SCIENTIFIC REPORTS

OPEN

Characterization of a male specific region containing a candidate sex determining gene in Atlantic cod

Received: 13 August 2018
Accepted: 25 November 2018
Published online: 15 January 2019

Tina Graceline Kirubakaran[1], Øivind Andersen[1,2], Maria Cristina De Rosa[3], Terese Andersstuen[1], Kristina Hallan[1], Matthew Peter Kent[1] & Sigbjørn Lien[1]

The genetic mechanisms determining sex in teleost fishes are highly variable and the master sex determining gene has only been identified in few species. Here we characterize a male-specific region of 9 kb on linkage group 11 in Atlantic cod (*Gadus morhua*) harboring a single gene named *zkY* for zinc knuckle on the Y chromosome. Diagnostic PCR test of phenotypically sexed males and females confirm the sex-specific nature of the Y-sequence. We identified twelve highly similar autosomal gene copies of *zkY*, of which eight code for proteins containing the zinc knuckle motif. 3D modeling suggests that the amino acid changes observed in six copies might influence the putative RNA-binding specificity. Cod zkY and the autosomal proteins zk1 and zk2 possess an identical zinc knuckle structure, but only the Y-specific gene *zkY* was expressed at high levels in the developing larvae before the onset of sex differentiation. Collectively these data suggest *zkY* as a candidate master masculinization gene in Atlantic cod. PCR amplification of Y-sequences in Arctic cod (*Arctogadus glacialis*) and Greenland cod (*Gadus macrocephalus ogac*) suggests that the male-specific region emerged in codfishes more than 7.5 million years ago.

The origin and evolution of sex chromosomes from autosomes, and the mechanism of sex determination have long been subjects of interest to biologists. In eutherian mammals and birds, the sex chromosomes are highly dimorphic and have degenerated with extensive gene losses[1–3]. However in teleost fishes, cytogenetically different sex chromosomes are found in less than 10% of the species examined[4,5], and their recent origin in several lineages makes them good models to study the early stages of divergence. The frequent turnover of sex chromosomes seems to be associated with the variety of sex determining genes, even in closely related species[4–6]. For example, in medaka fishes (genus *Oryzias*), the Y-chromosomal *dmY* or *dmrt1bY* copy determines the sex in *O. latipes* and *O. curvinotus*, while testicular differentiation in *O. dancena* and in *O. luzonensis* is initiated by male-specific regulatory elements upstream of *sox3* and *gsdf* (gonadal soma-derived factor), respectively[7–10]. The various mechanisms of genetic sex determination in fish range from sex–specific alleles of the anti-Müllerian hormone (Amh) in Nile tilapia (*Oreochromis niloticus*) and the Amh type 2 receptor (*Amhr2*) in tiger pufferfish (*Takifugu rubripes*)[11,12] to complex polygenic regulation involving several genomic regions as shown in the cichlid *Astatotilapia burtoni*[13]. The diversity of mechanisms is further increased by the insertion of X- and Y- sequences modulating the expression of the neighboring master sex determinant gene, as found in the sablefish (*Anoplopoma fimbria*)[14].

Sexual differentiation in vertebrates is initially marked by highly increased proliferation of the germ cells in the presumptive ovaries compared to that in testes[6,15–19]. The time course of the sex-dimorphic germ cell proliferation varies greatly among teleost species and occurs around hatching in medaka[20], at the start-feeding stage in Atlantic cod[21] and in late juvenile stages in sea bass (*Dicentrarchus labrax*)[22]. Male germ cell divisions are inhibited by Dmy in the medaka *O. latipes*[23], and the sex determinants Gdsf, Amh and Amhr2 of the TGFβ signaling pathway might play similar roles in inducing sex differentiation in other species[6]. However, the requirement of germ cells for gonadal development appears to vary among teleost species[24–26] reflecting that sex determination might be triggered by alternative mechanisms. Most sex determinants found in vertebrates are thought to have

[1]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Norwegian University of Life Sciences (NMBU), PO Box 5003, 1433, Ås, Norway. [2]Nofima, PO Box 5010, N-1430, Ås, Norway. [3]Institute of Chemistry of Molecular Recognition - CNR c/o Institute of Biochemistry and Clinical Biochemistry, Catholic University of Rome, 00168, Rome, Italy. Matthew Peter Kent and Sigbjørn Lien jointly supervised this work. Correspondence and requests for materials should be addressed to M.P.K. (email: matthew.peter.kent@nmbu.no) or S.L. (email: sigbjorn.lien@nmbu.no)

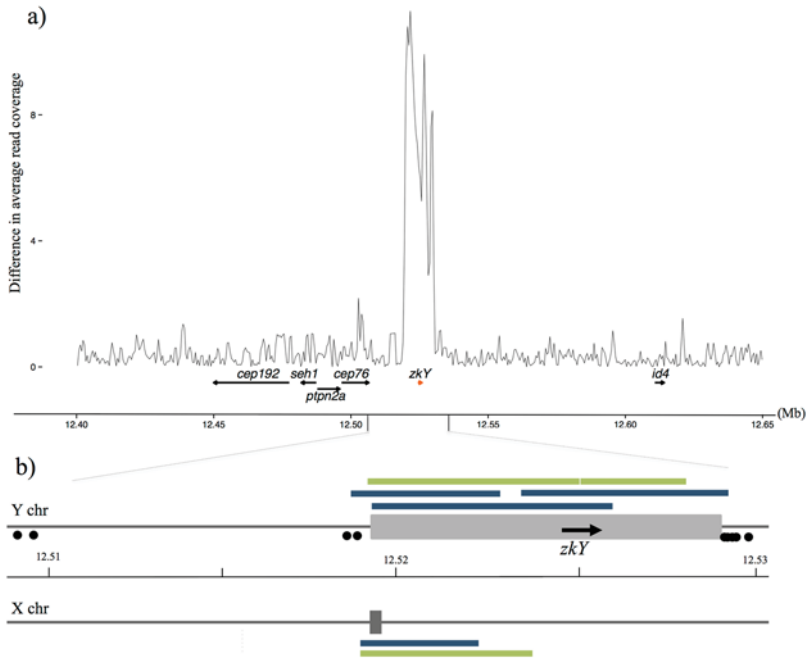

Figure 1. Genomic organization of the Y- and X-sequences on LG11 in Atlantic cod. (**a**) Comparison of read depth between male (n = 49) and female (n = 53) samples reveals an excess of Y-specific reads across a 9 kb interval on LG11 in the final and public gadMor2.1 assembly. Positions of cod *zkY* and the neighboring genes are indicated. (**b**) The Y- and X-sequences (grey boxes), including the Y-specific *zkY* gene, and the nine polymorphisms heterozygous in all males and homozygous in all females (black dots). Long reads confirming assembly integrity were generated using Oxford Nanopore (green) and Pacific Biosciences (blue) sequencing technologies.

acquired this role by being recruited from conserved downstream regulators of gonadal differentiation, although the function of some actors differs among lineages[6,27–29]. An exception is the novel salmonid sdY; a truncated male-specific copy of the interferon regulatory factor 9 (irf9), which is an immune-related gene not associated with sex[30]. Knowledge about how genes are functionalized and incorporated at the top of the sex regulatory cascade should broaden our understanding of the genetic regulation of sex determination and elucidate whether there are constraints on the types of genes that can be co-opted as master control switches.

Atlantic cod is an economically important cold-water marine species widely distributed in the North Atlantic Ocean. Due to the decline in many cod stocks, cod farming has attracted interest but production is hampered by precocious early sexual maturation, particularly in males. This could partly be solved by production of all-female triploid stocks[4,31,32]. Gynogenetic and sex-reversed cod populations have demonstrated a XX-XY sex determination system[21,33,34], but karyotyping of Atlantic cod and the closely related European hake (*Merluccius merluccius*) failed to reveal sex-linked chromosome heteromorphism[35,36]. Recently, whole genome sequence data from wild Atlantic cod was used to identify genotypes segregating closely with a XX-XY system in six putative sex determining regions distributed across five linkage groups[37]. Here we use whole genome re-sequencing data from 49 males and 53 females, together with long-read sequence data and Sanger sequencing of targeted PCR products, to characterize a Y-sequence of 9,149 base pairs on LG11 harboring a single gene which we have named zinc knuckle on the Y chromosome gene *(zkY)*. Gene expression data from early development stages and modeling of the zinc knuckle structure offer circumstantial evidence consistent with a function in Atlantic cod sex determination.

Results and Discussion

Identification of male specific sequence on linkage group 11. Two independent approaches were used to identify a male specific genomic region in Atlantic cod. Firstly, we searched whole genome sequence data from males and females for SNPs that segregate according to an XX-XY system. Illumina short-reads (approximately 10X coverage per individual) generated from 49 males and 53 females were mapped to an initial in-house developed gadMor2.1 genome assembly, followed by variant detection. When applying stringent criteria demanding that gender specific variants should be heterozygous in all 49 males and homozygous in all 53 females, we detected 9 variants all distributed within a 15 kb region on LG11 (Supplementary File 1 and Fig. 1b). BLAST alignments revealed that these variants fell inside the 55 kb region previously reported as showing most evidence for being involved in sex determination by Star *et al.*[37]. Secondly, we analyzed resequencing data to identify positions in the genome that displayed significant differences in read depth between males and females. Two regions showed an almost complete absence of female reads while displaying >500 reads from males, i.e. Y-specific characteristics. These regions were separated by an intervening sequence displaying roughly 2-times coverage in females versus males, i.e. a possible X-sequence. Both the Y- and X-sequence regions fell within an interval flanked by the nine sex-linked markers on LG11 lending further support to the significance of this region in sex

determination. Closer examination of the architecture of this region using IGV[38] revealed that despite a high read-depth, there was no evidence of single reads bridging, or read-pairs spanning, these differential read-depth sub-regions. Collectively these anomalies are suggestive of assembly errors in the initial gadMor2.1 assembly, and a hybridization of X- and Y-sequences, and underscores the importance of developing high-quality reference genomes.

In order to resolve this complex region we performed nested, long-range PCR to generate a Y-specific fragment, which was then sequenced using an Illumina MiSeq. Assembly construction using Gap Filler generated a 6 kb sequence, which was integrated into the initial gadMor2.1 assembly to produce a complete Y-sequence of 9,149 bp (Fig. 1 and Supplementary File 2). The sequencing revealed an imperfect repeated element positioned close to the end of the Y-sequence (see underlined sequence in Supplementary File 2). An X-sequence of 425 bp was constructed by Sanger sequencing PCR fragments generated from females using primers flanking the Y-sequence. To confirm the integrity of these manually curated X- and Y-sequences we aligned publicly available long-read data (Pacific BioSciences) derived from the male used to generate the reference (accession numbers at ENA (http://www.ebi.ac.uk/ena), ERX1787826-ERX1787972) and long reads generated in-house using an Oxford Nanopore MinION device, also from a male. Although overall coverage was low, we were able to identify long-reads that aligned specifically to either the Y- or X-sequences and, in combination, spanned their entire lengths (Supplementary File 3 and Fig. 1b).

The Y-sequence contains a single gene. X- and Y-sequences of the public gadMor2.1 assembly were annotated using standard gene-model predictive software and RNA-Seq data produced from early larval stages together with publicly available RNA-Seq data from Atlantic cod (See Material and Methods). This revealed a single gene (which we have named *zkY*) in the Y-sequence (Supplementary File 2 and Fig. 1), which is inserted in a synteny block that is highly conserved in teleosts and in spotted gar (*Lepisosteus oculatus*) (Supplementary Figure 1). The intronless cod *zkY* codes for a zinc knuckle protein characterized by the zinc knuckle consensus sequence Cys-X2-Cys-X4-His-X4-Cys (X = any amino acid). This motif is mainly found in the RNA-binding retroviral nucleocapsid (NC) proteins, but also in various eukaryotic proteins binding single-stranded nucleotide targets[39–42]. The flanking basic residues bind nucleic acids non-specifically through electrostatic interactions with the phosphodiester backbone of nucleic acids[43].

Zinc knuckle proteins are members of the large family of zinc finger proteins possessing a versatility of tetrahedral Cys- and His-containing motifs that bind to DNA and RNA target sites[44,45]. The DNA-binding domain of the DMRT transcription factors, including the sex determinant DmY in medaka, consists of intertwined CCHC and HCCC motifs binding in the minor groove of DNA[46]. The zinc finger protein ZFAND3 is essential for spermatogenesis in mice and the polymorphic tilapia *zfand3* was recently mapped in the sex determining locus[47,48]. Hence, the recruitment of zinc finger proteins as the sex determinant seems to have occurred several times in teleosts.

Multiple autosomal copies of cod zkY. The sex determining gene in several teleosts exhibits an autosomal copy that differs from the male-specific gene in spatio-temporal expression patterns and/or functionality[28,49–52]. We identified 12 autosomal copies of cod *zkY* mapping to seven different linkage groups and one unassembled scaffold. Premature stop codons and single base indels were found in four of the genes encoding putatively non-functional proteins lacking the zinc knuckle domain. The remaining eight genes code for zinc knuckle proteins named zk1 to zk8, which differ in length from 143 to 633 amino acids (Supplementary File 4). Sequence alignment revealed several amino acid substitutions in the zinc knuckle domain in addition to the variable size of an imperfect repeat of basic residues preceding the domain (Fig. 2A, Supplementary File 5 and Supplementary Figure 2).

Functional implications of the amino acid substitutions in the zinc knuckle domain of the autosomal proteins was inferred by exploring the predicted 3D structure and the interactions with a putative RNA oligonucleotide target. The suitability of the d(ACGCC) sequence in the structure modeling was supported by the stacking of the conserved Trp442 between the two bases G3 and C4 in agreement with the interactions between specific base moieties and Trp at the corresponding position in NC proteins[53–55]. The three cysteines Cys443, Cys436 and Cys446 together with His441 coordinate the tetrahedral binding of the Zn^{2+} ion in the modelled cod zinc knuckles. While the replacement of Met433 in zkY, zk1 and zk2 with the Ile residue in the four zk3-zk7 proteins maintains the hydrophobic interactions with Cys446, the Ile433 residue is predicted to interact with the G3 base of the pentanucleotide (Fig. 2C). The RNA-binding specificity of human HIV-1 was consistently altered by the Met- > Val and Met- > Lys substitutions in the same position of the second zinc knuckle[56]. The Asn435Lys change in the zk8 protein may affect the tetrahedral Zn^{2+} ion coordination by interacting with Cys446, while Gln443 is predicted to form a hydrogen bond with the phosphate group connecting A1 and C2 (Fig. 2D). Additionally, electrostatic interactions are predicted between Arg437 and His434, which are moved away from its C2 contacts observed in the other variants.

Altogether, the predicted alterations in the contacts between the replaced amino acids and the interactions with the oligonucleotide sequence template suggest functional differences in the putative RNA-binding activity of the zinc knuckle proteins examined. In addition, the functional role played by the flanking basic residues in the non-specific binding of nucleic acids[44] suggests that these interactions might be altered by the extended flanking repeat in the zk6 and zk7 proteins. Based on the identical zinc knuckle domain and the similar flanking sequences in the three proteins zkY, zk1 and zk2, we decided to compare the larval expression of these coding genes.

Increased transcription of cod *zkY* prior to onset of sex differentiation. A prerequisite for being a sex determining gene is to show pronounced transcription levels prior to onset of sex differentiation, which in Atlantic cod likely occurs at start feeding around 35 days post hatching (dph) based on the high number of germ

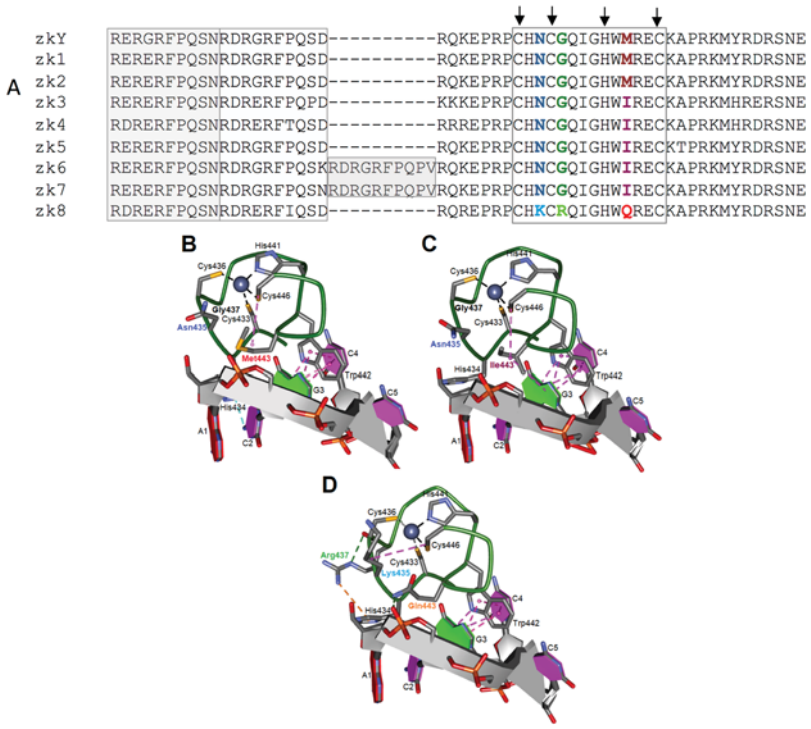

Figure 2. Amino acid substitutions in the zinc knuckle domain in cod zkY and its autosomal copies. (**A**) Sequence alignment of the zinc knuckle domain (boxed) and flanking regions in cod zkY and the eight autosomal protein variants. The characteristic Cys-Cys-His-Cys residues of the zinc knuckle are arrowed, substituted amino acids are indicated by colors and correspond to the labelled positions 435, 437, and 443 in (**B**-**D**). Repeated segments adjacent to the zinc knuckle are shaded. (**B**–**D**) Modeled structure of the three different zinc knuckle domains of cod zkY and the autosomal proteins. The Zn^{2+} ion is displayed together with the oligonucleotide d(ACGCC) template (see Methods). (**B**) Met443 variant of zkY, zk1, zk2, (**C**) Ile443 variant of zk3-zk7, (**D**) Lys435-Arg437-Gln443 variant of zk8. Hydrogen bonds are indicated by dotted green lines, Pi-donor hydrogen bonds in light blue, hydrophobic interactions by dotted magenta lines, and electrostatic bonds by dotted orange lines.

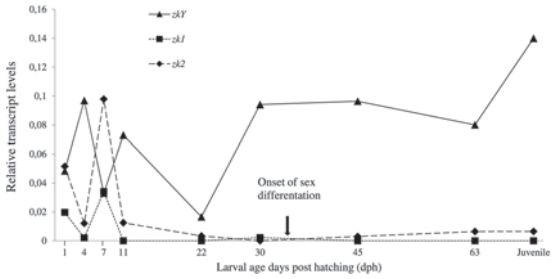

Figure 3. Larval expression of *zkY, zk1* and *zk2* from hatching to juvenile stage. Expression levels are given as relative transcript read numbers. The onset of sex differentiation is indicated[21].

cells in females compared to males together with the female-biased expression of *cyp19a1a*[21,57]. RNA-Seq analysis of cod *zkY* and the autosomal *zk1* and *zk2* copies showed that the three genes were expressed at variable levels from hatching until 22 dph at which stage the transcription levels of *zkY* increased and stabilized at relatively high levels from 30 dph onwards (Fig. 3). In contrast, the larval expression of *zk1* and *zk2* was almost undetectable or very low from 11 dph.

The increased larval expression of the male-specific *zkY* gene prior to the onset of sex differentiation is consistent with a possible function in sex determination. Multiple germline-specific RNA regulatory proteins are essential for germ cell specification, maintenance, migration and proliferation in various organisms[58]. While this regulatory network seems to be evolutionary conserved in nematodes, flies and mammals, the key role played by different sex determinants in the control of male germ cell proliferation in fish suggests the involvement of lineage-specific regulators[6]. Knockdown of the germline zinc knuckle helicases *glh1-2* in *C. elegans* and the *vasa* homolog in mice resulted in male infertility[59–61]. Intriguingly, the loss of zinc knuckles in the RNA-binding Vasa

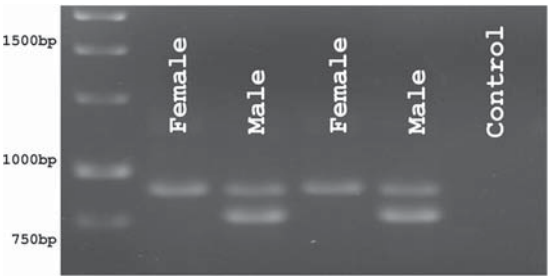

Figure 4. Agarose gel separation of PCR products from male and female Atlantic cod. (Cropped from a larger gel image presented in Supplementary Figure 4).

in vertebrates and insects was suggested to have coincided with the emergence of hitherto unknown zinc-knuckle cofactors conferring target specificity[62].

X- and Y-sequences in other gadoid species. To elucidate the emergence of this putative sex determining region in other cod fishes we scanned draft genome assemblies from 12 species within the family *Gadidae*[63] for sequences matching the Atlantic cod Y- sequence. In its entirety, the 9,149 bp male-specific region showed only partial hits suggesting that the Y- sequence is absent or incorrectly assembled in these draft assemblies. Therefore, to find evidence for the sex determination region in other codfish species we were obliged to use a PCR based approach. Although the multi-species sequence alignments were highly fragmented, a close examination revealed isolated regions within and immediately outside the male-specific region of Atlantic cod which were, to some extent, conserved across species. These regions allowed us to develop primers, which amplified a specific fragment of the Y-sequence in both Greenland cod (*Gadus macrocephalus ogac*) and Arctic cod (*Arctogadus glacialis*), in addition to Atlantic cod (see alignment in Supplementary File 6). The X-sequence was efficiently amplified and sequenced in Greenland cod, Arctic cod, polar cod (*Boreogadus saida*), haddock (*Melanogrammes aeglefinus*) and burbot (*Lota lota*) (see alignment in Supplementary File 7). This result documents that the Y-specific sequence was present prior to the separation of Arctic cod from Atlantic cod more than 7.5 MYA[63,64] while the presence of X-sequence in burbot suggests that this arrangement predates the *Lotinae* – *Gadinae* split 45 MYA (Supplementary Figure 3).

A diagnostic test for distinguishing male and female Atlantic cod is potentially useful for researchers and the emerging aquaculture industry. To efficiently determine gender, we developed a simple PCR reaction including two different forward primers (one Y-sequence specific and the other matching a common sequence upstream of the X- and Y- regions) and a common reverse primer (annealing to a sequence downstream of the X- and Y-regions) resulting in a single band in females and double in males (Fig. 4).

Sex determination is a complex process involving the coordinated actions of multiple factors, and a greater understanding of what genes are involved in the sex determination cascade is needed. Furthermore, this understanding must be developed in multiple species to learn whether these processes are species or taxa specific and to improve our knowledge of the evolution of sex determination. The main focus of this paper has been to improve our knowledge of the genomic basis for sex determination in Atlantic cod. Our results indicate that the *zkY* gene present within the male specific region on LG11 is involved in Atlantic cod sex determination, and documents the presence of X- and Y- sequences in other gadoids.

Materials and Methods

Ethics statement. To produce whole genome sequence data, fin clips were collected non-destructively in strict accordance with the welfare rules given by the National Animal Research Authority (NARA). Samples were collected as a part of an ongoing, long term project authorized and managed by Nofima AS which seeks to maintain a biobank of samples from parents of the National cod breeding program nucleus. For RNA sequencing (larvae 1–35 dph), permission for sampling biological material is not required for fish eggs and larvae collected before start-feeding according to NARA.

Sampling and Illumina whole-genome sequencing. Tissue was collected from 49 males and 53 females from the National Atlantic cod breeding program maintained by Nofima, Tromsø, Norway (available at https://www.ebi.ac.uk/ena, PRJEB12803 and Supplementary File 8). The fish were parents in year classes 2005 or 2006 and represented the second generation of cod produced in captivity. The original broodstock in the base population were sampled from different geographical areas along the Norwegian coast[65]. DNA was extracted using the DNeasy kit manufactured by QIAGEN (Germany) according to manufacturer's instructions. Libraries were prepared using the Truseq Library prep kit (Illumina, San Diego, USA), and sequenced using an Illumina HiSeq 2500 instrument. A total of 13.1 billion paired-reads (2×100 PE) were produced, averaging 129 million paired reads per individual with coverage from 12.4 to 19.0X.

Construction of the initial gadMor2.1 assembly and variant detection. The initial gadMor2.1 assembly was constructed by integrating a dense linkage map with two public draft assemblies (NEWB454 and CA454ILM https://figshare.com/articles/Transcript_and_genome_assemblies_of_Atlantic_cod/3408247) generated from sequencing the same male. The assemblies were constructed using different combinations of short-read sequencing data and different assembly programs (Newbler and Celera respectively) and displayed

different qualities with NEWB454 having longer scaffold N50s and more gaps than CA454ILM[66]. The linkage map containing 9354 SNP markers was produced after genotyping a pedigree of 2951 fishes. Chimeric scaffolds were broken at junctions between contigs containing SNPs from different LGs. Overlapping scaffolds were identified by comparing SNPs mapping to both assemblies and were merged using coordinates from alignment with LASTZ[67] generating scaffolds that were used to build the final chromosome sequences. Finally, all scaffolds were oriented, ordered and concatenated into a new chromosome sequence based on information from the linkage map.

Variants were detected by first filtering raw reads from each individual using Trimmomatic v0.32[68], and subsequently aligning reads to the unmasked initial gadMor2.1 assembly using Bowtie2 v2.3.2 with the parameter 'sensitive'[69]. The resulting BAM files were merged by gender using Sambamba v0.6.5[70] and GATK HaplotypeCaller v3.8[71] was used to call variants with the following parameters: gt_mode DISCOVERY, minPruning 3. A Perl script was used to report SNPs, and their positions, that display heterozygous genotypes in all males and homozygous genotypes in all females.

Read depth differences between males and females. For each sex, quality filtered reads from males (n = 49) and females (n = 53) were combined to generate two gender-specific bed files. Bedtools genomecov version 2.22.0[72] was then used to count read depth at each position in the public gadMor2.1 assembly. A Perl script was then used to calculate the average read depth per individual across successive 1000 bp intervals with an overlap of 500 bp (positions with 0 coverage in males and females were ignored). For each average, the absolute difference between the larger and smaller number was calculated and plotted. Across the genome, the region with the most extreme read depth differences was seen on LG11 (Fig. 1a).

Construction of Y- and X-sequences. A nested PCR strategy was used to amplify the Y-specific sequence lacking from the initial gadMor2.1 assembly. The initial reaction used the following primers and conditions, (Fwd; 5′-CCTAAACCACAGTCCTGGGC-3′, Rev; 5′-ACATTGTGCACACACATTGTATC-3′, PrimeSTAR GXL DNA Polymerase (Takara, Japan), cycling 30 times with 10 s at 98 °C, 15 s at 57 °C, 3 min at 68 °C, final extension 72 °C for 10 min. PCR-product was used as template in a subsequent reaction using the same conditions but with the nested PCR primers (Fwd; 5′-CTCTGTAGTTTGTGGTGGGGT-3′, Rev; 5′-ACATGACGAATGGCCTCCTTT-3′). The resulting PCR product was purified using a QIAquick PCR purification kit from QIAGEN (Germany) and quantified. DNA was prepared for sequencing using a Nextera XT kit from Illumina (USA) according to manufacturer's instructions and the resulting library sequenced using 2 × 250 nt sequencing chemistry using a MiSeq (Illumina, USA). After quality trimming the resulting reads were anchored to the existing flanking Y-sequences using Gap filler[73]. The resulting sequence contained a single gap that was filled by Sanger sequencing a PCR product generated using the following PCR primers and conditions (Fwd; 5′-ACACAACGCAGAGTCTGTCC-3′, Rev: 5′-TCAGCTAGTCTCGCAATGGC-3′, denaturation 15 min at 95 °C, then cycling 30 times with 10 s at 98 °C, 15 s at 98 °C, 60 s at 68 °C, final extension 10 min at 72 °C). An X-sequence was constructed from Sanger sequencing a PCR product produced using primers annealing up- and down-stream of the Y-sequence.

Generation X- and Y-sequences in other gadoids. Sequence alignments of male specific region in Atlantic cod with public assembly data from 12 gadid species[63], using MUMmer version 3.23[74], identified four rather short regions that showed good conservation across species, including two regions within the Y-sequence (B; 12,520,541–12,520,838 and C; 12,527,782–12,5280,27) and two regions (A; 12,518,827–12,519,231 and D; 12,528,480–12,529,285) immediately flanking the X- and Y-sequences. Primers were designed using Primer3[75] to amplify from within the Y-sequence conserved sequence C to flanking sequence D (i.e. a Y-specific PCR) and from flanking sequence A to D (an X-sequence specific PCR). Products from these reactions were Sanger sequenced and aligned using MAFFT v7.402[76] (Supplementary Files 6 and 7).

Diagnostic sex-test. A sex distinguishing duplex PCR was designed for Atlantic cod using the following primers and conditions (Fwd A; 5′-ACACACGGTCTGCTGTAGTG-3′, Fwd C; 5′-GGAGGGGAATTGTACAAACACG-3′, Rev D; 5′-GTGTGCCAAATGGATGCCAA-3′), denaturation for 15 min at 95 °C, cycling 35 times with 30 s at 94 °C, 60 s at 55 °C, 60 s at 72 °C, final extension 72 °C for 10 min (Supplementary File 9).

Nanopore Sequencing. High-molecular weight DNA was extracted using a phenol-chloroform method[77]. Sequencing libraries were prepared using SQK-RAD003 and SQK-LSK108 kits and protocols from Oxford Nanopore Technology (Oxford, UK) and sequenced on a MinION device generating a combined total of 3.9 Gb sequence data with an N50 read length of 4.2 kb. Reads were aligned to the public gadMor2.1 assembly using GraphMap aligner v0.5.2[78].

Transcript profiling of cod larvae. Cod larvae were hatched at the National Atlantic Cod Breeding Centre in Tromsø, Norway, after the incubation of fertilized eggs in seawater rearing tanks at 4.5 °C. For 1 and 7 dph larvae, RNA was extracted from a pool of 10 individuals using the QIAGEN AllPrep DNA/RNA/miRNA Universal kit (QIAGEN; Germany). Samples were prepared for sequencing using TruSeq Stranded mRNA kit from Illumina (USA) and sequenced using an Illumina HiSeq 2000 to produce 362 million reads. For the samples 12 and 35 dph samples, RNA was extracted using an RNeasy kit (QIAGEN, Germany) prepared for sequencing using TruSeq Stranded mRNA kit from Illumina (USA) and sequenced using an Illumina MiSeq (2 × 250nt) to produce 4.4 million reads. All reads were trimmed using Trimmomatic v0.32[68] before further analysis.

Total available short read data (http://www.ebi.ac.uk/ena, PRJEB18628; and https://www.ncbi.nlm.nih.gov/sra/?term = SRP056073) was binned based on days before hatching (dph) before being aligned to the public gadMor2.1 assembly using star aligner STAR v2.3.1z12 as described previously. Potential transcripts were constructed

using stringtie v1.3.3[79], while cuffmerge v2.2.1[80] was used to produce a GTF file containing key metrics. FPKM values for zkY, zky1, zky2, were calculated using only those reads with a mapping quality of ≥30.

Gene annotation. Data from various public sources was used to build gene models including (i) 3M transcriptome reads generated using GS-FLX 454 technology and hosted at NCBI's SRA (https://www.ncbi.nlm.nih.gov/sra/?term = SRP013269), (ii) >250 K ESTs hosted by NCBI (https://www.ncbi.nlm.nih.gov/nucest) (iii) 4.4 M paired-end mRNA MiSeq sequences from whole NEAC larvae at 12 and 35 dph (https://www.ebi.ac.uk/ena, PRJEB25591), (iv) Pacbio reads from (https://www.ebi.ac.uk/ena, PRJEB18628), (v) 362 M Illumina reads from 1 and 7 dph (https://www.ebi.ac.uk/ena, PRJEB25591) and (vi) approximately 1.7B Illumina reads from 4–63 dph as well as juvenile samples (https://www.ncbi.nlm.nih.gov/sra/?term = SRP056073). To enable model building, short Illumina reads (<250nt) were mapped to the public gadMor2.1 assembly using STAR v2.3.1z12. Because Illumina reads from 4 dph - juvenile were generated using an unstranded library, the parameter' outSAMstrandField intronMotif' was used in alignment. Long reads from PacBio were mapped using STARlong v2.5.2a[81] while 454 transcriptome reads were mapped using gmap v 2014-07-28[82] with '–no-chimeras' parameter in addition to default parameters. Cufflinks v2.2.1[80] was used to assemble the reads into transcript models for all alignments except for data from 4 dph - juvenile stage samples where stringtie v1.3.3[79] was used. Transcript models were merged using cuffmerge v2.2.1[80].

Gene models were tested by performing (i) open reading frame (ORF) prediction using TransDecoder[83] using both pfamA and pfamB databases for homology searches and a minimum length of 30 amino acids for ORFs without pfam support, and (ii) BLASTP analysis (evalue <1e-10) for all predicted proteins against zebrafish (*Danio rerio*) (v9.75) and three-spined stickleback (*Gasterosteus aculeatus*) (BROADS1.75) annotations from Ensembl. Only gene models with support from at least one of these homology searches were retained. Functional annotation of the predicted transcripts was done using blastx against the SwissProt database.

Modeling the zinc knuckle domain. The three-dimensional structure of the three different variants of the zinc knuckle domain were built using a threading approach which combines three-dimensional fold recognition by sequence alignment with template crystal structures and model structure refining. The query-template alignment was generated by HHPRED (https://toolkit.tuebingen.mpg.de) and then submitted to the program MODELLER[84] as implemented in Discovery Studio (Dassault Systèmes BIOVIA). Query coverage and e-value score were considered to define a suitable template structure. The structure of nucleocapsid protein NCp10 of retrovirus MoMuLV, which contains a single Cys-X2-Cys-X4-His-X4-Cys zinc knuckle domain bound to the oligonucleotide d(ACGCC) was selected as template (https://www.rcsb.org/structure/1a6b). Zinc ions and oligonucleotides were explicitly considered through molecular modeling steps. Fifty models, optimized by a short simulated annealing refinement protocol available in MODELLER, were generated and their consistency was evaluated on the basis of the probability density function violations provided by the program. Stereochemistry of selected models was checked using the program PROCHECK[85]. Visualization and manipulation of molecular images were performed with Discovery Studio (Dassault Systèmes BIOVIA).

Data Availability

The gadMor2.1 genome assembly, long-range PCR MiSeq Illumina data are available at https://figshare.com/s/313f8fe1fdcc82571a99. The 102 individual samples read data are available at ENA, with the study accession number PRJEB12803. The MiSeq whole NEAC larvae at 12,35 dph and 1,7 dph illumina samples are available at ENA, with the study accession number PRJEB25591. All data generated or analyzed during this study are included in this published article and its Supplementary Information files.

References

1. Graves, J. A. Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901–914 (2006).
2. Bellott, D. W. *et al.* Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–U613 (2010).
3. Peichel, C. L. Convergence and divergence in sex-chromosome evolution. *Nat Genet* **49**, 321–322 (2017).
4. Devlin, R. H. & Nagahama, Y. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* **208**, 191–364 (2002).
5. Volff, J. N., Nanda, I., Schmid, M. & Schartl, M. Governing sex determination in fish: Regulatory putsches and ephemeral dictators. *Sex Dev* **1**, 85–99 (2007).
6. Kikuchi, K. & Hamaguchi, S. Novel sex-determining genes in fish and sex chromosome evolution. *Dev Dyn* **242**, 339–353 (2013).
7. Nanda, I. *et al.* A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka. *Oryzias latipes. PNAS* **99**, 11778–11783 (2002).
8. Matsuda, M. *et al.* DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417**, 559–563 (2002).
9. Myosho, T. *et al.* Tracing the emergence of a novel sex-determining gene in medaka. *Oryzias luzonensis. Genetics* **191**, 163–170 (2012).
10. Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K. & Sakaizumi, M. Geographic variation and diversity of the cytochrome b gene in Japanese wild populations of medaka. *Oryzias latipes. Zoolog Sci* **20**, 1279–1291 (2003).
11. Kamiya, T. *et al.* A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *Plos Genet* **8**, e1002798 (2012).
12. Li, M. *et al.* A tandem duplicate of anti-mullerian hormone with a missense SNP on the Y chromosome is essential for male sex determination in nile tilapia. *Oreochromis niloticus. Plos Genet* **11**, e1005678 (2015).
13. Roberts, N. B. *et al.* Polygenic sex determination in the cichlid fish *Astatotilapia burtoni. BMC Genomics* **17**, 835 (2016).
14. Rondeau, E. B. *et al.* Genomics of sablefish (*Anoplopoma fimbria*): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. *BMC Genomics* **14**, 452 (2013).
15. Nakamura, Y. *et al.* Migration and proliferation of primordial germ cells in the early chicken embryo. *Poult Sci* **86**, 2182–2193 (2007).
16. Zust, B. & Dixon, K. E. Events in germ-cell lineage after entry of primordial germ-cells into genital ridges in normal and uv-irradiated xenopus-laevis. *J Embryol Exp Morph* **41**, 33–46 (1977).

17. Nakamura, M., Kobayashi, T., Chang, X. T. & Nagahama, Y. Gonadal sex differentiation in teleost fish. *J Exp Zool* **281**, 362–372 (1998).
18. Siegfried, K. R. & Nusslein-Volhard, C. Germ line control of female sex determination in zebrafish. *Developmental Biology* **324**, 277–287 (2008).
19. Saito, D. *et al*. Proliferation of germ cells during gonadal sex differentiation in medaka: Insights from germ cell-depleted mutant zenzai. *Developmental Biology* **310**, 280–290 (2007).
20. Kobayashi, T. *et al*. Two DM domain genes, DMY and DMRT1, involved in testicular differentiation and development in the medaka. *Oryzias latipes. Dev Dyn* **231**, 518–526 (2004).
21. Haugen, T. *et al*. Sex differentiation in Atlantic cod (*Gadus morhua* L.): morphological and gene expression studies. *Reprod Biol Endocrin* **10**, 47 (2012).
22. Saillant, E. *et al*. Sexual differentiation and juvenile intersexuality in the European sea bass (*Dicentrarchus labrax*). *J Zool* **260**, 53–63 (2003).
23. Herpin, A. *et al*. Inhibition of primordial germ cell proliferation by the medaka male determining gene Dmrt I bY. *BMC Dev Biol* **7**, 99 (2007).
24. Fujimoto, T. *et al*. Sexual dimorphism of gonadal structure and gene expression in germ cell-deficient loach, a teleost fish. *PNAS* **107**, 17211–17216 (2010).
25. Wargelius, A. *et al*. Dnd knockout ablates germ cells and demonstrates germ cell independent sex differentiation in Atlantic salmon. *Sci Rep-Uk* **6**, 21284 (2016).
26. Goto, R. *et al*. Germ cells are not the primary factor for sexual fate determination in goldfish. *Dev Biol* **370**, 98–109 (2012).
27. Cutting, A., Chue, J. & Smith, C. A. Just how conserved is vertebrate sex determination? *Dev Dyn* **242**, 380–387 (2013).
28. Matsuda, M. & Sakaizumi, M. Evolution of the sex-determining gene in the teleostean genus. *Oryzias. Gen Comp Endocr* **239**, 80–88 (2016).
29. Herpin, A. & Schartl, M. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO Rep* **16**, 1260–1274 (2015).
30. Yano, A. *et al*. The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evol Appl* **6**, 486–496 (2013).
31. Pandian, T. J. & Koteeswaran, R. Ploidy induction and sex control in fish. *Hydrobiologia* **384**, 167–243 (1998).
32. Penman, D. J. & Piferrer, F. Fish gonadogenesis. part I: Genetic and environmental mechanisms of sex determination. *Rev Fish Sci* **16**, 16–34 (2008).
33. Ottera, H. *et al*. Induction of meiotic gynogenesis in Atlantic cod, *Gadus morhua* (L.). *J Appl Ichthyol* **27**, 1298–1302 (2011).
34. Whitehead, J. A., Benfey, T. J. & Martin-Robichaud, D. J. Ovarian development and sex ratio of gynogenetic Atlantic cod (*Gadus morhua*). *Aquaculture* **324**, 174–181 (2012).
35. Ghigliotti, L. *et al*. Karyotyping and cytogenetic mapping of Atlantic cod (*Gadus morhua* Linnaeus, 1758). *Anim Genet* **43**, 746–752 (2012).
36. Garcia-Souto, D., Troncoso, T., Perez, M. & Pasantes, J. J. Molecular Cytogenetic Analysis of the European Hake *Merluccius merluccius* (Merlucciidae, Gadiformes): U1 and U2 snRNA Gene Clusters Map to the Same Location. *Plos One* **10**, e0146150 (2015).
37. Star, B. *et al*. Genomic characterization of the Atlantic cod sex-locus. *Sci Rep-Uk* **6**, 31235 (2016).
38. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013).
39. Summers, M. F. *et al*. Nucleocapsid zinc fingers detected in retroviruses: EXAFS studies of intact viruses and the solution-state structure of the nucleocapsid protein from HIV-1. *Protein Sci* **1**, 563–574 (1992).
40. Williams, M. C., Gorelick, R. J. & Musier-Forsyth, K. Specific zinc-finger architecture required for HIV-1 nucleocapsid protein's nucleic acid chaperone function. *PNAS* **99**, 8614–8619 (2002).
41. Guerrerio, A. L. & Berg, J. M. Design of single-stranded nucleic acid binding peptides based on nucleocapsid CCHC-box zinc-binding domains. *J Am Chem Soc* **132**, 9638–9643 (2010).
42. Benhalevy, D. *et al*. The human CCHC-type zinc finger nucleic acid-binding protein binds G-rich elements in target mRNA coding sequences and promotes translation. *Cell Rep* **18**, 2979–2990 (2017).
43. Mitra, M. *et al*. The N-terminal zinc finger and flanking basic domains represent the minimal region of the human immunodeficiency virus type-1 nucleocapsid protein for targeting chaperone function. *Biochemistry* **52**, 8226–8236 (2013).
44. Michalek, J. L., Besold, A. N. & Michel, S. L. J. Cysteine and histidine shuffling: mixing and matching cysteine and histidine residues in zinc finger proteins to afford different folds and function. *Dalton T* **40**, 12619–12632 (2011).
45. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struc Biol* **11**, 39–46 (2001).
46. Zhu, L. Y. *et al*. Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Gene Dev* **14**, 1750–1764 (2000).
47. de Luis, O., Lopez-Fernandez, L. A. & del Mazo, J. Tex27, a gene containing a zinc-finger domain, is up-regulated during the haploid stages of spermatogenesis. *Exp Cell Res* **249**, 320–326 (1999).
48. Ma, K. *et al*. Charactering the ZFAND3 gene mapped in the sex-determining locus in hybrid tilapia (*Oreochromis spp*.). *Sci Rep-Uk* **6**, 25471 (2016).
49. Bradley, K. M. *et al*. An SNP-based linkage map for zebrafish reveals sex determination loci. *G3-Genes Genom Genet* **1**, 3–9 (2011).
50. Reichwald, K. *et al*. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* **163**, 1527–1538 (2015).
51. Rondeau, E. B., Laurie, C. V., Johnson, S. C. & Koop, B. F. A. PCR assay detects a male-specific duplicated copy of anti-mullerian hormone (amh) in the lingcod (*Ophiodon elongatus*). *BMC Res Notes* **9**, 230 (2016).
52. Hattori, R. S. *et al*. A Y-linked anti-mullerian hormone duplication takes over a critical role in sex determination. *PNAS* **109**, 2955–2959 (2012).
53. Meric, C. & Goff, S. P. Characterization of Moloney murine leukemia virus mutants with single-amino-acid substitutions in the Cys-His box of the nucleocapsid protein. *J Virol* **63**, 1558–1568 (1989).
54. Urbaneja, M. A., McGrath, C. F., Kane, B. P., Henderson, L. E. & Casas-Finet, J. R. Nucleic acid binding properties of the simian immunodeficiency virus nucleocapsid protein NCp8. *J Biol Chem* **275**, 10394–10404 (2000).
55. Hashimoto, H. *et al*. Crystal structure of zinc-finger domain of Nanos and its functional implications. *EMBO Rep* **11**, 848–853 (2010).
56. Mark-Danieli, M. *et al*. Single point mutations in the zinc finger motifs of the human immunodeficiency virus type 1 nucleocapsid alter RNA binding specificities of the gag protein and enhance packaging and infectivity. *J Virol* **79**, 7756–7767 (2005).
57. Johnsen, H., Tveiten, H., Torgersen, J. S. & Andersen, O. Divergent and sex-dimorphic expression of the paralogs of the Sox9-Amh-Cyp19a1 regulatory cascade in developing and adult atlantic cod (*Gadus morhua L.*). *Mol Reprod Dev* **80**, 358–370 (2013).
58. Cinalli, R. M., Rangan, P. & Lehmann, R. Germ cells are forever. *Cell* **132**, 559–562 (2008).
59. Gruidl, M. E. *et al*. Multiple potential germ-line helicases are components of the germ-line-specific P granules of *Caenorhabditis elegans*. *PNAS* **93**, 13837–13842 (1996).
60. Kuznicki, K. A. *et al*. Combinatorial RNA interference indicates GLH-4 can compensate for GLH-1; these two P granule components are critical for fertility in *C-elegans*. *Development* **127**, 2907–2916 (2000).

61. Tanaka, S. S. *et al*. The mouse homolog of Drosophila Vasa is required for the development of male germ cells. *Gene Dev* **14**, 841–853 (2000).
62. Gustafson, E. A. & Wessel, G. M. Vasa genes: emerging roles in the germ line and in multipotent cells. *Bioessays* **32**, 626–637 (2010).
63. Malmstrom, M. *et al*. Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet* **48**, 1204–1210 (2016).
64. Bakke, I. & Johansen, S. D. Molecular phylogenetics of gadidae and related gadiformes based on mitochondrial DNA sequences. *Mar Biotechnol* **7**, 61–69 (2005).
65. Bangera, R., Odegard, J., Nielsen, H. M., Gjoen, H. M. & Mortensen, A. Genetic analysis of vibriosis and viral nervous necrosis resistance in Atlantic cod (*Gadus morhua* L.) using a cure model. *J Anim Sci* **91**, 3574–3582 (2013).
66. Torresen, O. K. *et al*. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* **18**, 95 (2017).
67. Harris, R. S. *Improved pairwise alignment of genomic DNA* PhD degree thesis, The Pennsylvania State University (2007).
68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
70. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
71. McKenna, A. *et al*. The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13**(Suppl 14), S8 (2012).
74. Kurtz, S. *et al*. Versatile and open software for comparing large genomes. *Genome Biol* **5** (2004).
75. Untergasser, A. *et al*. Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40** (2012).
76. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* (2017).
77. Russell, J. S. A. D. *Molecular cloning: A laboratory manual third edition*. 2344 pp (Cold Spring Harbor Laboratory Press, 2001).
78. Sovic, I. *et al*. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* **7**, 11307 (2016).
79. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
80. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515 (2010).
81. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
82. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
83. Haas, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).
84. Sali, A. & Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).
85. Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **12**, 345–364 (1992).

Acknowledgements

Thanks to Kim Præbel at The Arctic University of Norway (UiT) for providing DNA samples from haddock, Arctic cod, Greenland cod, polar cod and burbot. Thanks also to Nicola Barson for her valuable, critical review of the manuscript.

Author Contributions

S.L., M.P.K., O.A. designed the research. T.G.K. performed the genome assembly and analysis. S.L., M.P.K. and O.A. evaluated the analysis. S.L. and M.P.K. designed the experiments. T.A., K.H., M.P.K. carried out the experiments. M.C.D.R. performed protein modeling. T.G.K., O.A., M.P.K. and S.L. wrote the paper. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36748-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.