied to an Antarctic notothenioid fish (Nicodemus Johnson, 2010), did not yield similar success with the gadids. In early attempts, despite repeated attempts at cloning and electroporation optimizations, the yield of recombinant clones remained low. Similar problems have been reported for other marine animals (e.g. Pacific white shrimp) (Zhang et al., 2010). In tracing the cause of poor cloning yield, we found that the agarose plugs of embedded gadid RBC DNA exuded a viscous, gelatinous substance with time during storage in 20% NDS. We suspected that gadid RBCs may have large amounts of cell surface glycoproteins that were incompletely removed during the *in situ* lysis of agarose-embedded RBCs and preservation of DNA. This prompted the CTAB treatment of the agarose plugs. CTAB is a cationic detergent often used to release DNA from the complex polysaccharides and glycoproteins common in plants, but seldom used in DNA isolation from animal cells. Pretreating the gadid DNA plugs with CTAB/NaCl buffer plus Proteinase K greatly increased the yield of DNA fragments of desired size in subsequent preparative *Eco*RI partial digestions, whereby sufficient numbers of recombinant clones were achieved (Fig. 1). The source of our initial cloning failure therefore was very likely due to the incomplete removal of membrane glycoproteins associated with red blood cells during agarose plug preparation and processing. DNA trapped within the glycoprotein matrix expectedly would become much less accessible to enzymatic digest by *Eco*RI, leading to low yields of DNA fragments. In addition, the glycoproteins adhering to the eluted DNA fragments likely also hinder the ligase during ligation to the BAC vector, leading to poor recombinant yield.

3.3. Insert size distribution and genome coverage estimation

To evaluate the quality and coverage depth of the two gadid BAC libraries, we randomly sampled 0.7–0.9% BAC clones from each library to assess the percentage of true positive clones and to analyze clone

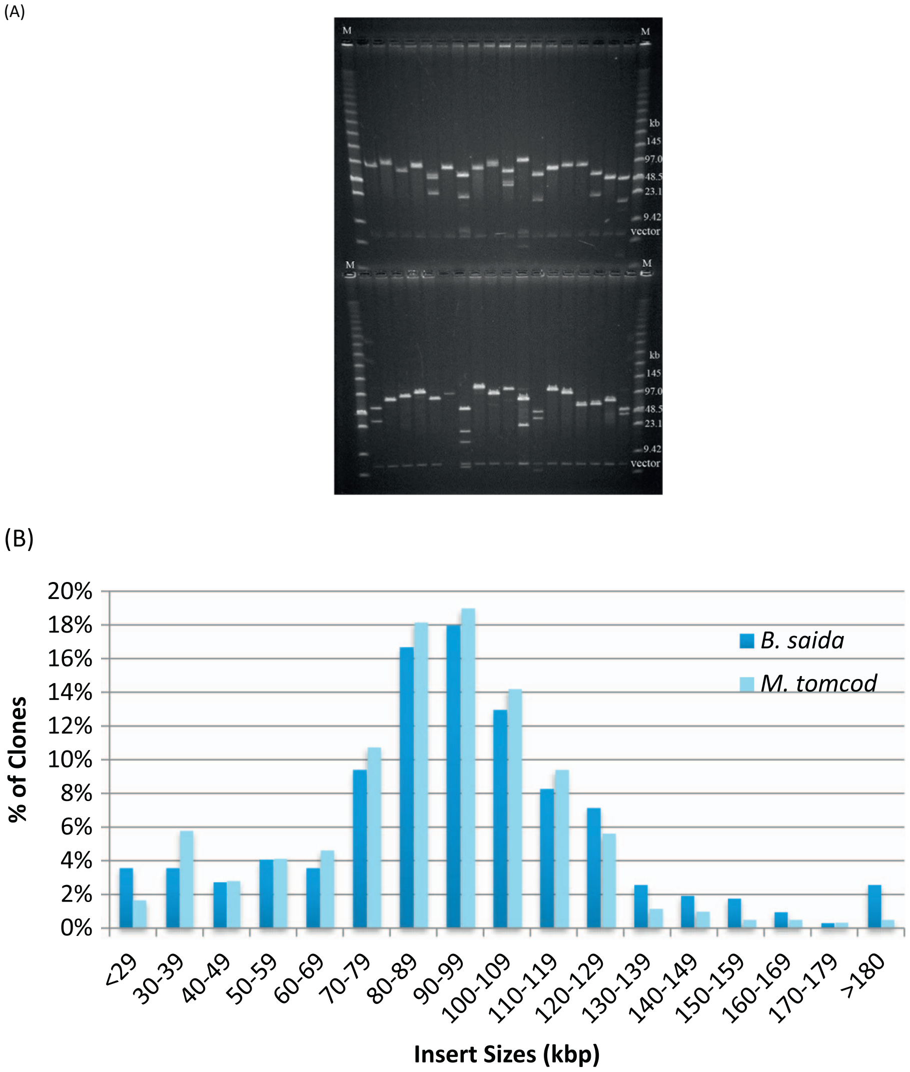

Fig. 2. Insert sizes of randomly selected clones from *B. saida* and *M. tomcod* BAC library. (A) BAC plasmids were digested with the restriction endonuclease *Not*I, which cleaves near the *Eco*RI cloning site in the vector, and the insert DNA and the vector were separated on 1% agarose pulsed field gel electrophoresis in 1 × TBE. MW standard lanes indicated as M contain Low Range PFG Marker (New England Biolabs). The vector band is labeled. The upper tier in the gel shows digested BAC plasmids of *B. saida* and lower tier shows those of *M. tomcod*. (B) Histograms show the insert size distributions of 645 and 637 randomly selected clones from the *B. saida* and *M. tomcod* library, respectively.

insert sizes. A total of 645 and 637 clones were randomly picked from the *B. saida* and the *M. tomcod* libraries respectively, and BAC plasmids were digested with a *Not*I restriction endonuclease. Internal *Not*I recognition sites within the BAC insert DNA, presented as more than one insert bands, were identified at substantial frequency, in 46.6% and 43.3% of the selected clones from *B. saida* and *M. tomcod* library respectively. The *Not*I recognition sequence is 5′GCGGCCGC3′, and thus high frequency of *Not*I sites in these two species suggests their genomes may contain GC-rich sequences.

The majority of cloned DNA fragments in the BAC libraries of *B. saida* and *M. tomcod* ranged from 70 kbp to 130 kbp (Fig. 2). Approximately 80% of the clones in both libraries contained inserts > 70 kbp. The average insert sizes of these two libraries were 94.7 kbp for *B. saida* and 89.6 kbp for *M. tomcod*. The insert size range of *M. tomcod* was narrower and the average size was slightly smaller than the those of *B. saida*, because the *M. tomcod* library was constructed using the DNA from only one gel size fraction, 70 kbp to 120 kbp, while the *B. saida* library was constructed using two gel size fractions, 90 kbp to 140 kbp and 120 kbp to 170 kbp. Apparently the DNA of sizes smaller than 130 kbp was more easily cloned, as although larger-size DNA gel fractions (up to 170 kb) were used in constructing the *B. saida* library, it did not increase the fraction of clones with larger insert sizes (> 130 kbp). The percentage of clones without an insert was 4.19% for *B. saida* and 4.87% for *M. tomcod*, indicating that > 95% of the clones in these libraries are recombinant clones.

We determined the genome coverage of the BAC libraries as follows: genome coverage of a library = (average insert size × total number of clones × percentage of recombinant) / genome size. Based on the data from the above analyses, the genome coverage of the *B. saida* and *M. tomcod* libraries is estimated to be 10-fold and 9.7-fold, respectively.

3.4. Estimation of genome coverage by hybridization with housekeeping genes

To validate the genome coverage of the libraries and test their utility for gene characterization, we hybridized each library with three nuclear gene probes, two of which are single-copy genes (*Zic1* and *Myh6*) and the other is a low-copy gene (β-*actin*) in the genome of these gadids (Table 1). Given the fact that these BAC libraries were constructed in *Eco*RI cloning sites and no *Eco*RI recognition site was found in any of these three gene fragments that we used as a probe, each of these gene fragments will not be subdivided into more than one BAC clone but will reside within a single clone. Thus the number of positive clones that hybridized with each gene probe represents the actual gene count in the library. If the probe is a single copy gene in the genome, this number would represent the fold of genome coverage encompassed by the library. The strongly positive clones identified from screening the *B. saida* and *M. tomcod* libraries with the Zic1 gene fragment were 12 and 6 respectively, and 10 and 8 respectively when screened with Myh6 (Table 1), which are close to the estimated genome coverage of the libraries based on cloning statistics. Besides these positive clones with strong hybridization signals, there are also many weakly hybridizing clones in both libraries with the Myh6 gene probe. These weakly hybridized clones probably contain the sequences that share low level sequence similarity with the probe sequence and thus could be distinguished and excluded from the strongly positive clones. The β-actin gene is a member of actin multigene family, and conserved domains in the probe sequence could cross-hybridize to paralogs and result in more positive clones than the actual β-actin containing clones. Teleost fishes were reported to contain a variety of six to nine actin genes (Hall et al., 2003). Using the primers designed based on the conserved regions of teleost β-actin (Table 1) in PCR amplification of genomic DNA, we isolated two β-actin genes or pseudogenes from *B. saida* and three from *M. tomcod*. Since these β-actin genes and pseudogenes share > 80% sequence similarity, it is difficult to differentiate them among the positive hybridized clones. If the number of homologous genes in the cod genome is also around six to nine, then based on the number of β-actin positive clones in each library (Table 1), the estimated genome coverage of *B. saida* is around six to ten, and *M. tomcod* is around seven to eleven. Therefore, the library screening results with either single-copy or low-copy genes were consistent with the genome coverage calculated from the average insert size and total clone numbers.

3.5. Identification of AFGP-positive clones

A total of 102 putative *AFGP*-positive clones were identified from the *B. saida* BAC library, and 101 of them were verified as true positive by Southern blot hybridization of the *Not*I digested clones with an AFGP DNA probe. Only one AFGP-positive BAC clone was identified and verified in the *M. tomcod* library, even though this library was additionally screened with AFGP gene probes specific to this particular individual using cloned *AFGP* sequences identified from the shotgun library after sequencing. The large number of *AFGP*-positive clones in *B. saida* is commensurate with a large *AFGP* genomic locus, while the single positive clone in *M. tomcod* is consistent with a small locus, > 500kbp and about 35 kbp respectively (detailed in a later section). The large number of *AFGP*-positive clones in *B. saida* corroborates a deep coverage (10 ×) of the genome or the AFGP genomic region in the BAC library. It is puzzling therefore that with comparable depth of genome coverage (9.7 ×) only a single positive clone was identified from the *M. tomcod* BAC library. A possible cause of this curious discrepancy between the numbers of *AFGP*-positive clones in these two species may lie in the distributions of *Eco*RI restriction sites within and/or surrounding their *AFGP* genomic regions. The two BAC libraries were constructed with *Eco*RI partially digested DNA fragments. Based on the sequences of their *AFGP* loci (detailed in a later section), no *Eco*RI site is present within the *M. tomcod AFGP* locus and only 10 sites flanking the locus, whereas 74 *Eco*RI sites were found in *B. saida AFGP* locus, in the intergenic sequences between AFGP genes and flanking the gene clusters. Consequently, the frequency of partial *Eco*RI digest fragments generated from the *M. tomcod* AFGP genomic region and their subsequent cloning would be much lower than from the *B. saida AFGP* locus. In addition, chromosomal FISH with the *AFGP* probe suggests that this particular *M. tomcod* individual is heterozygous for the *AFGP* loci (detailed in the next section), which would additionally reduce the cloning frequency. Regardless, it remains curious that despite 9.7× overall genome coverage, only a single *AFGP*-positive clone was represented in the *M. tomcod* BAC library. Given the multiple-fold genome coverage of both BAC libraries, confirmed by library statistics and screening with single-copy and low-copy genes, most or all regions in the genome should have been well represented. Thus we reasonably submit that the single *AFGP*-positive BAC clone from the *M. tomcod* library encompasses the whole *AFGP* genomic locus in this species.

3.6. Finger Printed Contig (FPC) analysis of the minimal tiling path and chromosomal localization of the AFGP locus

Screening of the *B. saida* BAC library produced 101 *AFGP*-positive clones. Since the estimated coverage of the library is around ten-fold, many of the 101 positive clones would be redundant and represent overlapping regions. We therefore used a FPC analysis of the restriction endonuclease fingerprints of these clones to reduce the redundant set to a minimum number of overlapping clones that would span the *AFGP* genomic region, *i.e.* the minimal tiling path (MTP). FPC clusters clones into contig groups based on the coincidence probability scores of shared restriction banding patterns. FPC assembled 90 of the 101 clones into six contig groups (Fig. 3). The remaining 11 clones were ungrouped singletons, which are of relatively small sizes and therefore likely produced inadequate number of shared bands for FPC to predict their overlap with the other clones. To ensure the MTP clones we selected for sequencing covered all AFGP genes in the genome, we performed an additional verification besides the conventional FPC analysis. The DNA bands in the endonuclease fingerprinting gel were transferred onto nylon membrane and hybridized with an AFGP coding sequence probe. The AFGP-positive bands of the 101 clones on the autoradiograph could be classified into 12 categories by fragment size (Fig. S1). The minimum number of overlapping clones encompassing all 12 size categories of positive bands were then identified from the FPC contig groups. We found eight *AFGP*-positive BAC clones to comprise the MTP set, red-highlighted in Fig. 3. The single *AFGP*-positive clone in the *M. tomcod* library was treated as the MTP clone.

FPC generated multiple, apparently non-overlapping contig groups of the *B. saida AFGP* locus. To determine whether these contig groups represented distinct loci in separate genomic locations or co-localized to a single chromosomal region, we mapped the *AFGP* locus in chromosomes by FISH using the repetitive tripeptide coding sequence of AFGP gene. *AFGP* hybridization localized to a single site in one pair of chromosomes in *B. saida* (Fig. 4A and B), indicating the clones of all the *AFGP*- containing FPC contig groups belonged to a single genomic region, with intervening distances below the spatial resolution of chromosomal FISH. In *M. tomcod*, *AFGP* hybridization mapped to a site in a single chromosome rather than a pair of homologous chromosomes (Fig. 4C and D), indicating the particular individual for which the BAC library was constructed was heterozygous for the *AFGP* locus. As indicated above, this could be a contributing factor to the presence of a single *AFGP*-positive BAC clone from *M. tomcod.* Of broader interest, the *AFGP* locus heterozygosity observed here suggests there is an interbreeding individual (or population) of this species that lacks the AFGP trait. Whether it belongs to the population in the nearby Hudson River, where the freshwater environment would not require antifreeze protection thereby leading to trait loss, is an interesting hypothesis to pursue.

3.7. Sequencing strategy for highly repetitive AFGP sequence

AFGPs are encoded as large polyprotein precursors composed of tandem repeats of the tripeptide Thr-Ala/Pro-Ala (Chen et al., 1997). The very long and highly repetitive 9-nucleotide coding sequences in the gene present difficult challenges in sequence assembly. Repetitive sequences are often underrepresented in genome assemblies from short-read sequences generated from Second Generation sequencers, because short tandem repeat sequences generally collapse and form contracted contigs with atypically high depths, producing gaps in regions where the repeat sequence should span (Alkan et al., 2011). The difficulty of assembling repetitive sequences led to the exclusion of the AFGP gene

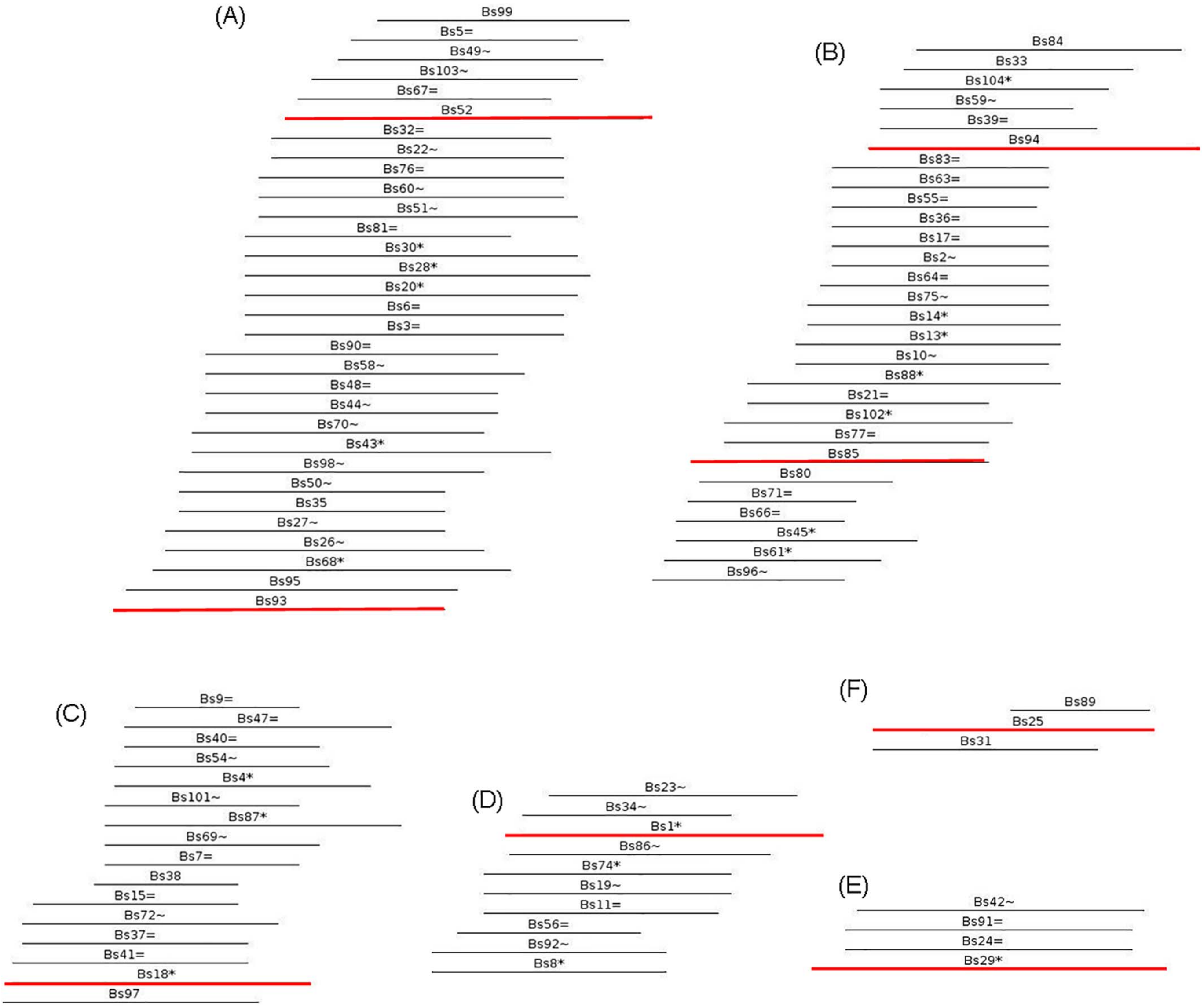

Fig. 3. FPC assemblies of *Hind*III + *Eco*RI fingerprints of 101 true *AFGP*-positive clones from the *B. saida* BAC library (Bs1–102; Bs12 is false positive and thus excluded). Ninety clones were assembled into six contig groups (A to F) and the remaining 11 clones were ungrouped singletons (not shown). A clone name ending with a "*" indicates that it has buried clone(s) shown in the contig group. A clone name ending with a "=" has all the same bands as the parent clone. A clone ending with a "~" has approximately the same set of bands as the parent clone. Red lines indicate the MTP clones selected for sequencing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

family (Zhuang et al., 2012) in the first draft genome (gadMor1) of the Atlantic cod *Gadus morhua* (Starr et al., 2011). The subsequently refined gadMor2 assembly was improved with Third Generation PacBio long reads (Tørresen et al., 2017) and was able to assemble the majority of AFGP sequences (Baalsrud et al., 2017). The assembly of tandem repeats can only be resolved when the sequence read length is sufficient to cover the entire repetitive region and extend into the flanking non-repetitive regions. Third generation sequencing platforms such as PacBio or Oxford Nanopore, capable of generating very long reads from HMW input DNA, can provide this solution, but the cost and capacity are not justified for much smaller subgenomic regions or BAC clones. For these, the default longest read option is Sanger sequencing, with the caveat that the necessary shotgun plasmid library preparation from the parent BAC clone, and low throughput 96-well sequencing format is labor-intensive and time-consuming. Second Generation sequencing such as the dominant Illumina platform has tremendous throughput and sequencing depth but the very short read lengths remain as an unresolved obstacle for assembling repeats. Whole genome Illumina shotgun sequencing and assembly of some AFGP-bearing gadids (including *B. saida*) was recently carried out (Malmstrøm et al., 2017). However, although some AFGP gene fragments can be identified by their 5′ non-repetitive portion (Baalsrud et al., 2017), the highly repetitive AFGP coding regions could not be assembled properly.

We decided to sequence the eight *B. saida* AFGP MTP clones using the Second Generation pyrosequencing platform Roche 454 GS-FLX Titanium, as it generates longer read lengths than Illumina and was still available at the start of this study at the Biotechnology Center at the University of Illinois. For the single positive BAC clone of *M. tomcod* clone, we used Sanger sequencing. For the *B. saida* clones, we sequenced both shotgun libraries and 3 kbp paired-end libraries. The shotgun sequencing gave longer reads, while the paired-end reads aided in ordering assembled contigs and forming scaffold(s). The average 300–400 nt read lengths of 454 pyrosequencing are shorter than the 800–1000 nt reads of Sanger sequencing, but the ~100× sequence coverage of the BAC insert and the reliable paired-end distance information enabled us to assemble most of the highly repetitive AFGP

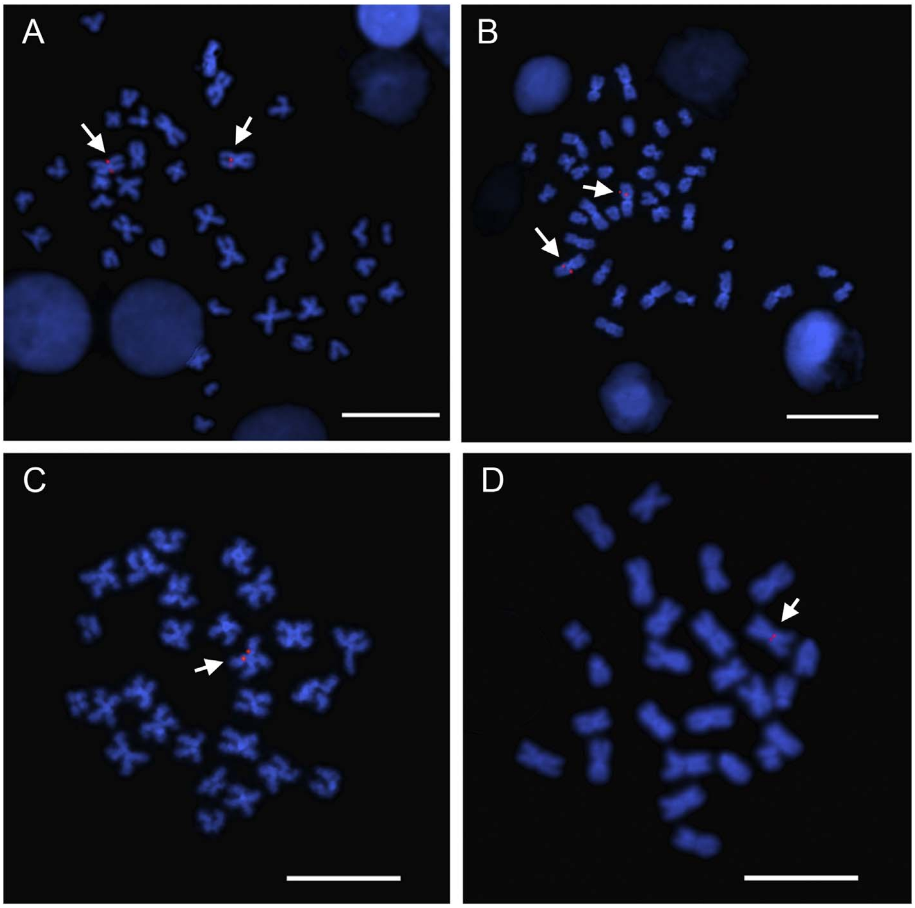

Fig. 4. Fluorescence in situ hybridizations (FISH) on *B. saida* (A and B) and *M. tomcod* (C and D) metaphase chromosomes using *B. saida* AFGP gene as probe.
Red dots are the hybridization signals. (A, B) AFGP genes map to a single chromosomal site in a pair of chromosomes in *B. saida*. (C, D) AFGP genes map to a single chromosomal site in one chromosome in *M. tomcod*. Scale bars = 10 μm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

coding regions. After additional efforts at gap closure, the *B. saida* AFGP MTP clones were assembled with only two to four gaps per clone, while the single *M. tomcod* clone using Sanger sequencing ended up with two gaps. Most of these gaps are in simple sequence repeats. Both *B. saida* and *M. tomcod* have a few unclosed gaps in the long repetitive AFGP tripeptide coding regions, as the highly repetitive sequences still cannot be completely resolved by 454 or Sanger sequencing. In comparing these two methods we used, the plasmid-based Sanger long read sequencing showed no obvious superiority over 454 paired-end sequencing in terms of the repeats assembly quality, while the 454 sequencing greatly exceeds Sanger sequencing by its high throughput and is therefore a better platform for sequencing a large number of BAC clones. However 454 has since been phased out by Roche, and the newer generations of the dominant Illumina have yet to reach the same read lengths.

3.8. Sequence assembly of the AFGP locus in two gadids

Since the gadids have diploid genomes, the clones in their genomic DNA BAC libraries should comprise the DNA from each of the two chromosome sets, and therefore two haplotypes of the *AFGP* locus would be expected in the sequence assembly. The sequence assembly of the eight MTP BAC clones of *B. saida* produced five sequence contigs encompassing three AFGP gene clusters (I-III) (Fig. 5). Cluster I contains seven AFGP genes, and the two sequence contigs containing this cluster represent two haplotypes. Haplotype 1 was assembled from MTP clones Bs93 and Bs52 with an overlap of 38 kbp; haplotype 2 was assembled from MTP clones Bs94 and Bs85 with 33 kbp overlap. The two sequence contigs of cluster I are 199 and 183 kbp in length respectively for haplotype 1 and 2. AFGP gene cluster II contains two AFGP genes, found in clones Bs29, Bs1 and Bs18. One haplotype was assembled from MTP clones Bs29 and Bs1 with 17 kbp overlap, and spans 233 kbp in total length. The other haplotype consists of only one clone Bs18, which is about 138 kbp in length. AFGP gene cluster III contains seven AFGP genes contained within a single MTP clone (Bs25) with a total length of 80 kbp. The other haplotype of cluster III was subsequently identified in FPC singleton clone Bs57, which then was also sequenced. The three *AFGP* clusters contain a total of 16 AFGP genes and span a combined distance of 512 kbp, indicating the total *B. saida* AFGP genomic locus is quite large.

These assembled sequence contigs from the MTP BAC clones are consistent with the contig grouping of these clones predicted by FPC analysis (Fig. 3). The only exceptions are clones Bs1 and Bs29, which both contain the two cluster II AFGP genes but mapped to separate contig groups (D and E respectively) in the FPC analysis. This was likely due to the overlapping region shared by the two clones being too short to generate sufficient restriction fragments for FPC to determine their overlap. The separate FPC contig groups also confirm the distinct haplotypes of each AFGP cluster. Clones belonging to the same haplotype should share 100% sequence identity in their overlapping regions, while clones from a separate haplotype would exhibit allelic variations. We found the homologous regions in the two haplotypes contain SNPs (single nucleotide polymorphism), VNTRs (variable number tandem repeat), insertions and deletions. These sequence variations between haplotypes could generate different restriction fingerprints in the FPC analysis, resulting in the separation of the two haplotypes into non-overlapped contig groups.

To determine the linear order and orientation of the three *B. saida* AFGP clusters, we attempted to find the BAC clones that could connect adjacent clusters. We therefore sequenced the two insert ends of 20 *AFGP*-positive BAC clones, including all 11 ungrouped singletons and nine potential “edge” clones from the FPC contig groups. These BAC

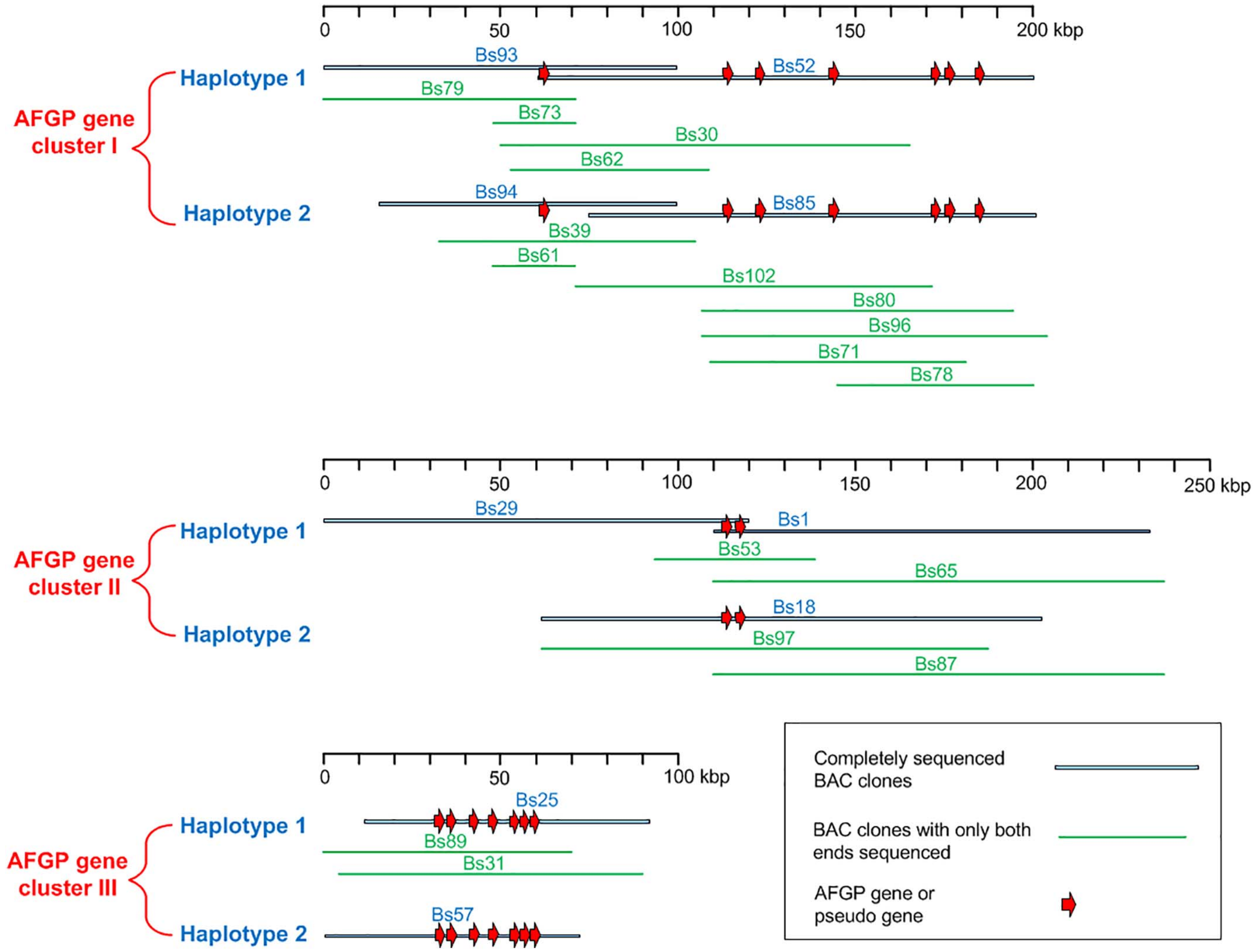

Fig. 5. Reconstruction of three AFGP genes clusters in the *B. saida* AFGP genomic locus. Blue lines represent assembled BAC insert sequences from eight MTP clones of the fingerprinted contig (FPC) groups (Bs93, Bs52, Bs94, Bs85, Bs29, Bs1, Bs18 and Bs25) and one ungrouped FPC singleton Bs57. Green lines represent the BAC clones whose BAC ends were sequenced. These BAC clones (green lines) are shown underneath the assembled haplotype (blue lines) they belong to. Red arrows denote AFGP genes or pseudogenes. Line locations in each cluster indicate their actual relative spatial positions. The corresponding regions in both haplotype assemblies are aligned vertically. Scale bars indicate the sizes of contigs and clones. Lines are drawn to scale but AFGP genes are not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

insert end sequences were mapped to the five sequence contigs of three AFGP clusters (Fig. 5) with 100% sequence identity. The distances of the paired BAC insert ends on the sequence contigs matched the BAC insert lengths estimated from *Not*I digest; this further supports the accuracy of the sequence assemblies. However, none of the 20 clones could bridge the contigs of any two AFGP clusters as they all largely overlap with the reconstructed sequence contigs. This suggests that the MTP clones we selected for sequencing and reconstructing the AFGP locus likely have covered the largest possible genomic region spanned by the FPC contig groups, and therefore the remaining unsequenced *AFGP*-positive clones may not provide additional sequence information to order the three *AFGP* clusters contigs. In other words, the BAC clones containing the sequences in between these cluster contigs are not present among the *AFGP*-positive clones. Thus we deduce that the three AFGP clusters are intervened by substantial distances of sequences containing no AFGP genes.

The assembly of the AFGP locus in *M. tomcod* was simple as it consisted of only one BAC clone. The assembled BAC insert sequence of this single *AFGP*-positive clone was approximately 75 kbp, encompassing an *AFGP* locus of approximately 35kbp that contains three AFGP genes and one AFGP pseudogene (Fig. 6).

The large AFGP gene family and thus a large AFGP gene dosage in *B. saida* could significantly contribute to the production of abundant protein. *B. saida* is one of the two true high Arctic gadid species (the other is the ice cod *Arctogadus glacialis*) and is principally distributed north of the Arctic Circle and into the Arctic basin (Cohen et al., 1990; Howes, 1991). These habitats represent extreme low temperatures and prolonged freezing conditions, likely driving the evolutionary expansion of the AFGP genotype in *B. saida* for synthesizing high levels of the protective AFGPs. *M. tomcod* inhabits relatively milder lower latitudes, from the coast of southern Labrador to Virginia (Howes, 1991), where shallow waters freeze only for short periods in the winter. *M. tomcod* would therefore require less protective AFGPs than *B. saida* and only in the winter. The seasonal cycle of AFGP levels in *M. tomcod* reaches its peak in January, starts to reduce when water temperatures begin to rise, and is at insignificant levels by late spring (Reisman et al., 1987). Even its peak winter AFGP levels (2 mg/ml) (Fletcher et al., 1982) are considerably lower than the high AFGP activities observed in *B. saida* (Denstad et al., 1987), which persist even in the Arctic summer (Enevoldsen et al., 2003). The AFGP phenotype is therefore commensurate with the size of the AFGP gene family and gene dosage determined for the two species - 16 genes in *B. saida* and four genes in *M.*

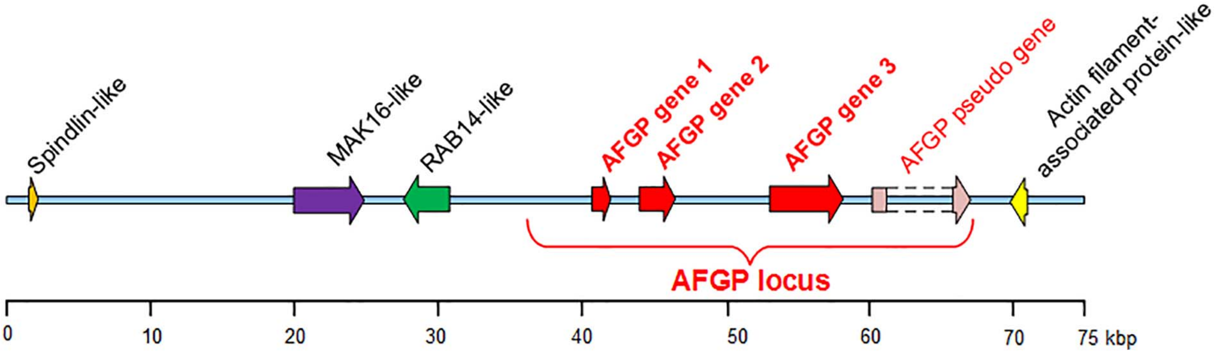

Fig. 6. Schematic of the AFGP locus and its neighboring genomic regions in *M. tomcod*. Three intact AFGP genes are denoted as red arrows. The fragments of one AFGP pseudogene are denoted as light-red box and arrow connected with dashed lines. Other hypothetical protein-coding genes or coding regions in the neighboring region are represented by arrows of different colors. Arrows point in the sense direction of the gene. All genes are drawn to scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tomcod – in this study.

4. Conclusion

Despite voluminous DNA sequences in databases, no homolog of gadid AFGP gene exists to aid in inferring its ancestry. To enable our determination of how and from where the gadid AFGP genotype evolved, we used a targeted approach that was essential for characterizing the highly repetitive AFPG genes and gene family. We constructed a BAC library for each of two AFGP-bearing gadid fishes, the polar cod *B. saida* and the Atlantic tomcod *M. tomcod*, and isolated their respective *AFGP* genomic regions for sequencing and assembly. We solved the initial problem of cloning inefficiency by pre-treating agarose-embedded erythrocyte DNA plugs with the cationic reagent CTAB (cetyltrimethylammonium bromide) to release DNA trapped in cell surface glycoproteins prior to partial enzymatic digestion. This added treatment would likely benefit other efforts at BAC cloning for teleost species, and/or using erythrocytes or other tissues with high glycoprotein or mucopolysaccharide content as the source of DNA. We estimated the genome sizes of the two gadid species using flow cytometry, and in conjunction with the average insert size of randomly sampled clones and the number of clones in each library, estimated the depth of genome coverage of the polar cod and Atlantic tomcod BAC libraries to be 10 and 9.7 fold respectively. We then isolated the *AFGP* loci from the libraries, sequenced the selected clones, and reconstructed the *AFGP* loci from sequence assemblies. We found the AFGP gene dosage of these two gadids is commensurate with the respective level of AFGP they synthesize, and consistent with the degree of environmental severity they encounter. By comparing the efficacy of conventional plasmid-based Sanger sequencing that generates long reads (used for the *M. tomcod AFGP* locus) with Second Gen (Roche 454) sequencing that generates shorter reads but in much greater depth (used for the *B. saida AFGP* locus) in subsequent assemblies of highly repetitive AFGP coding regions, we found that paired-end 454 pyrosequencing is better suited for a large numbers of BAC clones that contain such highly repetitive sequences. Unfortunately the Roche 454 platform has been phased out since this study. The reconstructed *AFGP* genomic loci of the two gadid species in this study provide the needed data at near-contiguous sequence level for our fine-scaled comparative analyses to decipher the definitive genomic origin and molecular mechanisms of gadid AFGP gene evolution, the evolutionary history of AFGP genes in the gadid lineage, and the mechanism of gene family expansion in the respective species. In addition, the gadid BAC libraries serve as useful resources for targeted, fine resolution sequencing and analysis of traits at the genomic level, as well as finishing draft genome assemblies and enabling investigations of the genetic changes associated with the evolutionary adaptations of gadids to the cold Arctic marine environments.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.margen.2018.02.003.

Acknowledgements

We greatly appreciate Dr. Chris Amemiya for his insightful suggestions on the construction of BAC libraries and for making his lab facility available for our use. We thank Andrew Stuart of the Amemiya lab for his technical advice. We sincerely thank Dr. Jørgen Christiansen for the opportunity to participate in the TUNU cruises to the East Greenland coast to collect polar cod. And again, we thank Dr. Howard Reisman for the collection of Atlantic tomcod. This work was supported by National Science Foundation grant DEB 0919496 to CHCC.

References

Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61–65. http://dx.doi.org/10.1038/nmeth.1527.

Baalsrud, H.T., Tørresen, O.K., Solbakken, M.H., Salzburger, W., Hanel, R., Jakobsen, K.S., Jentoft, S., 2017. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol. Biol. Evol. http://dx.doi.org/10.1093/molbev/msx311.

Bonillo, C., Coutanceau, J.-P., D'Cotta, H., Ghigliotti, L., Ozouf-Costaz, C., Pisano, E., 2015. Standard Fluorescence in situ Hybridization Procedures. In: Ozouf-Costaz, C., Pisano, E., Foresti, F., Foresti de Almeida Toledo, L. (Eds.), Fish Cytogenet. Tech. CRC Press, Taylor and Francis Group, London, pp. 103–117.

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

Cheng, C.-H.C., 1998. Origin and mechanism of evolution of antifreeze glycoproteins in polar fishes. In: Fishes of Antarctica. A Biological Overview, pp. 311–328.

Cheng, C.-H.C., Cziko, P.A., Evans, C.W., 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. http://dx.doi.org/10.1073/pnas.0603796103.

Cohen, D.M., Inada, T., Iwamoto, T., Scialabba, N., 1990. FAO species catalogue. In: v. 10: Gadiform Fishes of the World (Order Gadiformes). FAO Fisheries Synopsis (FAO).

Coulson, M.W., Marshall, H.D., Pepin, P., Carr, S.M., 2006. Mitochondrial genomics of gadine fishes: implications for taxonomy and biogeographic origins from whole-genome data sets. Genome 49, 1115–1130. http://dx.doi.org/10.1139/g06-083.

Craig, P.C., Griffiths, W.B., Haldorson, L., McElderry, H., 1982. Ecological studies of arctic cod (*Boreogadus saida*) in beaufort sea coastal waters, Alaska. Can. J. Fish. Aquat. Sci. 39, 395–406. http://dx.doi.org/10.1139/as-2016-0056.

Denstad, J.-P., Aunaas, T., Börseth, J.F., Vollan Aarset, A., Zachariassen, K.E., 1987. Thermal hysteresis antifreeze agents in fishes from Spitsbergen waters. Polar Res. 5, 171–174.

DeVries, A.L., 1983. Antifreeze peptides and glycopeptides in cold-water fishes. Annu. Rev. Physiol. 45, 245–260. http://dx.doi.org/10.1146/annurev.ph.45.030183.001333.

Dolezel, J., Bartos, J., Voglmayr, H., Greilhuber, J., 2003. Nuclear DNA content and genome size of trout and human. Cytometry 51, 127.

Enevoldsen, L.T., Heiner, I., DeVries, A.L., Steffensen, J.F., 2003. Does fish from the Disko Bay area of Greenland possess antifreeze proteins during the summer? Polar Biol. 26, 365–370. http://dx.doi.org/10.1007/s00300-003-0489-9.

Fletcher, G.L., Hew, C.L., Joshi, S.B., 1982. Isolation and characterization of antifreeze glycoproteins from the frostfish, *Microgadus tomcod*. Can. J. Zool. 60, 348–355. http://dx.doi.org/10.1139/z82-046.

Ghigliotti, L., Cheng, C.-H.C., Ozouf-Costaz, C., Vacchi, M., Pisano, E., 2015. Cytogenetic diversity of notothenioid fish from the Ross sea: historical overview and updates. Hydrobiologia 761, 373–396.

Gradinger, R.R., Bluhm, B.A., 2004. *In-situ* observations on the distribution and behavior of amphipods and Arctic cod (*Boreogadus saida*) under the sea ice of the High Arctic Canada Basin. Polar Biol. 27, 595–603. http://dx.doi.org/10.1007/s00300-004-0630-4.

Hall, T.E., Cole, N.J., Johnston, I.A., 2003. Temperature and the expression of seven muscle-specific protein genes during embryogenesis in the Atlantic cod *Gadus morhua*. J. Exp. Biol. 206, 3187. http://dx.doi.org/10.1242/jeb.00535.

Hardie, D.C., Hebert, P.D.N., 2003. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome 46, 683–706. http://dx.doi.org/10.1139/g03-040.

Howes, G.J., 1991. Biogeography of gadoid fishes. J. Biogeogr. 18, 595–622. http://dx.doi.org/10.2307/2845542.

Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7 (1), 44. http://dx.doi.org/10.1186/1471-2148-7-44.

Malmstrøm, M., Matschiner, M., Tørresen, O.K., Jakobsen, K.S., Jentoft, S., 2017. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Sci. Data 4, 160132. http://dx.doi.org/10.1038/sdata.2016.132.

Miyake, T., Amemiya, C.T., 2004. BAC libraries and comparative genomics of aquatic chordate species. Comp. Biochem. Physiol. C Toxicol. Pharmacol. 138, 233–244. http://dx.doi.org/10.1016/j.cca.2004.07.001.

Nicodemus Johnson, J.D., 2010. Analysis of the Antifreeze Glycoprotein Containing Genomic Locus in the Antarctic Notothenioid Fish Dissostichus Mawsoni. Doctoral dissertation. University of Illinois at Urbana-Champaign.

Reisman, H.M., Fletcher, G.L., Kao, M.H., Shears, M.A., 1987. Antifreeze proteins in the grubby sculpin, *Myoxocephalus aenaeus* and the tomcod, *Microgadus tomcod*: comparisons of seasonal cycles. Environ. Biol. Fish 18, 295–301. http://dx.doi.org/10.1007/BF00004882.

Sambrook, J., Russell, D.W., 2001. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press.

Soderlund, C., Humphray, S., Dunham, A., French, L., 2000. Contigs built with fingerprints, markers, and FPC V4. 7. Genome Res. 10, 1772–1787. http://dx.doi.org/10.1101/gr.GR-1375R.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrom, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., et al., 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477 (7363), 207–210.

Teletchea, F., Laudet, V., Hänni, C., 2006. Phylogeny of the Gadidae (sensu Svetovidov, 1948) based on their morphology and two mitochondrial genes. Mol. Phylogenet. Evol. 38 (1), 189–199 (Jan 1).

Tørresen, O.K., Star, B., Jentoft, S., Reinar, W.B., Grove, H., Miller, J.R., Walenz, B.P., Knight, J., Ekholm, J.M., Peluso, P., 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics 18, 95. http://dx.doi.org/10.1186/s12864-016-3448-x.

Zhang, X., Zhang, Y., Scheuring, C., Zhang, H.B., Huan, P., Wang, B., Liu, C., Li, F., Liu, B., Xiang, J., 2010. Construction and characterization of a bacterial artificial chromosome (BAC) library of Pacific white shrimp, *Litopenaeus vannamei*. Mar. Biotechnol. 12, 141–149. http://dx.doi.org/10.1007/s10126-009-9209-y.

Zhuang, X., 2014. Creating Sense From Non-Sense DNA: De Novo Genesis and Evolutionary History of Antifreeze Glycoprotein Gene in Northern Cod Fishes (Gadidae). University of Illinois at Urbana-Champaign.

Zhuang, X., Yang, C., Fevolden, S.-E., Cheng, C.C., 2012. Protein genes in repetitive sequence—antifreeze glycoproteins in Atlantic cod genome. BMC Genomics 13, 293. http://dx.doi.org/10.1186/1471-2164-13-293.